Observing Web Archives

The Case for an Ethnographic Study of Web Archiving

Jessica Ogden
University of Southampton
Southampton, UK
jessica.ogden@soton.ac.uk

Susan Halford
University of Southampton
Southampton, UK
susan.halford@soton.ac.uk

Leslie Carr
University of Southampton
Southampton, UK
lac@ecs.soton.ac.uk

ABSTRACT

This paper makes the case for studying the work of web archivists, in an effort to explore the ways in which practitioners shape the preservation and maintenance of the archived Web in its various forms. An ethnographic approach is taken through the use of observation, interviews and documentary sources over the course of several weeks in collaboration with web archivists, engineers and managers at the Internet Archive - a private, non-profit digital library that has been archiving the Web since 1996. The concept of *web archival labour* is proposed to encompass and highlight the ways in which web archivists (as both networked human and non-human agents) shape and maintain the preserved Web through work that is often embedded in and obscured by the complex technical arrangements of collection and access. As a result, this engagement positions web archives as places of knowledge and cultural production in their own right, revealing new insights into the performative nature of web archiving that have implications for how these data are used and understood.[1]

KEYWORDS

web archiving, knowledge production, STS, materiality, information labour

ACM Reference format:
Jessica Ogden, Susan Halford, and Leslie Carr. 2017. Observing Web Archives. In *Proceedings of WebSci '17, Troy, NY, USA., June 25–28, 2017,* 10 pages.
https://doi.org/10.1145/3091478.3091506

1 INTRODUCTION

The World Wide Web has emerged as the preeminent mechanism for global communication, political, economic and cultural exchange and more. Yet, at the same time, the Web is ephemeral. For a medium that has become pre-eminent, its dynamism and transience has become increasingly worrisome. These concerns have been illustrated in various longitudinal studies of link rot [55] and investigations which found that during a period between 2009 and 2012, on average 11% of online resources shared on social media failed to resolve one year later [60]. In this context, it is increasingly claimed that the ephemerality of the Web demands intervention to preserve web content - in web archives - that reconstruct sites and the 'web experience' for posterity [2]. However, there has been rather less attention to the nature of this intervention: how it is done and why this matters. This paper explores the critical decisions being made now that will shape future generations' ability to understand the history of the Web.

1.1 Web Archiving

Web archiving has roots in a wider digital preservation movement which emerged in the 1980s-1990s, led by memory institutions to develop strategies for addressing the rise of personal computing and the impact of digital artefacts on their abilities to capture and preserve 'records of social phenomena' [61]. This was particularly fuelled by fears over the so-called 'digital dark ages,' a term first used by Kuny [41] to describe a scenario where the development pace of technologies (used to produce digital objects) outweighs that of the investment in technologies, infrastructures and policies to preserve them long-term. As the world's information and communication platforms are increasingly born-digital and online, a diverse community of practitioners have positioned web archives as key to capturing and preserving digital cultural heritage, ensuring stability and access to pre-existing web resources and facilitating new knowledge via scholarly research. Web archives in their various forms - including social media archives - have thus become a sort of 'prosthesis' for the Web and a necessary pre-condition for any research into the Web(s) of the past and near-present.[2]

The history of web archiving has been documented to varying degrees in existing overviews [9, 13, 77] which chart the emergence of a field of practice around web archiving. Each have used a series of factors to characterise the domain over time, including: the tools and technologies used, the frequency and scale of selection/collection methods (e.g. broad versus targeted) and the various motivations behind the creation of web archives. These motivations may reinforce and represent, at least in part, a continuation of classical interpretations and analogue conceptions of the value and role of libraries and archives as institutions that provide access to cultural heritage, information and knowledge resources; facilitate evidence-based accountability and promote community memory and identity, amongst others [18, 26].

Web archiving projects have spanned from the large-scale collection of web resources by organisations such as the Internet Archive

[1] This paper is based on data collected and fieldwork undertaken by the first author as part of their PhD research.

WebSci '17, June 25–28, 2017, Troy, NY, USA.

ACM ISBN 978-1-4503-4896-6/17/06.
https://doi.org/10.1145/3091478.3091506

[2] Inspiration for this analogy is taken from Derrida's [20] treatment of 'technological devices for archiving' as prostheses for memory formation and storage.

and national libraries and archives, to networked 'rogue' collectives such as Archive Team[3] and DataRefuge,[4] and the individual efforts of activists and scholars creating web archives for their own purposes. Surveys of the field reveal a growing number of organisations that are web archiving, albeit with limited staff-time resources and with the majority of respondents using some form of external service to collect or manage their web archives [3, 52].

Much of the focus of the web archiving community has been on the continued development of technologies and practices for web collection development [35], with an increased attention in recent years on facilitating the scholarly use of web archives [21]. Web archiving has thus become representative of an interdisciplinary space where practitioners and scholars from a range of backgrounds (e.g. libraries, archives, information science, engineering and computer science) come together with scholars from a variety of disciplines (humanities, social sciences and computer science) to study the Web, its application and the users who derive value from its use.

This research examines this space from the perspective of web archival practitioners with the aim of documenting the practices that determine and create the archived Web within the context of the Internet Archive. From the point of view of Web Science, web archiving should not be seen as merely a series of technical solutions to the short-comings of the Web's infrastructure, but rather as an assemblage of contingent sociotechnical practices that shape what is known about the Web. These practices - and the entanglement of both human and non-human actors - are important for understanding the affordances of web archives and their implications for interpreting an increasingly dynamic Web. This type of engagement raises wide-ranging epistemological questions concerning the role of the web archive (and the archivist) in shaping knowledge about the Web. From this perspective, the mechanisms and circumstances surrounding the production of web archives are therefore fundamental to understanding them as 'new forms of social data' [45, p.31].

1.2 Problematising Web Archival Practices

The field of web archiving, the practices and technologies used to facilitate the creation, maintenance and use of web archives are all continually evolving, but not without their issues. The different legal, economic, technical and ethical challenges to preservation and access presented by both 'web documents' and social media have led to a mix of overlapping and divergent collection and access strategies. Existing overviews of the field [9, 48], literature reviews [54] and best practice documents [8, 56] provide models for mapping the general components of web archiving. However, some of this documentation has failed to keep up to date with the pace of web/archiving technology development, and is often limited to specific software, tools and professional contexts (e.g. library or bibliographic approaches to web archiving). Surveys of the field have yielded insights into the types and number of organisations participating in web archiving, the tools and services used, the types of content being archived, and the availability of institutional policy documents and staff-time resources [3, 29, 52]. These surveys provide useful information for understanding the broader landscape of web archiving. However, by virtue of the methods used, they do not provide in-depth information about the day-to-day decisions, activities and processes that facilitate web archiving in practice.

Web technologies are evolving at a faster pace than that of web preservation technologies and practice [21], presenting challenges for preserving an 'infinite stream with finite resources' [42]. The limitations of web harvesters in the face of new markup languages (e.g. HTML5), executable content (e.g. JavaScript and Flash) and other dynamic content (e.g. streamed multimedia, database-driven or password-protected) all often lead to missing elements in the representation of web resources in archives [56]. Problems for use are often presented as issues with the 'quality' of web archives, as measured by the relative 'likeness' between archived web objects and the 'live web.' As Brügger [14, p.108] has described, a combination of both collection decisions and 'technical problems' leads to archived web objects that are not *copies* of the live web, but rather *contingent constructions* where 'the process of archiving itself may change what is archived, thus creating something that is not necessarily identical to what was once online.'

There is an abundance of research within the field pertaining to improving the efficiency and quality of web crawlers, detecting change in web resources and automating frequency decisions associated with captures [64], as well as technical overviews of crawling technologies at the time of production [51]. Yet, little research exists around the interactive nature and structuring effects of algorithmic and automated agents in decisions around what and when to archive. Recent calls for the study of 'archival algorithmic systems' [70, p.11] point towards the need to further consider the performative nature of crawlers and other web archiving technologies, as well as the ways in which different environmental contexts shape the technological development [40] driving collection and access tools. Here web crawlers (non-human, automated agents, bots, algorithms, code) are conceived as not merely passive or objective participants in the collection of web resources [47], but are intricately implicated in the active shaping of the 'doing' of web archiving.

At the heart of conflicts over the collection and use of web archives are often problems with defining the boundaries of the web object to be captured and studied. Whilst recognising that web archives are by nature, inherently and necessarily incomplete [28], they are also highly 'subjective reconstructions' [11, 12] of what exists on the live Web at any given time. As argued by Dougherty and Meyer [21], often the specific ontological and epistemological assumptions made during the collection and curation of web archives are either not made explicit to potential users or are seen as an impediment to their use and/or re-use [21]. Collection decisions, whether thematic or broad, domain-based harvesting [57], the temporal dimensions of when to capture and for how long [46] and issues over geographic [72] and language [1] coverage and representation of the global Web(s), have all highlighted methodological concerns over provenance, the subjectivity of records, the lack of transparency and metadata for harvester algorithms and problems with the generalisability of potential research findings based on web archives.

[3] http://archiveteam.org
[4] https://www.datarefuge.org

Issues of provenance in web archives - the why, when, and how web archives are collected - have inspired calls for greater documentation around intent, particularly around what to preserve and why [76]. Others have focused on calling for greater transparency in how web archives are built, some with a particular focus on the Internet Archive [43]. However, practices that surround the capture and maintenance of web archives remain relatively understudied, and for initiatives that sit outside of mainstream memory institutions, continue to exist almost wholly unexamined. Investigations into how curation strategies and collection tools structure the nature of collections - for example, the timing, frequency and length of collection - have the potential to yield insights into the contingencies that lead to the archived Web(s). Recent qualitative research on web archival appraisal practices by Summers and Punzalan [70] underlines the value of such an approach for both situating web archiving within wider institutional/archival paradigms, as well as exposing undocumented practices largely missing from the archival record. Taking a different approach, Milligan et al. [50] contribute to a discussion of curatorial practices by reverse engineering selection through a comparison of algorithmic, manual and social media-generated web archives associated with the 2015 Canadian Elections. These studies, plus the previously mentioned work of Dougherty and Meyer [21], can be seen as complementary to this research both in methodology and in their aims to address how collection practices structure the nature of web archival engagement. The following sections outline the ways in which this paper aims to extend current knowledge of web archival practice.

1.3 Aims and Outline

This paper aims to examine the ways in which web archives and associated practices - as sociotechnical phenomena - are structured and organised by an array of actors and environmental factors that actively shape the practice of archiving. This research contends that these (often) undocumented activities and processes are critical for interpreting the affordances of web archives as contingent reconstructions of the previously live Web.

The argument for a practice-based approach to web archiving is further examined through a theoretical framework. This framework draws on aspects of archival theory and Science and Technology Studies (STS) investigations into the *materiality* of information mediation practices. Here, examining web archiving from the perspective of the materiality of practice enables two paths of inquiry. The first acknowledges the ways in which new knowledge is embedded and produced as part of the creation of web archives; the second addresses how practice transforms knowledge through the maintenance of web archival systems.[5]

A case for the chosen methodology is then made along with a description of the data collection, in an effort to highlight the ways in which this research may have implications for a Web Science approach to examining sociotechnical practices. The theoretical approach is then worked through empirical examples and discussion drawn from vignettes of ethnographic fieldwork at the Internet Archive. A brief summary discussion is provided that highlights some of the broad findings at this stage of the empirical research.

[5]For further inspiration into the types of inquiries enabled by a focus on the materiality of media technologies, see Gillespie et al. [25].

2 FRAMING PRACTICE ENGAGEMENT

Postmodernism and its application in archival and social theory lays the groundwork for critically engaging with web archiving as knowledge production. From this perspective, there has been a longstanding interest in knowledge production activities where the record is often positioned as 'evidence of process, of activity, [and] of transaction' [34, p.12].

First coined by Stoler [68], 'the archival turn' denotes a shift from 'archive as source' to 'archive as subject', signalling wide-ranging epistemological questions concerning the role of the archive (and the archivist) in shaping and legitimising knowledge and particular ways of knowing. Cook [16, p.4-5] argues that postmodern archival theory represents a fundamental paradigm shift within a community of practice that hitherto had been largely grounded in scientific rationalism, 'archival science,' the merits of record stability and the objective role of archivists; towards one which recognises the incompleteness of records and attends to context and the interpretive role of archivists in the construction of social memory. This signified a theoretical move away from the framing of archives as 'sites of knowledge retrieval' towards a recognition that archives are deeply reflective of and implicated in the production of knowledge [68, p.90]. The conditions of historical narrative-making are intrinsically tied to the processes of archival construction, where certain narratives are privileged and others marginalised through the active reshaping by the archivist [62, 71, 74]. These usually invisible practices involved in the maintenance of archives, have ramifications for the ways in which archival holdings are then (re)presented as being a view from nowhere or of 'all possible statements' rather than 'the law of what can be said' [6].

In contrast, Brown and Davis-Brown [10] focus on the 'technical-rational work' of archivists - or the everyday decisions and practices related to the collection and maintenance of archives. They characterise a profession where the 'explicitly political *who* is often reduced to the technically instrumented *how*' - a sentiment also echoed by critical information studies [7], as well as practitioners from within the web archiving field [76]. In light of this, an engagement with the political nature of web archival practice would then include an examination of what comes to count as the 'professional decision-making' [10] involved in a host of activities that mark the everyday tasks of (organisational) archivists.

The turn also ascribes the importance of unpicking digital information technologies and not falling into the trappings of either equating them to analogue technologies, nor essentialising their capabilities or potentialities for capturing the cultural record. Cook [17], describes the impact of the transition to electronic archival records and the role of postmodern critique in strengthening the role of the archivist in the digital age:

> We will move from databases to knowledge bases. We will move, in the language of the post-modernists, to re-contextualize our activities: we will reorient ourselves from the content to the context, and from the end result to the original empowering intent, that is, from the artifact (the actual record) to the creating processes behind it, and thus to the actions, programmes, and functions behind those processes [17, p.410-411].

As such, archives can be further considered in light of the digital technologies that are intimately tied to both the production and preservation of web resources. Waterton [75, p.653-654] and others make the case for examining digital archives and the 'generative capabilities' of technologically-enabled data, information, and knowledge which are in an 'eternal process of becoming' [33]. Questions are therefore raised regarding how prior conceptualisations of significance and potential use are manifested in the ways in which digital 'objects' are collected and preserved. Pinch and Henry [58]'s notion of the 'materiality of knowledge' is useful here for considering the ways in which knowledge is also 'embedded in physical artefacts, technologies, and ways of doing things'. This is manifested in how practice produces 'material bodies' [4, p.808-809] but also how materiality is bound to and embedded in practice. In this instance Barad [4] is referring to how the body (e.g. the anatomy and physiology) actively contributes to the processes of 'materialisation,' but in the case of web archives it warrants an examination of how the materiality of technologies (platforms, tools, interfaces, code, algorithms) is both implicated in the production of archives but also potentially produced through practice.

In this theoretical context, we extend the focus on materiality and practice, to include the role that labour plays in the production and maintenance of web archives. Whereas practice embodies the action, artefacts and tacit knowledge that gives meaning to both [15]; labour is conceptualised here to encompass the work it takes to produce and transform web archives into information sources. To frame our analysis, we draw on Downey's [23] concept of '*information labor*', or the human and algorithmic labour that '[enables] and [constrains] the constant circulation of information.' Downey argues that a focus on information labour, particularly in the context of media literacy, reveals the contingent 'social relations' that exist between information producers and consumers, through the process of acknowledging the work of agents that are often obscured by the technical arrangements of access. Here, by acknowledging and exploring the interconnected collection and maintenance work of both human agents (web archivists, engineers, users) and non-human agents (algorithms, bots, code) in web archiving, we open the doors for new enquiry into the value and role of the work of web archivists in the production of knowledge.

3 METHODOLOGY

Ethnographic methods were chosen to document the routine activities of archival practices and the 'typical patterns of work' [32, p.169] through the use of observation, interviews and documentary sources for understanding the day-to-day activities of web archivists.

3.1 The Case for an Ethnographic Approach

An 'upward trend' has been observed in the use of ethnographic methods in library and information science research [38]. Although there are comparatively fewer examples of ethnographic methods being used within the context of archives, ethnography has been identified as a means for conducting in-depth, comparative and cross-cultural studies of archival practice [49]. Gracy's [31] ethnographic work of archives provides the stimulus and methodological framework for an 'archival ethnography', or rather, the study of archival practice *in situ*:

> Archival ethnography is a form of inquiry which positions the researcher within an archival environment to gain the cultural perspective of those responsible for the creation, collection, care and use of records [30].

Specifically focused on record creation, other studies have used ethnographic observation to understand archives as a form of knowledge production in scientific laboratories [63]; to investigate communication and organisational accountability in record-keeping [79]; and to understand both the technical and the social apparatuses that facilitate record creation and maintenance within organisational contexts [73]. The motivations for these studies, as well as their findings, have all emphasised the importance and effect of context on the production and management of the archival record and reinforce the value of ethnographic methods for documenting situated practices and the wider interactions between archivists, archival institutions and users of archives.

Ethnographic methods have been argued to enable a more 'complex' appreciation of the development, use and role of technologies in society - beyond a view of technologies as simply 'functional instruments' [59, p.110]. One key assumption of this research design is that the direct observation of technologies and their use in web archiving is central to understanding and documenting practices. Suchman [69] provides relevant assistance here, defining *technology* as 'the assemblage of skilled practices and associated logics characteristic of modern industrial societies' and *artefacts* as the 'material production of skilled practice'. This approach allows for an exploration of the materiality of web archiving through a discussion of the relationship between practice and the production of digital artefacts, as well as the role of the environment (the policies, activities, infrastructure, and communities) that actively inform them.

Our aim, clearly, is not to fetishise web archives as either technological objects or sociomaterial culture, a point mirrored in STS debates which attend to the pitfalls of determinism and advocate for the avoidance of reductionist conceptualisations of either social or technical agents as essentially defined by the other [69, p.165]. Rather, we understand web archives as produced by an assemblage that require attention be paid to both the structure and agency of technical actors, as well as the sociocultural elements of technical practices - as evidence for the ways in which 'cultural values are enacted, produced, shared, reified, represented and reaffirmed' [22].

3.2 Data Collection and Analysis

A combination of interviews, observation records and documentary sources were collected over a four week period in collaboration with web archivists, engineers and managers at the Internet Archive. Both organisational (for observations) and individual consent (for interviews) were sought and received prior to data collection. In accordance with this consent all individual participant names reported here are pseudonyms.

Ethnographic interviews are a form of research interview that are distinguished by not only the types of questions that are asked, but also the ways in which the interviews are broken down into

'ethnographic elements' or 'speech acts' designed to elicit specific types of cultural responses from informants [65]. They were used in combination with observation as a mechanism for developing rapport with informants, as well as to clarify existing observation records and focus subsequent observation activities. In an effort to not to predetermine what practices were discussed, the interviews took a largely unstructured approach using a combination of descriptive, structural and contrasting questions in direct response to the answers provided by informants within the context of the interview. A number of participant-led 'walkthroughs' were undertaken as part of the interviews in order to allow practitioners to narrate their daily activities in real-time. In total, 16 interviews were conducted with 11 staff members at the Archive. The length of each interview varied, ranging from 20 minutes to over 2 hours, all of which were audio recorded. Most interviews took place in person, barring two which took place on Skype as the informants worked remotely. All interview data were fully transcribed.

The purpose of ethnographic observations is to provide access to 'practices and actions as they unfold' [5, p.55], in the form of 'non-elicited data' that can allow insights into the implicit and embodied activities that form everyday life. This is predicated on the notion that some actions cannot be articulated by participants/or 'insiders' through other research methods such as interviews. Boellstorff [5] and others [53] have long argued that 'elicitation methods' (such as interviews) cannot be a substitute for observation as there are inherent differences and disconnects between *what people do* and *what they say they do*. Furthermore, people do not always have the perspective or ability to report on all aspects of the processes (particularly cognitive processes) that underly the decisions and activities that make up practice [53]. Observation therefore, offers another window into understanding the relationship between meaning and action in everyday practice.

Ethnographic records were made documenting participation activities - including all-staff and tool development meetings, staff and application training seminars, troubleshooting and quality assurance tasks with participants - all with the aim of providing the basis for 'thick descriptions' [24] of practices. Observation pro-forma were not used, however ethnographic records were created for observations describing: *what was done* - action, activities, *what was made and used* - 'cultural artefacts' and *what was said* - speech acts, discursive activities [66, p.10-12]. Varying degrees of researcher participation was undertaken as part of the observations (as deemed appropriate by the host), where 'passive' participant observation was used during staff meetings, more 'active' participant observation took place during workshop and community events held at the Internet Archive.[6]

Documentary sources in various forms were collected to supplement the observations and interviews. 'Documents' are used here to refer to various types of materials (not just textual) produced by organisations, communities and individuals to describe procedures, policies and preferred ways of practicing web archiving. The use of documentary sources here offers insights into aspects of (otherwise) implicit knowledge that underlies practices [15, p.401].

Strategies for synthesising ethnographic data advocated by Spradley [65] were used to develop the analytical themes. This involved the identification of 'things informants know' in an effort to elicit everyday practices through the various kinds of participant knowledge (e.g. knowledge about crawl behaviours, scoping rules, reporting tools) and their connections to ways of doing things (e.g. maintenance and quality assurance tasks, de-duplication techniques). All data were analysed together in order to identify common themes present across the dataset. Lists and groupings were compiled by repeatedly listening to the audio recordings, reading the transcripts and observation records, and comparing the subject matter across the data. The methods used provided the opportunity for mapping heterogenous data and highlighted particular groupings of practitioner knowledge and activities for further analysis. The themes presented here are not exhaustive but rather, subjective and reflective of both the aims and current state of this research.

4 FINDINGS AND DISCUSSION

In what follows we explore the theoretical framework described above through an initial analysis of data collected from fieldwork carried out at the Internet Archive ('The Archive'). This is the result of four weeks spent at the Archive in October-November 2016 and February-March 2017. Before addressing the findings, some background is provided to set the scene. Through the presentation of findings that follows, the concept of *web archival labour* is examined and explored through the *knowledge work*, *breakdown* and *repair* activities that shape web archiving at the Archive.

4.1 Site Background and Setting

In 1996, Brewster Kahle and Bruce Gilliat established the Internet Archive as a nonprofit organisation alongside *Alexa Internet*, a commercial web indexing service [39]. The Internet Archive headquarters are based in San Francisco, California, a state which officially designated them a library in 2007 [37]. Since 1999, the Internet Archive has expanded their holdings beyond archived web content to provide web-based access to both digitised and born-digital resources, including but not limited to: books, audio, film/video, images, documents, software and video games [78]. According to web traffic statistics provided by Alexa Internet (now a subsidiary of Amazon), the Internet Archive regularly ranks between the 250-350th most visited website globally, with roughly 30% of search engine results driven by users looking for the Wayback Machine.[7]

The Archive forms a central component of the web archiving landscape through their provision of both large-scale web indexing and preservation services, as well as steering the direction of standards and practice through the development of tools and technologies that support web archiving at scale. The Archive is widely considered the largest web archive in the world, containing in excess of 15 petabytes of 270+ billion captures of web content [27]. They are one of the few institutions to employ 'exhaustive' global web crawling, which harvests web resources both within and beyond national domain boundaries or thematically-driven captures for the primary purpose of preservation (rather than indexing for other search or commercially-driven purposes). The collection of web archives at the Archive was originally (solely) facilitated

[6] For a further breakdown of the types of participation levels common to participant observation, see Spradley [66, p.58].

[7] http://www.alexa.com/siteinfo/archive.org [Accessed: 3 February 2017]

by the Alexa Internet 'toolbar', a browser-based plugin developed for the purposes of improving early-web navigation and analytics [44, p.274]. Based on user navigation, the toolbar captures and preserves each web page as it is visited, subsequently donating it to the Archive with a six month access embargo (a practice that continues to this day).

The Archive released the Heritrix web crawler as open source in 2002 alongside leading the development of the Wayback Machine software which currently forms the predominant mechanism by which users collect and access web archives. In 2006, the Archive began supporting *Archive-It*, the widely used subscription-based web archiving service. The Archive-It team work as part of the 'web group' at the Archive, providing technical and storage support for service subscribers and partner organisations. As such, this case study offered the opportunity to observe some of the ways the Archive-It team supports the development of tools and practice for web archiving.

4.2 Mapping Practice Roles and Activities

The web archival activities of the Archive can be broken down into three broad areas: *crawling*, *access* and *tool development*. The Archive is engaged in crawling at many different levels and thus this area includes the crawl activities that are self-directed, those that are fully undertaken at the direction of other organisations and those crawls that are initiated and directed by other organisations using their subscription service Archive-It. The Archive facilitates access and hosting for crawls undertaken by themselves and others through the Wayback Machine, including the provision of tools that allow others to deposit and donate their own WARCs to the collection. And lastly they facilitate web archiving through their ongoing tool and technology development for crawling and replay. In practice, these three areas do not sit in isolation of one another and represent a working environment of overlapping roles, tasks and activities at the Archive. The work that makes up *web archival labour* permeates each of these activity areas, explored further in the following sections.

4.3 Knowledge Work: Crawling and Curating

Elsewhere, others have acknowledged a certain pre-occupation with abundance and 'plenitude' at 'universal archives' such as the Internet Archive, an observation which De Kosnik [19, p.95] argues implicitly denies, or at least distracts from, attention to any selectivity in archiving. Although it is clear that the Archive is overtly concerned with abundance and scale in their endeavours to capture more, our observations point strongly towards efforts on their part to increasingly shape and prioritise the web resources that they capture. This point was made clear at the Archive's 20th Anniversary Party (2016) when Kahle announced that the Archive had (at that point) archived '273 billion webpages from over 361 million websites' with the help of robots and '1000 librarians.' Here and in an interview with Kahle, the Archive signals the creation of the Archive-It subscription service as a significant step towards archiving more selectively, by providing librarians with the tools to save web resources. However, the selection narrative is more complex than this, or as an informant indicated when I enquired about the Archive's appraisal practices: 'the process is strategically mish-mashed.' Collectively, these strategies (some of which are discussed below) can be seen as one component of what Downey [23] calls the '*knowledge work*', or the 'high value labour' that goes into the production of information that is obscured, or marketed as either automated or infinite. The notion of knowledge work is explored below as it relates to the curation of seed lists and what one informant called the 'hybrid' crawling activities of the Archive.

One informant relayed that a common misconception about the Archive's crawling activities is that they employ 'one giant crawler' to archive the global Web.'[8] This is repeated in research literature and popular media articles around the perceived automation of the Archive's crawling activities. At any given moment, in fact, the Archive has an (unknown) number of crawlers engaged in selective archival activities. A key priority was to begin to map these crawling events during fieldwork, in an effort to understand each as a contingent, sociotechnical assemblage. These crawl events or '*crawl modalities*' [70] define what is collected in the Archive. Each is shaped by different motivations, priorities and approaches to web archiving.[9]

Looking further back in history to 2010, a moment of significance can be observed when the Wayback team started engaging in and directing their own global crawls. Several informants described the motivations that led to this shift in direction, which was driven (at least in part) by the perceived inadequacies of the crawl data that were seeded, crawled and donated by Alexa Internet to the Archive. The Alexa crawling approach[10] is raised here as it is representative of a historical focus on the *popularity* of sites as a factor in the selection of seeds, one which (until recently) continued to influence the ways in which crawls were prioritised - even in those crawls directed by the Archive itself. The use of Alexa's 'top million' sites as the starting seeds for the 'wide crawls' (see below) was discussed as a common place practice but often resulted in the over-prioritising of popularity as an indicator of the value in capturing certain web resources - with Gregory claiming that: 'over 50% of our wide crawl was from 2,000 websites.' Further complaints were relayed about the quality of Alexa crawls as they do not capture images or embedded dynamic resources, often leading to web archives with extensive missing elements.

Various crawl events have subsequently become associated with the Archive's global crawling efforts, including for example, the 'survey crawls' and 'wide crawls.' Survey crawls are being used to supplement wide crawls by taking a snap shot of the home page of every domain/host ever identified by the Archive. Wide crawls are run twice a year over 4-6 months, though as Arthur described they had originally envisaged the crawl cycle to run 4-6 times a year. Wide crawls start from a seed list (initially the Alexa top million, as described above) and are then allowed to run autonomously, the bot

[8] personal communication

[9] The term 'crawl modality' is borrowed from Summers and Punzalan [70] who in a study of web archival appraisal practices describes the different ways in which crawling activities were broadly conceptualised and implemented. Although different (but overlapping) crawling modalities were identified at the Archive, the focus is similarly on determining how these modalities come to pass and how they shape what is collected.

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

following each outlink until 'it doesn't produce any interesting data any more.' When asked how wide crawls were stopped, Arthur said they have to regularly check on each Heritrix instance by manually go into the machine in question and looking at the logs to see what is actually being captured, a process they described as 'daunting.' Alex described what they were looking for when they examine the logs, which largely involves a visual inspection of the domain names contained within the capture URLs, watching out for strings that resemble 'calendar traps,' pornography and endless Facebook sites. In direct result of some of the manual labour required to shape and monitor the large-scale crawls at the Archive, engineers began developing various ad-hoc tools to mitigate the need for interacting with the harvester logs and other shell scripts. One such tool is something Arthur calls the 'Domain Browser tool' used in conjunction with *Hericrawler*, a crawl queue management system the Archive developed for orchestrating large-scale crawls:

> The domain browser manual tool is for identifying undesirable domains. It's used to establish and prioritise 'shades of gray,' for example only crawl this site if there are no other sites to crawl. It's used as a ranking mechanism for prioritising domains based on time, resources and place in the queue, as certain important URLs can get blocked by many instances of unimportant URLs. For example people linking to Facebook pictures can create an infinite loop of queued Facebook links because of the nature of the graph. These types of sites are really slow to crawl as they are hosted on a single site, which must be crawled in succession because of the nature of the Heritrix crawler. Each domain/host is assigned a budget and the crawler is paused if it reaches its budget.

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

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

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

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

> I do a lot of link analysis where I study the hyperlink structure of our crawls and try to figure out in certain pockets, use some rank methodologies to figure out 'oh these are important resources,' for instance they have a lot of links to them or traffic is really high - let's seed the crawl with those. The most recent wide crawl I took the most linked to pages from every single website, so 230 million websites [...] and instead of crawling the Alexa top million, let's crawl this bit. Sort of like a hybrid survey and wide model [...]. And we found resources that we had never crawled before. I'm not saying one is better than the other, I'm saying that hybridising this process might be one way of balancing the scales a little bit.

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

> The way I think of it [...] there's three branches, there's *popularity*, there's *novelty* and there's the *risk of going away*. How do you achieve that balance? You want to get stuff that people are using - not just junk that is on the Web that you're just filling up the servers with that won't ever be found useful, like calendar pages, things like that, crap [...] - there's no novelty. It's new? We want to make sure it's preserved because it just came out, it's a new article, it's a new website. And then there's the risk of going away [...] - if you're going to shut down this service - Vine is going away - we jump

> in and crawl. So as we're crawling the Web can we do a good job of sort of achieving that balance? We don't quite know what the solution is to achieving that balance.

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

These methods highlight some of the ways the Archive is leveraging the power and labour of 'the crowds' - through the users of Twitter, Wikipedia and Save Page Now - to not only diversify and 'balance' the domains/types of resources that are archived, but also (implicitly) co-opt and transform these users into potential stakeholders. Furthermore, the Archive is in multiple ways, leveraging the web archives amassed in the Wayback Machine project over the last 20 years to continuously increase their net resources.

4.4 Breakdown and Repair: Maintaining Archives

Many informants described scenarios where human mediation was required in either otherwise (seemingly) automated processes. Examples that require manual intervention include: running *patch crawls* when bots failed to crawl designated seeds, or in the event of a *crawler trap* where bots are trapped in an infinite loop of seed requests (discussed above), or when *missing* or *altered* elements are observed in the playback of archived web pages. Fundamentally, each of these boil down to issues surrounding either the capture or replay of web resources. Borrowing from Star and Ruhleder [67], these moments could be considered a form of *breakdown* which in turn reveals the contingent assemblage of processes that for example, enable the capture of intended web resources or provide 'high fidelity' access to the archived collections.

The issues around the *quality assurance* of collection capture and playback is not an unknown dimension of web archival practice, however, the processes that are undergone to mitigate these issues are not well documented. Here we draw on the work of Jackson [36] who advocates for an examination of the moments of *breakdown*, *repair* and *maintenance* of technologies, in that they redirect attention to the act and 'ethics of care' and embody the creation of value by their maintainers. In other words, the practice and processes involved in fixing and maintaining technologies can be used as evidence for their worth by those who sustain these practices over time. These moments of repair in web archiving are present throughout the study, including the repair and maintenance of crawl data and crawling technologies, Wayback and access tools, as well as the repair of broken links on the live Web by the Archive. Some of these practices are discussed further below, as they highlight some of the factors that influence the decision-making and technical processes that enable the repair of web archives (and the technologies that enable access).

A few training sessions for a web archivist on the Archive-It team, Karen were observed. Through listening to Karen's Q&A with other team members, certain junctures were highlighted where support staff are regularly required to prioritise activities, particularly in response to the 'quality' or 'completeness' of web archives (and as raised by partner organisations). To set the scene a bit, we draw on interpretive notes following an observation session:

> *In the second session we all sat on the couch in the pit. Lydia asked Karen if she wanted to continue the training they began in the morning, which was aimed at addressing a recent support request that came in from a partner. When the conversation got a bit technical - as a result of Karen asking increasingly detailed questions about how certain seeds and test crawls are rendered in playback - Mike (a support engineer) was called over by Lydia. Mike walked over and leaned on the couch and began to explain some common differences between capture and playback issues. The consensus from Lydia and Mike seems to be that the first task in the support role is to determine whether the issue is related to capture or playback. Karen is concerned about waiting to solve issues based on partner requests, advocating for the team to be more actively QA'ing collections in case they are at risk of disappearing. Mike responds that there are different issues at play (including time and resources) and that it's key to understand that capture issues will always take precedent over playback issues - for exactly that reason.*

From this vignette several points can be interpreted. First, we can see that Karen is getting to grips with a key aspect of the role of the web archivist at Archive-It, which (in conjunction with support engineers) is to determine the difference between playback and collection issues and respond accordingly. Second, the comment that capture problems will always trump replay problems is insightful. It emphasises the goal of capture and reflects the underlying motivation driving activities - the fear of disappearance. This observation also emphasises the active role of the web archivist and support engineer in the processes that shape the 'fidelity' of web archives. They are implicitly driven by time and resource constraints but they are also active participants in the practice of choosing which support issues are prioritised. Mike indicated that specific repair tasks are prioritised in each daily 'stand-up' where team members report on upcoming development goals. Web archivists and programme managers will mark tasks as high priority either because the content is at risk of changing or going way or because the request comes from a high priority partner (as both were the case for the whitehouse.gov and End of Term archives that were being captured at the time).

5 SUMMARY AND FUTURE WORK

The aims of the paper were to examine how web archives are structured by the practices associated with collecting and maintaining archives. The use of an ethnographic approach, with a focus on observing practice as web archival labour at the Internet Archive has revealed a number of insights. The data points towards a complex system of knowledge and maintenance work for prioritising which web assets to collect and repair. The Archive is leveraging their extensive existing archives for understanding networked linking behaviour in an effort to balance the breadth and depth of crawling activities, while discovering new sources for identifying websites to crawl based on measures of popularity, 'novelty' and sites that are endanger of going offline. The team has devised multiple mechanisms for identifying different types of 'undesirable domains,' including rule-based link pattern-matching and the development of 'gamified' tools for the manual curation of sites.

Collectively, the efforts of the Archive can be seen as knowledge work, and these activities, seen in combination with other practices around the prioritisation, repair and maintenance of tools and archives all have ramifications for how web resources are transformed for use. It is the labour of non/human agents that enables the preservation and ingestion of information from the Web into the Archive, and then once again back to the Web where archives are reassembled via the Wayback Machine. Although imperfect, this labour is increasingly recognised as an essential element of the web architecture. The information labour and knowledge work of potential web archival users is therefore intimately tied to the web archival labour of the Internet Archive. As the global Wayback Machine currently provides access to billions of webpages - often inaccessible elsewhere - editorial decisions have implications for not only the fidelity of archived captures, but indeed whether or not certain parts of the Web are preserved at all. Future work will aim to address any shortcomings in the ethnographic methods presented by further examining the labour of algorithmic and other non-human actors (e.g. hyperlinks, software) implicated in these processes. By continuing to make these practices visible, both the contingencies and value of this labour are revealed and open a new window on this increasingly vital activity.

ACKNOWLEDGMENTS

This work was supported by the UK Engineering and Physical Sciences Research Council and the Web Science Centre for Doctoral Training, Grant No. EP/G036926/1. The authors would also like to thank the Internet Archive and staff for opening their doors and being so generous with their time and feedback.

REFERENCES

[1] Ahmed AlSum, Michele C. Weigle, Michael L. Nelson, and Herbert Sompel. 2014. Profiling web archive coverage for top-level domain and content language. *International Journal on Digital Libraries* 14, 3 (2014), 149–166. DOI: https://doi.org/10.1007/s00799-014-0118-y

[2] William Y. Arms, Roger Adkins, Casey Ammen, and Allene Hayes. 2001. Collecting and Preserving the Web: the MINERVA Prototype. *RLG DigiNews* 5, 2 (April 2001). http://worldcat.org/arcviewer/2/OCC/2009/08/11/H1250005040496/viewer/file2.html

[3] Jefferson Bailey, Abigail Grotke, Kristine Hanna, Cathy Hartman, Edward McCain, Christie Moffatt, and Nicholas Taylor. 2014. *Web Archiving in the United States: A 2013 Survey.* Technical Report. The National Digital Stewardship Alliance. 1–24 pages. http://www.digitalpreservation.gov/documents/NDSA_USWebArchivingSurvey_2013.pdf

[4] Karen Barad. 2003. Posthumanist Performativity: Toward an Understanding of How Matter Comes to Matter. *Signs* 28, 3 (2003), 801–831. http://www.jstor.org/stable/10.1086/345321

[5] Tom Boellstorff. 2012. Rethinking Digital Anthropology. In *Digital Anthropology*, Heather A. Horst and Daniel Miller (Eds.). Bloomsbury Publishing, London, 39–60.

[6] Geoffrey C. Bowker. 2005. *Memory Practices in the Sciences.* MIT Press, Cambridge, MA.

[7] Geoffrey C. Bowker and Susan Leigh Star. 1999. *Sorting Things Out: Classification and Its Consequences* (paperback ed.). MIT Press, Boston, MA.

[8] Molly Bragg and Kristine Hanna. 2013. *The Web Archiving Life Cycle Model.* Technical Report. The Archive-It Team and Internet Archive. http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf

[9] Adrian Brown. 2006. *Archiving Websites: A Practical Guide for Information Management Professionals.* Facet, London.

[10] Richard Harvey Brown and Beth Davis-Brown. 1998. The making of memory: the politics of archives, libraries and museums in the construction of national consciousness. *History of the Human Sciences* 11, 4 (1998), 17–32.

[11] Niels Brügger. 2008. The Archived Website and Website Philology. *Nordicom Review* 29, 2 (2008), 155–175.

[12] Niels Brügger. 2009. Website history and the website as an object of study. *New Media & Society* 11, 1-2 (2009), 115–132. DOI: https://doi.org/10.1177/1461444808099574

[13] Niels Brügger. 2011. Web Archiving - Between Past, Present, and Future. In *The Handbook of Internet Studies*, Mia Consalvo and Charles Ess (Eds.). Wiley-Blackwell, Oxford, 24–42.

[14] Niels Brügger. 2012. When the Present Web is Later the Past: Web Historiography, Digital History, and Internet Studies. *Historical Social Research* 37, 4 (2012), 102 – 117.

[15] Christian Bueger. 2014. Pathways to practice: praxiography and international politics. *European Political Science Review* 6, 3 (Aug. 2014), 383–406. DOI: https://doi.org/10.1017/S1755773913000167

[16] Terry Cook. 2001. Archival science and postmodernism: new formulations for old concepts. *Archival Science* 1, 1 (2001), 3–24. DOI: https://doi.org/10.1007/BF02435636

[17] Terry Cook. 2007. Electronic Records, Paper Minds: The Revolution in Information Management and Archives in the Post-Custodial and Post-Modernist Era. *Archives & Social Studies: A Journal of Interdisciplinary Research* 1, 0 (March 2007), 399–443.

[18] Terry Cook. 2013. Evidence, memory, identity, and community: four shifting archival paradigms. *Archival Science* 13, 2 (2013), 95–120. DOI: https://doi.org/10.1007/s10502-012-9180-7

[19] Abigail De Kosnik. 2016. *Rogue Archives: Digital Cultural Memory and Media Fandom.* MIT Press, Cambridge, Massachusetts; London, England.

[20] Jacques Derrida. 1995. *Archive Fever: A Freudian Impression.* University of Chicago Press, Chicago and London.

[21] Meghan Dougherty and Eric T. Meyer. 2014. Community, Tools, and Practices in Web Archiving: The State-of-the-Art in Relation to Social Science and Humanities Research Needs. *Journal of the Association for Information Science and Technology* 65, 11 (2014), 2195–2209.

[22] Paul Dourish and Genevieve Bell. 2011. *Divining a Digital Future: Mess and Mythology in Ubiquitous Computing.* MIT Press, Cambridge, MA.

[23] Gregory J. Downey. 2014. Making Media Work: Time, Space, Identity, and Labor in the Analysis of Information and Communication Infrastructures. In *Media Technologies: Essays on Communication, Materiality, and Society*, Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot (Eds.). MIT Press, Cambridge, Massachusetts; London, England, 141–165.

[24] Clifford Geertz. 1973. *The Interpretation of Cultures.* Basic Books, New York.

[25] Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot. 2014. Introduction. In *Media Technologies: Essays on Communication, Materiality, and Society*, Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot (Eds.). MIT Press, Cambridge, Massachusetts; London, England.

[26] Anne J. Gilliland-Swetland. 2000. *Enduring Paradigm, New Opportunities: The Value of the Archival Perspective in the Digital Environment.* Technical Report 89. Council on Library and Information Resources, Washington, D.C. https://www.clir.org/pubs/reports/pub89/pub89.pdf

[27] Vinay Goel. 2016. Defining Web pages, Web sites and Web captures. (Oct. 2016). http://blog.archive.org/2016/10/23/defining-web-pages-web-sites-and-web-captures/

[28] Daniel Gomes, Sérgio Freitas, and Mário J. Silva. 2006. Design and Selection Criteria for a National Web Archive. In *Research and Advanced Technology for Digital Libraries*, Julio Gonzalo, Costantino Thanos, M. Felisa Verdejo, and Rafael C. Carrasco (Eds.). Lecture Notes in Computer Science, Vol. 4172. Springer Berlin Heidelberg, 196–207. http://dx.doi.org/10.1007/11863878_17

[29] Daniel Gomes, João Miranda, and Miguel Costa. 2011. A Survey on Web Archiving Initiatives. In *Research and Advanced Technology for Digital Libraries*, Stefan

Gradmann, Francesca Borri, Carlo Meghini, and Heiko Schuldt (Eds.). Lecture Notes in Computer Science, Vol. 6966. Springer Berlin Heidelberg, 408–420. http://dx.doi.org/10.1007/978-3-642-24469-8_41

[30] Karen F. Gracy. 2001. *The Imperative to Preserve: Competing Definitions of Value in the World of Film Preservation.* PhD. University of California, Los Angeles.

[31] Karen F. Gracy. 2004. Documenting Communities of Practice: Making the Case for Archival Ethnography. *Archival Science* 4, 3 (2004), 335–365. DOI: https://doi.org/10.1007/s10502-005-2599-3

[32] Martyn Hammersley and Paul Atkinson. 2007. *Ethnography: Principles and Practice* (third ed.). Routledge, London and New York.

[33] Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14, 3 (1988), 575–599.

[34] Verne Harris. 2000. Law, Evidence and Electronic Records: Strategic Perspective from the Global Periphery. *S. A. Archives Journal* 41 (2000), 3–19.

[35] Helen Hockx-Yu. 2014. Access and Scholarly Use of Web Archives. *Alexandria: The Journal of National and International Library and Information Issues* 25, 1-2 (2014), 113–127. DOI: https://doi.org/10.7227/ALX.0023

[36] Steven J. Jackson. 2014. Rethinking Repair. In *Media Technologies: Essays on Communication, Materiality, and Society*, Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot (Eds.). MIT Press, Cambridge, Massachusetts; London, England, 221–239.

[37] Brewster Kahle. 2007. Internet Archive officially a library. (June 2007). https://archive.org/post/121377/internet-archive-officially-a-library

[38] Michael Khoo, Lily Rozaklis, and Catherine Hall. 2012. A survey of the use of ethnographic methods in the study of libraries and library users. *Library & Information Science Research* 34, 2 (2012), 82 – 91. DOI: https://doi.org/10.1016/j.lisr.2011.07.010

[39] Michele Kimpton and Jeff Ubois. 2006. Year-by-Year: From an Archive of the Internet to an Archive on the Internet. In *Web Archiving* (first ed.), Julien Masanès (Ed.). Springer, Berlin, Heidelberg, 201–212.

[40] Rob Kitchin. 2016. Thinking critically about and researching algorithms. *Information, Communication & Society* 20, 1 (2016), 14–29. DOI: https://doi.org/10.1080/1369118X.2016.1154087

[41] Terry Kuny. 1997. A Digital Dark Ages? Challenges in the Preservation of Electronic Information. In *Proceedings of the 63rd International Federation of Library Associations and Institutions*. Copenhagen, Denmark.

[42] Kalev Leetaru. 2015. Why It's So Important To Understand What's In Our Web Archives. (Nov. 2015). http://onforb.es/1VDPHPH

[43] Kalev Leetaru. 2016. The Internet Archive Turns 20: A Behind The Scenes Look At Archiving The Web. (Jan. 2016). http://www.forbes.com/sites/kalevleetaru/2016/01/18/the-internet-archive-turns-20-a-behind-the-scenes-look-at-archiving-the-web/#747db6257800

[44] Jessica Livingston. 2007. *Founders at Work: Stories of Startups' Early Days.* Apress, United States of America.

[45] Deborah Lupton. 2015. *Digital Sociology.* Routledge, London.

[46] Peter Lyman. 2002. Archiving the World Wide Web. In *Building a National Strategy for Preservation: Issues in Digital Media Archiving*. Council on Library and Information Resources and the Library of Congress, 38–51. http://www.clir.org/pubs/reports/pub106/web.html

[47] Noortje Marres and Esther Weltevrede. 2013. Scraping the Social? *Journal of Cultural Economy* 6, 3 (2013), 313–335. DOI: https://doi.org/10.1080/17530350.2013.772070

[48] Julien Masanès (Ed.). 2006. *Web Archiving* (first ed.). Springer-Verlag Berlin Heidelberg.

[49] Sue Mckemmish and Anne Gilliland. 2013. Archival and recordkeeping research: Past, present and future. In *Research Methods: Information, Systems and Contexts*, K. Williamson and G. Johanson (Eds.). Tilde Publishing, Prahran, Victoria, 79–112.

[50] Ian Milligan, Nick Ruest, and Jimmy Lin. 2016. Content Selection and Curation for Web Archiving: The Gatekeepers vs. the Masses. In *JCDL âĂŹ16, June 19 - 23, 2016, Newark, NJ, USA*. ACM, Newark, NJ. DOI: https://doi.org/DOI:http://dx.doi.org/10.1145/2910896.2910913

[51] Gordon Mohr, Michael Stack, Igor Ranitovic, Dan Avery, and Michele Kimpton. 2004. An Introduction to Heritrix: An open source archival quality web crawler. In *Proceedings of the 4th International Web Archiving Workshop*. Bath, UK. https://webarchive.jira.com/wiki/download/attachments/5441/Mohr-et-al-2004.pdf

[52] National Digital Stewardship Alliance. 2012. *Web Archiving Survey Report.* Technical Report. NDSA Content Working Group. http://www.digitalpreservation.gov/documents/ndsa_web_archiving_survey_report_2012.pdf

[53] Richard E. Nisbett and Timothy DeCamp Wilson. 1977. Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review* 84, 3 (May 1977), 231–259.

[54] Jinfang Niu. 2012. An Overview of Web Archiving. *D-Lib Magazine* 18, 3/4 (April 2012). http://www.dlib.org/dlib/march12/niu/03niu1.html

[55] Fatih Oguz and Wallace Koehler. 2015. URL decay at year 20: A research note. *Journal of the Association for Information Science and Technology* 67, 2 (2015), 477–479. DOI: https://doi.org/10.1002/asi.23561

[56] Maureen Pennock. 2013. *Web-Archiving.* Technical Report Technology Watch Report 13:01. Digital Preservation Coalition, Great Britain. 1–50 pages. http://dx.doi.org/10.7207/twr13-01

[57] Margaret E. Phillips. 2005. What Should We Preserve? The Question for Heritage Libraries in a Digital World. *Library Trends* 54, 1 (2005), 57–71. http://muse.jhu.edu/journals/library_trends/v054/54.1phillips.html

[58] Steven Pinch and N Henry. 1999. Discursive Aspects of Technological Innovation: The Case of the British Motor-Sport Industry. *Environment and Planning A* 31, 4 (1999), 665–682. DOI: https://doi.org/10.1068/a310665

[59] P. Prasad. 1997. Systems of Meaning: Ethnography as a Methodology for the Study of Information Technologies. In *Information Systems and Qualitative Research: Proceedings of the IFIP TC8 WG 8.2 International Conference on Information Systems and Qualitative Research, 31st May - 3rd June 1997, Philadelphia, Pennsylvania, USA*, Allen S. Lee, Jonathan Liebenau, and Janice I. DeGross (Eds.). Springer US, Boston, MA, 101–118. http://dx.doi.org/10.1007/978-0-387-35309-8_7

[60] Hany M. SalahEldeen and Michael L. Nelson. 2012. Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost? In *Theory and Practice of Digital Libraries: Second International Conference, TPDL 2012, Paphos, Cyprus, September 23-27, 2012. Proceedings*, Panayiotis Zaphiris, George Buchanan, Edie Rasmussen, and Fernando Loizides (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 125–137. http://dx.doi.org/10.1007/978-3-642-33290-6_14

[61] Steve Schneider and Kirsten Foot. 2008. Archiving of Internet Content. In *The International Encyclopedia of Communication*, Wolfgang Donsbach (Ed.). Wiley Publishing.

[62] Joan M. Schwartz and Terry Cook. 2002. Archives, Records, and Power: The Making of Modern Memory. *Archival Science* 2 (2002), 1–19.

[63] Kalpana Shankar. 2004. Recordkeeping in the Production of Scientific Knowledge: An Ethnographic Study. *Archival Science* 4 (2004), 367–382.

[64] Marc Spaniol, Dimitar Denev, Arturas Mazeika, Gerhard Weikum, and Pierre Senellart. 2009. Data Quality in Web Archiving. In *Proceedings of the 3rd Workshop on Information Credibility on the Web*. ACM, New York, NY, USA, 19–26. DOI: https://doi.org/10.1145/1526993.1526999

[65] James P. Spradley. 1979. *The Ethnographic Interview.* Holt, Rinehart and Winston, United States.

[66] James P. Spradley. 1980. *Participant Observation.* Wadsworth/Thomson Learning, United States.

[67] Susan Leigh Star and Karen Ruhleder. 1996. Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Spaces. *Information Systems Research* 7, 1 (March 1996).

[68] Ann Laura Stoler. 2002. Colonial archives and the arts of governance. *Archival Science* 2, 1-2 (2002), 87–109. http://rd.springer.com/article/10.1007%2FBF02435632

[69] Lucy A. Suchman. 2001. Building Bridges: Practice-based Ethnographies of Contemporary Technology. In *Anthropological Perspectives on Technology*, Michael Schiffer (Ed.). University of New Mexico Press, Albuquerque, 163–177.

[70] Ed Summers and Ricardo Punzalan. 2016. Bots, Seeds and People: Web Archives as Infrastructure. *The Computing Research Repository* abs/1611.02493 (2016). http://arxiv.org/abs/1611.02493

[71] Diana Taylor. 2003. *The Archive and the Repertoire: Performing Cultural Memory in the Americas.* Duke University Press, London.

[72] Mike Thelwall and Liwen Vaughan. 2004. A fair history of the Web? Examining country balance in the Internet Archive. *Library & Information Science Research* 26 (2004), 162–176.

[73] Ciaran B. Trace. 2002. What is recorded is never simply 'what happened': Record keeping in modern organizational culture. *Archival Science* 2, 1 (2002), 137–159. DOI: https://doi.org/10.1007/BF02435634

[74] Michel-Rolph Trouillot. 1995. *Silencing the Past: Power and the Production of History.* Beacon Press, Boston.

[75] Claire Waterton. 2010. Experimenting with the Archive: STS-ers As Analysts and Co-constructors of Databases and Other Archival Forms. *Science, Technology, & Human Values* 35, 5 (2010), 645–676.

[76] Collin Webb, David Pearson, and Paul Koerbin. 2013. 'Oh, you wanted us to preserve that?!' Statements of Preservation Intent for the National Library of Australia's Digital Collections. *D-Lib Magazine* 19, 1/2 (Feb. 2013). http://www.dlib.org/dlib/january13/webb/01webb.print.html

[77] Peter Webster. 2017 (forthcoming). Users, technologies, organisations: towards a cultural history of world web archiving. In *Web 25: histories from the first 25 years of the World Wide Web*, Niels Brügger (Ed.). Peter Lang.

[78] Aaron Ximm. 2014. Active Personal Archiving and the Internet Archive. In *Personal Archiving: Preserving Our Digital Heritage*. Information Today, Medford, US, 187–213.

[79] Elizabeth Yakel. 2001. The Social Construction of Accountability: Radiologists and Their Record-Keeping Practices. *The Information Society* 17, 4 (2001), 233–245. DOI: https://doi.org/10.1080/01972240175333083