



Ogden, J. (2021). "Everything on the Internet can be saved": Archive Team, Tumblr and the cultural significance of web archiving. *Internet Histories*. <https://doi.org/10.1080/24701475.2021.1985835>

Publisher's PDF, also known as Version of record

License (if available):
CC BY

Link to published version (if available):
[10.1080/24701475.2021.1985835](https://doi.org/10.1080/24701475.2021.1985835)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the final published version of the article (version of record). It first appeared online via Taylor and Francis at <https://doi.org/10.1080/24701475.2021.1985835>. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: <http://www.bristol.ac.uk/red/research-policy/pure/user-guides/ebr-terms/>

“Everything on the internet can be saved”: Archive Team, Tumblr and the cultural significance of web archiving

Jessica Ogden

To cite this article: Jessica Ogden (2021): “Everything on the internet can be saved”: Archive Team, Tumblr and the cultural significance of web archiving, Internet Histories, DOI: [10.1080/24701475.2021.1985835](https://doi.org/10.1080/24701475.2021.1985835)

To link to this article: <https://doi.org/10.1080/24701475.2021.1985835>



© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 21 Oct 2021.



Submit your article to this journal [↗](#)



Article views: 1632




View related articles [↗](#)



View Crossmark data [↗](#)

“Everything on the internet can be saved”: Archive Team, Tumblr and the cultural significance of web archiving

Jessica Ogden 

School of Sociology, Politics and International Studies, University of Bristol, Bristol, United Kingdom

ABSTRACT

This article frames the cultural significance of web archiving through an ethnographic study of Archive Team and their efforts to archive “Not Safe for Work” posts on the popular social media platform, Tumblr. This research first sheds light on the origins and organisation of Archive Team, a long-running site of web archiving and “loose collective” of volunteers dedicated to saving websites in danger of going offline. I outline two Archive Team “tenets of practice” that reflect and frame an approach to web archiving centred on cultural values dedicated to the preservation of access. Using examples from their efforts to archive Tumblr NSFW, I examine how the entanglement of practice, participants and platform resistance ultimately shapes what was deemed worth saving (and conversely, not). I argue that web archiving is a transformative force that requires attentiveness to who is archiving, but also the cultural dimensions of practice that inform everyday decisions about how the Web is “saved.”

ARTICLE HISTORY

Received 25 February 2021

Revised 4 September 2021

Accepted 22 September 2021

KEYWORDS

Tumblr;
web archiving;
Archive Team;
NSFW

1. Introduction

Recent trends in communication and Internet studies research show a growing interest in the significance and impact of content moderation practices on the rise and fall of social media platforms. As just one example, Tumblr’s (2018) removal of so-called “adult content” and “Not Safe for Work” (NSFW) posts provides evidence for the effects of content moderation policy changes on the platform’s vitality. Where Tumblr was once home to thriving LGBTQ+ communities, the platform has reportedly since been reduced by “one-fifth” (Cuthbertson, 2019). Whilst a range of scholarship has attended to the effects of content moderation on Tumblr communities specifically, this article turns a critical eye towards the ways that archivists are intervening and shaping access to platform content in the face of its removal.

Since the mid-1990s, web archivists at the Internet Archive (IA), national libraries and archives, university research libraries and laboratories (and beyond) have been

CONTACT Jessica Ogden  jessica.ogden@bristol.ac.uk  School of Sociology, Politics and International Studies, University of Bristol, 11 Priory Road, Bristol BS8 1TU, United Kingdom.

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

preserving parts of the Web in “web archives” (Brügger, 2018). Web archives have been widely positioned as necessary pre-emptive interventions for preserving access to public webpages, social media and other forms of online communication for future use, as well as a key resource for Digital Humanities, social science and historiographical Web/Internet Studies research (Brügger, 2018; Milligan, 2019; Weber, 2020). In addition, a growing body of research has illustrated the importance of critically examining *how* the Web is archived and the related implications for scholarship using web archives. This research has centred methodological approaches that “reverse engineer” or reconstruct practices through interviews with practitioners and forensic studies of web archives (Ben-David & Amram, 2018; Maemura, Worby, Milligan, & Becker, 2018; Milligan, Ruest, & Lin, 2016; Summers & Punzalan, 2017), as well as ethnographic methods that demonstrate the value of observing practice (Ogden, Halford, & Carr, 2017; Summers, 2020b).

Building on this previous research, this article examines the subjective nature of web archiving through the lens of *culture*. Inspired by Kelty’s (2008) work on the cultural significance of free and open source software (F/OSS), I argue that web archiving is situated within particular cultural worlds which advocate for “moral and technical orders” that materially shape how the Web is archived. The argument is rooted in a practice theory approach to culture that grounds cultural processes in the “social relations of people” (Ortner, 2006, p.3) and technologies, in order to render visible the “human infrastructure” or “people, organisations, networks and arrangements” (Lee, Dourish, & Mark, 2006, p.484) that underpin web archiving activities. The culture-as-practice point of view therefore invokes and foregrounds the diverse imaginaries, “strategies of action” and “toolkit(s) of symbols, stories, rituals and worldviews” (Swidler, 1986, p.273) that shapes archiving practices in different cultural, professional and organisational contexts. By foregrounding culture in this way, I illustrate the ways that cultural priorities become a set of situated moral commitments in web archiving which fundamentally shape how the Web will be understood in future.

The article advances the topic and special issue on “death and dying platforms” through two contributions which frame the cultural significance of web archiving. The first illuminates the work of Archive Team (AT), a long-running site of web archiving and self-described “loose collective” of volunteers dedicated to saving websites in danger of going offline. Since forming in 2009, AT have organised efforts to archive such well-known (but now defunct) platforms as GeoCities, Friendster, Vine, Google+ and most recently, Parler. Despite the scale of these operations and the relative fanfare they received in popular media, AT’s activities have yet to receive significant scholarly attention.¹ Taking an ethnographic approach, this research draws on a broader doctoral research project on web archiving practices (Ogden, 2020) to shed light on the origins and organisation of AT and outline two tenets of practice that drive their approach to web archiving.

The second contribution demonstrates the ways that web archiving practices constrain how we will come to know and understand dead and dying platforms in future. Through examples drawn from AT’s efforts to archive Tumblr NSFW, I examine how the entanglement of practice, participants and platform resistance shapes which parts of the site were deemed worth saving (and conversely, *not*). The analysis emphasises the contradictions and dilemmas that arise through the practice of web archiving

and highlights the marked divergence between AT and their counterparts based in conventional memory institutions. I argue that web archiving is a transformative force that requires attentiveness to *who* is archiving, but also the cultural dimensions of practice that inform everyday decisions about *how* the Web is “saved.” In short, this article asks: what happens when web archiving is under the dynamic direction of a self-described group of “rogue archivists, programmers, writers and loudmouths dedicated to saving our digital heritage” (Archive Team, [n.d.-e](#))?

The next section provides background into the Tumblr NSFW case study and events that led to AT’s interventions before describing the methods used to study them. Section 3 discusses the origins of AT and proposes two tenets of practice that guide the collective, before examining how these tenets manifested through archiving Tumblr in Section 4. I conclude by reflecting on the ethical implications of AT’s work and the Tumblr NSFW web archives, both for the NSFW communities themselves and the study of dead and dying platforms.

2. Background and methodology

2.1. Tumblr: a dead and dying platform?

Since its launch in 2007, Tumblr became a widely used micro-blogging platform with an estimated 455 million blogs and over 168 billion posts by December 2018 (*About | Tumblr*, 2019). Tumblr’s relatively permissive content moderation policies have been positioned as one feature that led to the platform’s popularity (Fink & Miller, 2014); creating networked publics for the curation and circulation of sexually explicit images, nude selfies and other forms of “adult content.” The policies and platform affordances supported “a sense of community” and belonging (Tiidenberg, 2014, p.9) amongst users and blogs dedicated to “counterpublics” considered fringe and marginalised elsewhere (Renninger, 2015). Much of this content fell under the “Not Safe for Work” (NSFW) category; an internet colloquialism for links and media that may contain nudity, sexuality, profanity or violence and therefore, not suitable for public or workplace settings. Tumblr encouraged users to self-tag their blogs/posts as NSFW which created a hybrid platform browsing experience that failed to distinguish between (for example) “artistic, casual or pornographic nudity” (Gillespie, 2018, p.175).

Between 2013-2018, in what has widely been interpreted as a response to changes in the US regulatory environment, increased pressures from the Apple iOS app store and a desire to boost ad sales (Gillespie, 2018; Tiidenberg & Nagel, 2020), Tumblr escalated steps to filter and remove adult/NSFW content from view. Using a combination of algorithmic filtering and user-generated tags as training data, Tumblr first blocked NSFW with a site-wide “safe mode” in February 2018 (Koebler & Cole, 2018) (removing NSFW content from public view without a login), before automatically tagging users, blogs and posts for removal in December.

Tumblr framed the December policy change in a post entitled “A better, more positive Tumblr,” announcing that the platform would “no longer [allow] adult content, including explicit sexual content and nudity” and adding that this content would be removed from view later that month (D’Onofrio, 2018). Tiidenberg and Nagel (2020, p.74) describe the “deep sense of betrayal” expressed by NSFW users and content

creators in reaction to the ban, with one claiming the announcement was the moment the platform “died.” Further illustrated by the “RIP Tumblr” memes that trended during the weeks surrounding the ban, these reactions speak to the anger and erosion of trust over the creation of “safe spaces” for sexy content only to subsequently “[evict] communities wholesale” (Tiidenberg & Nagel, 2020, p.74). Given the fourteen days between Tumblr’s announcement and subsequent enforcement, and their limited attempts to provide users with the means to archive their own content (Liao, 2018), the NSFW case calls into question what responsibilities and obligations platforms should have to their communities of users (Gillespie, 2018), as well as the ways users and researchers will come to understand both the processes of content moderation and the platforms themselves in future.

2.2. Methods

The fieldwork and data underpinning this article followed an ethnographic approach, including both participant and non-participant observations online, the use of Internet Relay Chat (IRC) logs and documentary research. As part of a broader study of AT and web archiving, between December 2018 - January 2019, I closely followed AT’s activities as they occurred on public IRC channels on EFNet (#archiveteam, #archiveteam-bs and #tumbledown), AT GitHub repositories,² the social media content of AT founder Jason Scott and AT accounts on Twitter, Tumblr and Reddit, and in the media coverage and journalistic interviews with AT participants (Archive Team, n.d.-d). As will be discussed in Section 3, the social organisation of AT consists of a revolving collective of hundreds of volunteer participants, as well as a number of gatekeepers and an estimated 30-40 core members who maintain source code and infrastructure, manage projects and moderate channels of communication (in addition to web archiving). This analysis focuses in part on IRC communications amongst the mix of newcomers, core AT members and project participants who actively carried out the work of archiving Tumblr NSFW. Ad-hoc IRC chat and interviews with AT core organisers before, during and after the project provided insights and clarifications into decisions made regarding the Tumblr project and the wider practices of AT.

Further context was added to real-time observations through the analysis of information provided on AT’s wiki and historical logs of public IRC channels.³ The IRC logs can be seen as a form of what Kozinets (2010, pp.104-106) describes as “archival net-nographic data,” or archived data of community interactions in online forums, providing insights into historical discussions and practices that surrounded other web archiving projects over time. The combination of interviews, chat logs and documentary sources offers insights into tacit knowledge that underlies practices (Bueger, 2014, p.401), as well as transactionally documents the extent to which certain actors and contributors (that would otherwise be obscured) participate in the shaping of AT’s web archiving practices. As part of the broader study of web archiving, qualitative coding and thematic analysis were used to collate and interpret the relationship between actions, artefacts and tacit knowledge (Bueger, 2014) that make up AT’s web archiving practices.

The project adhered to a two-tiered consent process, where prior permission from core AT organisers was sought before observing and using the historical log data,

and individual informed consent was sought for the use of interview and one-on-one chat data with participants. As receiving informed consent from the hundreds of participants in the live and historical log data was not practical, efforts have been made to protect individual identities of informants through the allocation of additional gender-neutral pseudonyms for the IRC “handles” quoted in this article. This research received ethical approval from the University of Southampton.

3. Archive Team

The AT collective formed in 2009 in response to a series of blog posts by Jason Scott Sadofsky, a computer historian and self-described “free range archivist” more commonly known as Jason Scott. As a longtime collector and advocate for net culture, Scott was incensed by the closure of AOL Hometown, an early web-hosting platform which allowed AOL users to build websites with little-to-no coding expertise. The shuttering and loss of AOL Hometown in 2008, a service that hosted an estimated 14 million websites, was likened to “a mass eviction” (Scott, 2008) that amounted to a loss of people’s “information, hopes, dreams [and] history” (Scott, 2009). The closure spawned a vision for the AT by Scott, partly illustrated in this excerpt:

ARCHIVE TEAM would be like CERT (the Computer Emergency Response Team) used to be, where it was a bunch of disparate people working together to solve a problem in a nimble and networked fashion. They’d find out a site was going down, and they’d get to work (Scott, 2009).

This excerpt provides a working imaginary for AT, as well as symbolic commentary on the collective loss of web history and the perceived erosion of rights for platform users. Scott draws a direct comparison to CERT, a group of computer experts at Carnegie Mellon University that was established in the 1980s to respond to computer security incidents in the US. AT is envisioned to be a similar “emergency response team” that would dedicate their technical skills and hardware to archiving websites for platform users: an “A-Team” in the battle against corporations who fail to provide users with sufficient time and means to save their own content. The vision prioritises a distributed, “nimble” and “networked” team and goes on to foreground a sense of loss and deep mis-trust in hosted web services that is notably similar to reactions from Tumblr NSFW users described above. Scott paints a vision for the AT as “vigilante teams of mad archivists” for the Web, who work to not only archive digital content, but also to “publicize [the] demise” of failing platforms (Scott, 2009).

In the months and years that followed Scott’s posts, AT has since become a semi-stable distributed collective of c. 30-40 core organisers that are supplemented by hundreds (and sometimes thousands) of volunteers. AT participants are based around the world and range from those with professional programming experience to those with a general interest in archiving, computer history and digital preservation. AT convenes and uses a number of communication channels to organise and mobilise participation, including a wiki, IRC, GitHub and social media. AT maintains a “Deathwatch” wiki page that contains a running catalogue of projects and “dead and dying” websites going back to 2001 (Archive Team, n.d.-b), and regularly employs a

range of tools for monitoring and soliciting nominations for sites in need of archiving through the use of IRC, social media, Google alerts and bots that “listen to” Wikipedia categories such as *Deaths* and *Disestablishments*.

In 2011, Scott joined the IA as a “free range archivist” and staff member, establishing an unofficial partnership between AT and the IA Wayback Machine (IAWM).⁴ The dissemination and storage of AT projects is facilitated by this collaboration with the IAWM; one which is manifested through IA staff member participation in AT projects, and a mutually beneficial relationship whereby the IAWM collection coverage is significantly extended by acting as a reliable, long-term repository for AT. Under this arrangement, AT contributed more than 9.3 petabytes to the collection between 2009-2018 (Figure 1).⁵

3.1. Tenets of practice

As a loose “anti-bureaucratic” collective, AT creates organisational norms and shared notions of membership that drive web archiving through the use of satire, distributed working strategies and the transmission and enforcement of their own practice conventions to newcomers. AT simultaneously critiques and variably aligns itself with the broader principles of liberalism, alongside recognisable commitments to some “contemporary ethics and aesthetics of hacking” that pervade hacker communities of many stripes (Coleman, 2013, p.17-19), including: a dedication to information freedom, freedom of expression, the meritocracy of hacking and the potential of computers for making a better world (Levy, 2010, pp.39-46). The following briefly surfaces and proposes two additional “tenets of practice” that build on these ethics and distinguish AT from the web archiving practices and commitments of conventional memory

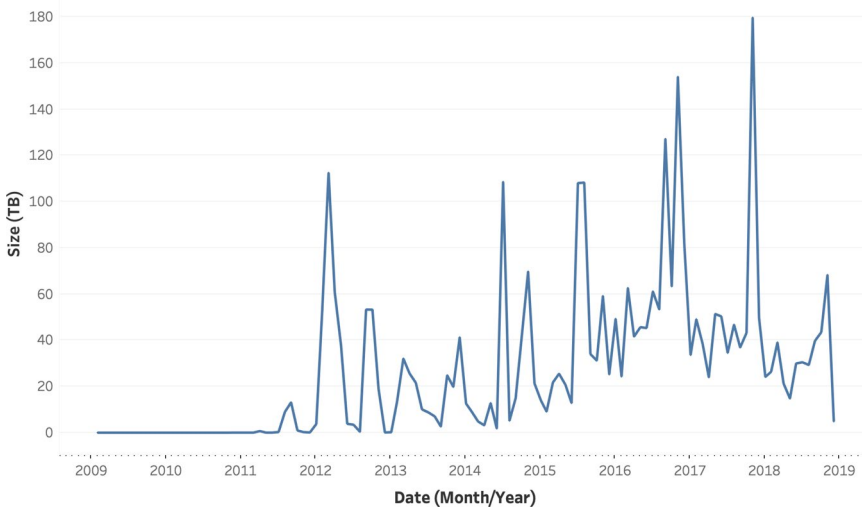


Figure 1. Archive Team collection activities between 2009 and 2018, in terabytes over time. The data was generated by scraping the Archive Team Internet Archive collection metadata (Summers & Ogdén, 2021).

institutions (discussed further below). Each of these tenets will be further explored through the Tumblr NSFW project.

3.1.1. *Everything online is created equal*

Fundamentally, AT frames web archiving as a tool for enabling the preservation of free and open access to online culture and information. In addition, AT espouses to treat all websites with equal priority, while positioning the collective as a non-partisan protector of web content and mobilising a community of practice around notions of history, heritage and the future of the Web. As Phillips (2015) observes in other online community cultures, in this context the “we” of web archiving is generative and similarly creates a sense of identity, belonging and community around the moral imperative of “saving [the Web] for the people” (Scott, 2009) whilst simultaneously attempting to depoliticise the practice of archiving. Although questions surrounding the power and politics of selection have historically preoccupied the archives profession for decades, these observations point to a community that is generally resistant to any discussion of “selectivity.” When I asked about how sites get selected for archiving, drew shared this passage from AT’s Wikipedia page:

According to Jason Scott, ‘Archive Team was started out of anger and a feeling of powerlessness, this feeling that we were letting companies decide for us what was going to survive and what was going to die’. Scott continues, ‘it’s not our job to figure out what’s valuable, to figure out what’s meaningful. We work by three virtues: rage, paranoia and kleptomania’. (*Archive Team* | *Wikipedia*, 2021)

The passage positions AT’s projects in opposition to platforms that shutter their services without warning, emphasising the role of corporate business models and policies in deciding “what was going to survive” of user-generated content. By implication, Scott positions the role of AT through a lens of objectivity that was echoed by AT core members in my observations, championing a general neutrality towards deciding the value of collecting certain websites over others. This particular “moral and technical order” (Kelty, 2008) is extended by drew’s reflection that archives must be representative of the diversity of the online experience so that in future “we have a good view” of “how the world was;” later indicating that “everything on the Internet can be saved.” Each of these observations are rooted in the values of neutrality and objectivity in web archiving, and work to provide context for how AT frames their role in saving the Web.

3.1.2. *Archive first, ask questions later*

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

AT makes it clear that they largely do not ask permission when archiving public websites. The question of seeking permission is frequently raised by newcomers in IRC and other public forum discussions about AT’s activities. “Special Archive Team IRC rules” (quoted in the excerpt above) attempt to reinforce the importance of staying

on-topic in IRC, but also explicitly discourages participants from questioning the premise of AT's archiving pursuits. And yet, the legal and ethical boundaries of web archiving are frequent topics of discussion raised by new joiners. The issue of consent is indicative of a fundamental difference between AT, some conventional (institutional) web archives, "community archives" projects (Flinn, Stevens, & Shepherd, 2009), and practitioners and Internet researchers who have questioned the ethics of crawling in the absence of consent from content creators (Lomborg, 2019). For better or worse, this action-oriented "brute force" approach has enabled AT to proceed where institutions like national web archives are subject to their own mandates and legislative environments that constrain the nature of what can be collected, stored and made accessible.⁶

Scott has argued that AT's "archive first, ask questions later" approach emphasises urgency and action over bureaucracy or philosophical debates about best practices and ultimately operates to "keep the discussion going" (Findlay, 2011). This is best exemplified by AT's disregard for the robots.txt protocol, a strategy outlined in Scott's AT wiki entry, irreverently titled: "Robots.txt is a suicide note" (Scott, 2017). Despite their aims, AT's tactics have both opened and closed the lines of communication between themselves and the platforms they archive, including times when AT inadvertently performs what amounts to a distributed denial of service (DDoS) attack. In one example, DNSHistory posted a permanent banner notification accusing AT of having a "self-righteous attitude" by ignoring robots.txt; describing their tactics as tantamount to "abuse" (*DNS History*, n.d.).

3.2. Rogue web archiving

Whilst there is limited space to fully address the broader field of web archiving in the context of this short article, it is worth emphasising how AT's tenets of practice compare to approaches taken in conventional memory institutions and community-centred archiving projects, more generally.⁷ As detailed elsewhere, web archiving in institutional contexts (e.g. national libraries/archives and university environments) is frequently directed by professional standards of archiving practice and records management, collection mandates and legal constraints (e.g. non-print legal deposit and copyright restrictions), as well as limited staffing and technical resources relative to the scale of the task (Hockx-Yu, 2011). Despite the long history of appraisal and collection development practices in archives, the digital realm has disrupted conventional mechanisms for establishing consent in archival collection where, even in institutional settings, the "modes of acquisition" often prevent (or at least, actively discourage) web archivists from interacting with content owners and even the records themselves (Summers, 2020a). This has created a landscape with very few fully open access web archives, whereby most are strictly bound by their statutory responsibilities, and therefore have very little appetite for risk regarding the boundaries of selection, copyright and individual rights to privacy (Winters, 2019).

In contrast, AT's web archiving projects are indicative of a particular type of community-oriented memory work that has transitioned from being the sole charge of state-based actors or practitioners within the libraries/archives profession, to that which is reliant on what De Kosnik (2016, pp.51-53) calls "techno-volunteerism," or

the dedicated labour of volunteer amateur archivists. AT practices can be seen as a form of “rogue archiving” (De Kosnik, 2016) or web archiving in service of creating digital archives that are openly and freely available online, unrestricted by copyright and founded by non-institutional actors dedicated to persistent publication and long-term preservation. Whereas other types of “community archives” projects emphasise localised custody and control of archives, an ethics of care, and the active participation of community members themselves in documenting their own histories (Flinn et al., 2009); here, “community” foregrounds the ways web archiving *itself* acts as a community building tool, and in turn mobilises different types of cultural priorities in practice. With its emphasis on the preservation of access above all else, in this case, rogue web archiving, AT’s tenets of practice and the Tumblr NSFW project, diverges from community/activist-oriented approaches that centre ethics and the often slow and collaborative work of increasing the accessibility of historically marginalised or minoritised community content (Christen & Anderson, 2019). It is perhaps no coincidence, therefore that AT’s approach to web archiving is most aligned with the strategies employed by the IAWM, who regularly ignores the robots.txt protocol (M. Graham, 2017) and operates an opt-out approach to collection consent.⁸ In general, the IA also takes a liberal approach to copyright/IP and (in addition to their paid staff) regularly engages the volunteer labour of amateur collectors in service of the organisation’s aspirations to “universal access to all knowledge.”

In summary, AT’s two connected tenets of practice frame an approach to web archiving that is informed by the cultural values of a community of practice centred on participatory action and the preservation of access. And though the collective is one that is clearly influenced by parallel values drawn from F/OSS and hacking ethics, my observations point to the ways that AT as a collective distinctly positions themselves and their work as a form of *activist archiving*. From a culture-as-practice (Swidler, 1986) point of view, the tenets of practice form the basis to which AT participants look to deploy strategies for rogue web archiving, demonstrating how these tenets sit in contrast to both institutional and other types of community archives approaches to archiving. Although depoliticising the process of web archiving may be a goal of such strategies, this positioning can be seen as an act of politics in and of itself that is both contradictory and indeed generative of further politics when put into practice. Tensions in each were observed through the Tumblr project, where on the one hand, (inevitable) selection decisions were both observed and challenged by participants, and the “brute force” approach was met by ethical dilemmas and repeated attempts by Tumblr to ban AT from crawling. The following discusses these challenges and their implications for an understanding of AT, the Tumblr platform itself and future web historiographies of Tumblr NSFW using these archives.

4. Transforming Tumblr NSFW

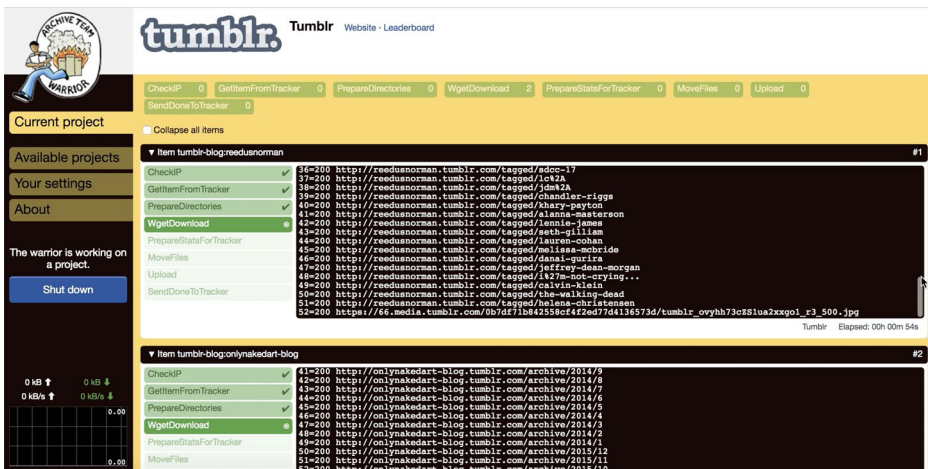
On the 8th of December, 2018, AT launched the Tumblr NSFW project. After a week of testing custom scripts, AT began publicising the project using Twitter, Reddit and the Tumblr platform itself to both “name and shame” Tumblr, as well as invite volunteers to nominate content and get involved in crawling. AT has developed a number of scripts and F/OSS to facilitate distributed, participatory web archiving by volunteers

willing to devote their hardware and cloud infrastructure to crawling the Web. Although methods differ, in general, web archiving relies on a mix of API-based applications (as is often the case in social media archiving) and semi-automated web crawlers that recursively index and download content. Crawls are determined by “seed” instructions (including the target URLs, “depth” and scope of capture); with crawlers writing outputs in the WARC file format, the accepted ISO-standard for web archives.

The *Tracker* and the *Warrior* virtual machine (Figure 2) were both used to archive Tumblr NSFW, creating a distributed resource pipeline between volunteer “workers,” AT staging servers and the IA. The *Warrior* is a desktop application that uses the host to crawl projects centrally managed by AT; standardising the captures and enabling the enrolment of participants with little-to-no technical expertise. The *Tracker* allows AT to control the allocation and rate of items (usernames, subdomains, URLs) given to each *Warrior*/worker instance, but also acts as a live leaderboard that displays each user’s upload rates in realtime. At the height of the Tumblr NSFW project, the *Tracker* recorded an estimated 1,525 handles working to archive the estimated 700,000 NSFW items. Although some proportion of these handles are duplicates of users with multiple machines and concurrent “workers,” the sheer volume is indicative of the power of the Tumblr case to mobilise participation in AT, and representative of a form of crowdsourced web archiving currently unprecedented in the context of institutional web archives. The next Section discusses how AT’s tenets of practice manifested first through the dynamic selection decisions about what to collect, and then in the tactics taken to continue crawling the platform despite Tumblr’s resistance.

4.1. The work of selection

Despite their idealistic stance towards selectivity, AT members are (unavoidably) making selections. Decisions concerning what gets archived are contingent on the time and infrastructure (bandwidth, IPs, workers) available, the value-judgements and priorities of a dynamic combination of stakeholders (AT participants, the IA and “the



The screenshot shows the Warrior application interface. On the left is a sidebar with navigation options: 'ARCHIVE TUMBLR WARRIOR', 'Current project', 'Available projects', 'Your settings', and 'About'. The main area displays a 'Leaderboard' for the 'Tumblr' website. At the top, there are progress indicators for various tasks: CheckIP (0), GetItemFromTracker (0), PrepareDirectories (0), WgetDownload (2), PrepareStatsForTracker (0), MoveFiles (0), and Upload (0). Below this, a list of items is shown, each with a status icon and a URL. The first item is 'tumblr-blog:reedsunorman' and the second is 'tumblr-blog:onlynakedart-blog'. The URLs are listed in a table format with columns for item ID, status, and URL. The status icons include checkmarks, green circles, and red circles. The URLs are mostly from the Tumblr domain, with some from the media subdomain. The interface also shows a 'Shut down' button and a 'Send Done To Tracker' button. At the bottom, there are bandwidth usage indicators for upload and download, both showing 0 KB/s.

Item	Status	URL
36=200	✓	http://reedsunorman.tumblr.com/tagged/edoc-17
37=200	✓	http://reedsunorman.tumblr.com/tagged/1c32a
38=200	✓	http://reedsunorman.tumblr.com/tagged/jdmk2A
39=200	✓	http://reedsunorman.tumblr.com/tagged/ghandier-tiggs
40=200	✓	http://reedsunorman.tumblr.com/tagged/khary-payton
41=200	✓	http://reedsunorman.tumblr.com/tagged/ahimee-materson
42=200	✓	http://reedsunorman.tumblr.com/tagged/iannie-james
43=200	✓	http://reedsunorman.tumblr.com/tagged/seth-gilliam
44=200	✓	http://reedsunorman.tumblr.com/tagged/lauren-cohan
45=200	✓	http://reedsunorman.tumblr.com/tagged/melissa-mcbride
46=200	✓	http://reedsunorman.tumblr.com/tagged/danal-quira
47=200	✓	http://reedsunorman.tumblr.com/tagged/jeffrey-dean-morgan
48=200	✓	http://reedsunorman.tumblr.com/tagged/1477m-not-cycling...
49=200	✓	http://reedsunorman.tumblr.com/tagged/calvin-klein
50=200	✓	http://reedsunorman.tumblr.com/tagged/kbe-walking-jed
51=200	✓	http://reedsunorman.tumblr.com/tagged/helene-christensen
52=200	✓	https://66.media.tumblr.com/0b7df71b842558c74f2ed7d4136573d/tumblr_ovyhh73c281ua2zxp01_r3_500.jpg

Figure 2. The Warrior downloading Tumblr NSFW blogs.

crowd”) and the sociotechnical affordances of the targeted platform. By observing practices, strategies of action are revealed through dilemmas that evidence a prioritisation of “abundance,” various “folk theories” about platform functionality and value claims surrounding what parts of Tumblr to archive and why.

4.1.1. Seeds

Multiple strategies were employed to source (and select) NSFW/adult domains, including the use of pre-existing curated lists and URL nominations solicited from the public using a Google Form. Seeds were captured from other web scraping efforts, including the pushshift.io Reddit comments and submissions API, and the *Majestic Million* SEO index of the top million domains with “the most referring subnets” (*Majestic Million*, n.d.). Seeds were also derived from NSFW community-based curators such as those copied from the Derpibooru *My Little Pony: Friendship is Magic* (MLPFiM) fan community and nominations submitted by community members in IRC.

The seed lists reveal different aspects about AT’s work and the ramifications for the archives they produce. These practices highlight a culture of scavenging and reuse, where curated lists from fan communities like MLPFiM, and other collated data sets are re-purposed for seeding web archives. The seeds also reveal the labour involved in converting these lists into usable datasets for the Warrior infrastructure. AT participants collated, de-duplicated and filtered millions of links in order to avoid wasting time and effort repeatedly capturing the same URLs, but also worked to filter the lists to only include blogs that were tagged as NSFW and “adult.” However, there were risks to this approach, both for representative coverage and for volunteer participants themselves. For example, confusion ensued concerning the difference between adult and NSFW content, raising questions about whether the selection of one category would subsume the other. These deliberations (along with those surrounding “notes,” discussed below) revealed the local, platform-specific knowledge required to disentangle Tumblr’s affordances (and patterns of use) in order to archive it.

Further concerns were also raised by some volunteers about the inadvertent selection of content that contained evidence of child sexual exploitation. Throughout the project a number of unanswered questions were raised in IRC about the legality of passing this type of content through the Warrior application to the IA, including speculation on whether or not the IA has review mechanisms for filtering and removing suspected sites before making them available in the IAWM. Some participants asked questions in relation to their own legal exposure and liability (for example, querying how and when the Warrior application would pass and delete content from their machine), while others focused on the “meta” legal and ethical boundaries between deletion and removal of such content by the IA. One notable exchange occurred, when after spotting a “highly suspect” blog/domain name being archived by their machine, the participant informed the channel they were quitting the project altogether.

4.1.2. Embeds

Beyond seed selection, there was much debate concerning the boundaries of what should be included and excluded throughout the process of archiving NSFW blogs. The project scope was discussed repeatedly on IRC - discussions which were steeped

in an impending (and ever-increasing) sense of urgency as they crept closer to the deadline. One particular discussion regarding blacklisting external image-hosting sites was indicative of concerns that “hot-linked” images (or images embedded in Tumblr posts but hosted elsewhere) were “beyond scope”:

<grayson>are we supposed to insert content from other domains into the [Internet Archive]? that feels funny and beyond scope

<frankie>actually yeah why are we pulling that anyway, that’s not generally at elevated risk of deletion

<hayden>a lot of these image upload sites are at risk of disappearing without warning

<hayden>so if people are embedding images from them I do think it’s a good idea to grab them

<frankie>but *now*, in *this* project?

<@ezra>Well...if it recurses to them why not?

<frankie>anyway does tumblr even let you embed content still? if not roughly 100% of 4chan links should 404

<grayson>sure, but they should be a project in themselves

<grayson>random other sites from years ago an invitation for stalled crawls

<grayson>the crawler only has a single path of execution right now

This discussion about blacklisting continued as the channel grappled with the most efficient ways to code a mechanism for excluding external sites/URLs that were clogging crawls and contributing to slow progress. This exchange illustrates just one instance of the dynamic scoping decisions made by those adjusting the crawlers and highlights the urgency of crawling under the imminent threat of content removals - a situation characteristically referred to as being “on fire.” grayson’s comment alludes to unanswered questions about what (if anything) was agreed with the IA about scoping the Tumblr NSFW collection. Here they are balancing the representative value of grabbing embedded images, the perceived likelihood that image hosting sites like 4chan are ephemeral (and therefore likely unavailable) and the effects this has on crawler speed and completion success rates.

4.1.3. Notes

Overriding concerns about the time needed to complete the crawls before Tumblr’s removal of NSFW sites were ever-present in the IRC. Although the above example about blacklisting implies a priority placed on archiving platform components deemed important by its users (e.g. “if people are embedding images [...] I do think it’s a good idea to grab them”), Tumblr “notes” presents another example of the tensions between the impending deadline and the realtime collective negotiation of what to keep (and why). Seko and Lewis (2018) explains the role of notes on Tumblr:

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

In order to decide whether to archive notes, participants discussed their form and function on Tumblr (which given Tumblr’s unique system of re-blogging, was not readily apparent to all involved in archiving the platform). Over the course of several days, participants made the case for different options: to continue archiving notes in full (despite the time constraints), to partially archive notes (e.g. only the first 50 on any post) or to fully exclude them from the archive (with ezra characterising the exclusion as a necessary “sacrifice”). Ultimately, in a move to speed up the captures and in the absence of definitive technical paths for conducting a partial capture of notes, several participants announced what they deemed “The Decision™”: they would exclude notes from the crawls. Despite “The Decision,” several days later participants were still discussing the value of notes as a proxy for understanding the social dynamics of Tumblr NSFW:

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

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

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

<kyle> agree with jamie

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

Here participants are balancing the desire to collect more (a trait I call “abundance”) with the value of collecting particular components of the platform. The continued discussion reveals how AT participants both contested “The Decision” and framed the value of the collection in relation to a desire to produce a complete and representative record of “the social network.” This desire is also steeped in opposition to Tumblr’s removal of NSFW; an act of “destruction” that comes at the detriment of a future understanding of this community if viewed via web archives that only partially capture the experience of Tumblr. In the context of the NSFW crawls, this highlights how the value of abundance works in tension with a second set of values related to traditional archival notions of “completeness” and “integrity,” and their implications for the future study of Tumblr.

This example also works to illustrate AT concerns that slow progress would turn away possible volunteers (who would rather see the Tracker speeding along). grayson reflected that it was “frustrating to see slow crawls” with an added concern that “people might not come back.” Over the course of the Tumblr grab, participants

periodically tested the limits of crawler speeds, in terms of bandwidth capacity and the potential likelihood Tumblr would rate-limit particular user-agents if used at scale. As one participant reflected, the slow decay of many dying sites targeted by AT (neglected by time and infrastructural resources) often makes them incapable of supporting the bandwidth required to enable simultaneous large-scale access. In this case, however, progress was not impeded by the slow decay of Tumblr's server infrastructure, but rather, what was interpreted by AT as concerted efforts (on the part of Tumblr) to block them from archiving NSFW.

4.2. Circumventing the Ban(s)

AT's archiving efforts were met by considerable resistance from Tumblr over the course of the fourteen-day project. There was widespread speculation about why Tumblr was actively resisting efforts to archive NSFW. Without direct explanation from Tumblr, and despite theories the bans were only triggered automatically due to excessive access rates (Cole, 2018), AT identified a mix of manual interventions that included permanent IP/subnet bans for all volunteers using official scripts, throttling and rate-limiting of common user-agents and evidence that "everyone got banned, regardless of how much or little they had crawled, and regardless of whether they used a residential or a datacenter IP" (Archive Team, n.d.c).

The continuous stalling led AT volunteers with the requisite technical skills to experiment with and implement work-arounds in order to keep crawling. The stalls and flurry of disparate responses to the repeated bans led to much confusion, with one AT newcomer reflecting that the project reminded them of a Defcon "capture the flag" (CTF) contest: "just as paranoid, just as disorganised." The comparison to CTF, a hacking competition where teams test their cybersecurity skills attacking and defending each other's networks, is apt. It draws attention not just to the semi-chaotic atmosphere of the NSFW crawl, but also to how the circumstances of the ban appealed to AT hacker inclinations for problem-solving and "craftiness" in pursuit of cleverly outwitting the technical constraints of crawling Tumblr before the takedown.⁹

It is a regular occurrence for AT to receive pushback from the sites it archives. AT typically uses a bot that is identified through the archiveteam user-agent which makes web masters aware of the origins of increased access requests to their servers. As drew explained, when AT is banned "what we usually do is throw more IP addresses at them [and] use more common user-agents so it's a little harder to automatically ban." Breakdowns in crawling forced repeated discussions around whether AT permitted the use of different (more common) user-agents and login cookies during crawling. Logins or the use of cookies that enabled crawler bots to mimic "real users" became a topic of extensive debate that revealed an ethical dilemma for AT members who attempted to balance a desire for "abundance" with ensuring the integrity of captures. For newcomers like grayson, it became unclear whether or not the question and use of login cookies was in fact a *technical* issue or one of *policy*:

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

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

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

[...]

<austin>Using login cookies apparently crosses a line

<austin>Then we're acting like people

<austin>And it gets into the WARCs

Only through further questioning ezra was I able to understand some of the reasons underpinning the debate about logins. Here, the discussion of logins reflects a previously undocumented AT "policy":

With logins/cookies we are sort of contaminating the resulting WARC with data that is not really meant to be there. [...] with private profiles we risk grabbing data that is not meant for public consumption, such as a users email address, or an users mobile phone number (that would be disastrous) [...] However if a site was previously publicly viewable, and they now have a login wall, we may circumvent that by using a login, but we don't like doing it as we are linking an account to what is supposed to be anonymous data. (ezra)

ezra outlines the reasons why AT tends to steer clear of using logins, namely to avoid the risk of collecting components of websites that are typically only viewable through the use of access controls. When by-passing these access controls, AT risks inserting information into the WARCs that was once only viewable behind a login. But in the case of Tumblr, (and after much debate) AT decided the risks of using login/cookies outweighed the prospect of having to stop collecting NSFW in the face of perpetual banning. After using and abandoning ad-hoc "burner accounts" (that quickly led to cookies getting "burned" when used at scale), a "cookie factory" was eventually constructed to support the dynamic generation of cookies that bypassed Tumblr's "safe mode" and EU GDPR consent forms.

The choice to proceed with logins reveals a hierarchy of priorities for AT participants and several observations about community practice and culture. First, the login saga makes visible a kind of "cultural politics of hacking" (Coleman, 2013, p.18), where despite a "policy" against using logins, the fixation on continued crawling (in pursuit of "abundance") ultimately trumped issues of archival integrity and the risks imposed by breaching Tumblr's access restrictions. When faced with technical roadblocks, AT consistently turned to hacking alternative approaches to keep the project going. The login saga also highlights the contribution of culture "in sustaining existing strategies of action and its role in constructing new ones" (Swidler, 1986, p.278). The example highlights how AT relies on the socialisation and transmission of practice conventions by core organisers to new participants, as well as the ways practice is adapted during moments of rupture or disagreement - ethical processes that (Coleman, 2013, p.124-125) calls "enculturation" and "punctuated crisis," respectively. The Tumblr case highlights a mutable AT collective that (though committed to the tenets of practice) is both open to modification and willing to negotiate the often fuzzy boundaries that define what constitutes the "public Web" in web archiving.

5. Conclusions

In the end, according to the Tracker, AT archived c. 355,000 of the estimated 700,000 NSFW items before Tumblr removed platform access. AT quickly concocted a follow-up project to capture cached media from the content delivery networks (CDNs) - but inevitably NSFW's "Day of Death" came on the 17th of December, 2018. While AT continued to work to circumvent the ban and capture, collate and deposit archives in the IA, NSFW Tumblr was declared officially dead in copious "RIP Tumblr" memes by users and tweets by Scott.

AT and the Tumblr NSFW case study reflect the dynamic and performative ways that web archiving is both shaped and sustained by the cultural worlds from which they stem. This research provides a window into how AT uses web archiving as a tool for agitating for a particular "moral and technical order" (Kelty, 2008) for the Web; one which is shaped by a "web imaginary" that is persistent, free and open access. Whilst the tenets of practice point to AT's overt desire to de-politicise their work, the cultural politics of web archiving were revealed through emergent practice dilemmas and collective negotiations over the selection of sites, media and platform components, as well as issues surrounding the breach of Tumblr's access restrictions. The juxtaposition between the values of abundance, completeness and integrity revealed some of the contingencies of web archiving; where decisions, time constraints and platform resistance all had consequences for what was saved.

In general, collective discussions about individual stances on the project were short-lived and trended towards positive consensus that archiving NSFW was the right thing to do in the face of Tumblr's planned removals - a sentiment neatly captured by a comment on a DataHoarder subreddit post that AT was "doing God's work" (SaltSnorter, 2018). Beyond many quips about archiving porn, there were in fact very few conversations that engaged with the importance of saving NSFW content, *specifically*. Participants framed the imminent removals as both an act of censorship and corporate "suicide to remove everything that/could/be porn," making references to Tumblr's (spectacularly poor) attempts to algorithmically flag adult content using image recognition technologies (Matsakis, 2018). Some were worried that this, in combination with insufficient time for downloading and/or appealing incorrectly flagged blogs, would lead to communities losing access to posts (whether it was in fact NSFW or not). These observations can be seen as in keeping with AT's moral commitment to preserving access and combating "the feeling of powerlessness" in the face of changes to access, but they do not necessarily illuminate the situated ethical arguments *for* (or indeed, as some would argue, *against*) archiving NSFW content.

This raises further issues about the implications and potential disconnects between the work of web archiving, the (now "evicted") Tumblr NSFW communities of use and potential future historiographies of the Tumblr platform, leading to the question: who are these web archives for? One possible response is that this "digital afterlife" serves some purpose for Tumblr NSFW communities themselves; a view that is supported by AT's framing of their own efforts to "save it for the people," as well as evidence that NSFW users and advocates participated in archiving (e.g. through the submission of seed nominations). However, as identified by the field of critical archive studies

and through increased attention in the field of web archiving (P. M. Graham, 2017; Lomborg, 2019), greater care is needed to contextualise the ways that web archiving interacts with the situated (and often, carefully crafted) ethics and communities of use that online platforms afford. Whereas the article focused on AT's tenets and practice, for Tumblr NSFW, additional work is needed to examine how these tenets and the NSFW web archives may in fact be in conflict with the diverse cultural values and ethical commitments of the communities they aimed to help.

In conclusion, this research has highlighted the ways that AT is actively shaping access to dead and dying platforms, as well as creating a community of practice centred on the preservation of access and "rogue archiving" strategies for saving the Web. Despite their use of common tools and standards, these practices will be seen in stark contrast to other risk-averse approaches to web archiving taken by conventional memory institutions, and community archives projects that centre an ethics of care for content creators and future users. AT's interventions simultaneously illustrate the possibilities of participatory web archiving at scale and the potential risks of such approaches in the face of platform resistance, rights and privacy concerns. Given the scale of AT's collecting activities and their impact on the coverage of the IAWM, understanding their practices offers insights into how the Web is transformed through web archiving, as well as their critical ethical implications for how these platforms are studied in future.

Notes

1. Existing work on AT has been constrained to brief accounts of their role in archiving Vine (Summers & Wickner, 2019), GeoCities (Milligan, 2017), and Webster's (2017) inclusion of AT in a "cultural history of web archiving." In addition, De Kosnik (2016) frames AT as a form of "rogue archiving," which I examine further in Section 3.
2. <https://github.com/ArchiveTeam>
3. <http://wiki.archiveteam.org>; Logs for EFNet IRC channels #archiveteam and #archiveteam-bs (amongst others) are hosted at https://archive.fart.website/bin/irclogger_logs, providing a searchable interface for publicly-logged channels and chat. Soon after the data collection for this research in 2018/2019, AT moved to the hackint IRC network and started logging their IRC channels elsewhere.
4. <http://web.archive.org>
5. <https://archive.org/details/archiveteam>
6. See Winters (2019) for discussion of how legal deposit constrains both collection and access in the national web archive context.
7. See Webster (2019) for a recent overview of the field of web archiving.
8. For further discussion of web archiving at the IA, see Ogden et al. (2017).
9. See Coleman (2013) for detailed account of the "craftiness of hackers."

Acknowledgments

I am grateful to various colleagues who have informed and shaped this article through their feedback and support, including: Susan Halford, Ed Summers, Shawn Walker, the Special Issue editors and four anonymous reviewers for their extensive engagement with the article and constructive feedback. I would also like to thank key informants and participants from Archive Team, without which this research would not be possible. Any errors of interpretation are my own.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by the UK Engineering and Physical Sciences Research Council and the Web Science Centre for Doctoral Training, Grant No. EP/G036926/1. Further support was provided by the UK Economic and Social Research Council under *The Social Life of Web Archives* project, Grant No. ES/VO12177/1.

Notes on contributor

Dr Jessica Ogden is a Senior Research Associate (ESRC Postdoctoral Fellow) in the School of Sociology, Politics and International Studies at the University of Bristol, and a Fellow at the Bristol Digital Futures Institute. Jessica received a PhD in Web Science from the University of Southampton. Her research focuses on digital culture, the politics of data/archives, web archiving and their implications for digital scholarship. She is Principal Investigator on the UK Economic and Social Research Council grant-funded project *The Social Life of Web Archives* looking at the broader impact of web archives online.

ORCID

Jessica Ogden  <http://orcid.org/0000-0003-4696-7340>

References

- About | Tumblr. (2019). <http://web.archive.org/web/20190112191359/https://www.tumblr.com/about>
- Archive Team | Wikipedia. (2021). | *Wikipedia*. (January). https://en.wikipedia.org/w/index.php?title=Archive_Team&oldid=999937182
- Archive Team. (n.d.-a). *Archiveteam:IRC* [Wiki]. https://wiki.archiveteam.org/index.php?title=Archiveteam:IRC#Special_ArchiveTeam_IRC_rules
- Archive Team. (n.d.-b). *Deathwatch* [Wiki]. <https://wiki.archiveteam.org/index.php/Deathwatch>
- Archive Team. (n.d.-c). *Frequently Asked Questions — tumblr-grab*. Archive Team. Retrieved July 29, 2016, from <https://github.com/ArchiveTeam/tumblr-grab/blob/c3b723ceec7476ab8f760afec-7c029ae4b90c9c6/FAQ.md>
- Archive Team. (n.d.-d). *In The Media* [Wiki]. https://wiki.archiveteam.org/index.php/In_The_Media
- Archive Team. (n.d.-e). *Main Page* [Wiki]. http://www.archiveteam.org/index.php?title=Main_Page
- Ben-David, A., & Amram, A. (2018). The Internet Archive and the socio-technical construction of historical facts. *Internet Histories*, 2(1–2), 179–201. <https://doi.org/10.1080/24701475.2018.1455412>
- Brügger, N. (2018). *The archived web: Doing history in the digital age*. MIT Press.
- Bueger, C. (2014). Pathways to practice: praxiography and international politics. *European Political Science Review*, 6(3), 383–406. <https://doi.org/10.1017/S1755773913000167>
- Christen, K., & Anderson, J. (2019). Toward slow archives. *Archival Science*, 19(2), 87–116. <https://doi.org/10.1007/s10502-019-09307-x>
- Cole, S. (2018). Archivists say Tumblr IP banned them for trying to preserve adult content. Motherboard. <https://www.vice.com/en/article/d3bekm/archivists-say-tumblr-ip-banned-them-for-trying-to-preserve-adult-content>
- Coleman, E. G. (2013). *Coding freedom: The ethics and aesthetics of hacking*. Princeton University Press.

- Cuthbertson, A. (2019). Tumblr has lost 20 per cent of traffic since its porn ban. *The Independent*. <https://www.independent.co.uk/life-style/gadgets-and-tech/news/tumblr-porn-ban-nsfw-verizon-yahoo-adult-content-a8817546.html>
- D'Onofrio, J. (2018). *A better, more positive Tumblr*. <https://staff.tumblr.com/post/180758987165/a-better-more-positive-tumblr>
- De Kosnik, A. (2016). *Rogue archives: Digital cultural memory and media fandom*. MIT Press.
- DNS History. (n.d.). <https://dnshistory.org/>
- Findlay, C. (2011). *Where do old websites go to die? with Jason Scott of Archive Team - Podcast*. <https://rkroundtable.org/2011/06/25/where-do-old-websites-go-to-die-with-jason-scott-of-archive-team-podcast/>
- Fink, M., & Miller, Q. (2014). Trans media moments: Tumblr, 2011–2013. *Television & New Media*, 15(7), 611–626. <https://doi.org/10.1177/1527476413505002>
- 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. <https://doi.org/10.1007/s10502-009-9105-2>
- Gillespie, T. (2018). *Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
- Graham, P. M. (2017). Guest editorial: Reflections on the ethics of web archiving. *Journal of Archival Organization*, 14(3–4), 103–110. <https://doi.org/10.1080/15332748.2018.1517589>
- Graham, M. (2017). *Robots.txt meant for search engines don't work well for web archives* [Blog]. <https://blog.archive.org/2017/04/17/robots-txt-meant-for-search-engines-dont-work-well-for-web-archives/>
- Hockx-Yu, H. (2011). *The past issue of the web* [Paper presentation]. Proceedings of the 3rd International Web Science Conference (pp. 1–8). New York, NY, Association for Computing Machinery, June. <https://doi.org/10.1145/2527031.2527050>
- Kelty, C. M. (2008). *Two bits: The cultural significance of free software*. Duke University Press.
- Koebler, J., & Cole, S. (2018). *Apple sucked Tumblr into its walled garden, where sex is bad*. https://motherboard.vice.com/en_us/article/a3mjxg/apple-tumblr-porn-nsfw-adult-content-banned
- Kozinets, R. V. (2010). *Netnography: Doing ethnographic research online*. SAGE.
- Lee, C. P., Dourish, P., & Mark, G. (2006). *The human infrastructure of cyberinfrastructure* [Paper presentation]. Proceedings of the 2006 Conference on Computer Supported Cooperative Work (pp. 483–492). New York, NY, ACM. <https://doi.org/10.1145/1180875.1180950>
- Levy, S. (2010). *Hackers: Heroes of the computer revolution*. O'Reilly.
- Liao, S. (2018). *How to back up your Tumblr before the porn ban* [Magazine]. <https://www.theverge.com/2018/12/7/18126312/how-to-save-tumblr-porn-ban-export-tool-backup>
- Lomborg, S. (2019). Ethical considerations for web archives and web history research. In N. Brügger & I. Milligan (Eds.), *The SAGE handbook of web history* (pp. 99–111). SAGE Publications Ltd.
- 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://doi.org/10.1002/asi.24048>
- Majestic Million. (n.d.). <https://majestic.com/reports/majestic-million>
- Matsakis, L. (2018). *Tumblr's porn-detecting AI has one job and it's bad at it*. <https://www.wired.com/story/tumblr-porn-ai-adult-content/>
- Milligan, I. (2017). Welcome to the web: The online community of GeoCities during the early years of the World Wide Web. In N. Brügger & R. Schroeder (Eds.), *The web as history: Using web archives to understand the past and present* (pp. 137–158). UCL Press.
- Milligan, I. (2019). *History in the age of abundance?* McGill-Queen's University Press.
- Milligan, I., Ruest, N., & Lin, J. (2016). Content selection and curation for web archiving: The gatekeepers vs. the masses. In *JCDL '16, June 19 - 23, 2016*. ACM.
- Ogden, J. (2020). *Saving the Web: Facets of web archiving in everyday practice* [PhD Thesis]. University of Southampton. Retrieved from <http://eprints.soton.ac.uk/id/eprint/447624>
- Ogden, J., Halford, S., & Carr, L. (2017). *Observing web archives: The case for an ethnographic study of web archiving* [Paper presentation]. *Proceedings of WebSci17*, Troy, NY, June 25–28, 2017 (pp. 299–308). ACM. <https://doi.org/10.1145/3091478.3091506>

- Ortner, S. B. (2006). *Anthropology and social theory: Culture, power, and the acting subject*. Duke University Press.
- Phillips, W. (2015). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. MIT Press.
- Renninger, B. J. (2015). Where I can be myself...where I can speak my mind": Networked counterpublics in a polymedia environment. *New Media & Society*, 17(9), 1513–1529. <https://doi.org/10.1177/1461444814530095>
- SaltSnorter. (2018). You're *doing God's work*. Don't let anyone tell you different. [Comment on the online forum post: Archive Team is currently rushing to download as much of Tumblr as possible before Monday and then upload it to the Internet Archive. Any help would be appreciated.]. https://www.reddit.com/r/DataHoarder/comments/a6j7rk/archive_team_is_currently_rushing_to_download_as/
- Scott, J. (2008). *Eviction, or the Coming Datapocalypse* [Blog]. <http://ascii.textfiles.com/archives/1617>
- Scott, J. (2009). *Datapocalypse!* [Blog]. <http://ascii.textfiles.com/archives/1649>
- Scott, J. (2017). *Robots.txt is a suicide note* [Wiki]. <https://www.archiveteam.org/index.php?title=Robots.txt>
- Seko, Y., & Lewis, S. P. (2018). The selfharmed, visualized, and reblogged: Remaking of self-injury narratives on Tumblr. *New Media & Society*, 20(1), 180–198.
- Summers, E. (2020a). Appraisal talk in web archives. *Archivaria* 89 (May), 70–103. Retrieved from <https://archivaria.ca/index.php/archivaria/article/view/13733>
- Summers, E. (2020b). Legibility machines: Archival appraisal and the genealogies of use [PhD], University of Maryland. <https://doi.org/10.13016/u95c-qayr>
- Summers, E., & Ogden, J. (2021). *ArchiveTeam Ingest Rate* (v1.0.0). [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.5548129>
- Summers, E., & Punzalan, R. (2017). *Bots, seeds and people: Web archives as infrastructure* [Paper presentation]. Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 821–834). New York, NY, ACM. <http://doi.acm.org/10.1145/2998181.2998345>
- Summers, E., & Wickner, A. (2019). Archival Circulation on the Web: The Vine-Tweets Dataset. *Journal of Cultural Analytics*, 4(2), 1–14.
- Swidler, A. (1986). Culture in action: Symbols and strategies. *American Sociological Review*, 51(2), 273–286. <https://doi.org/10.2307/2095521>
- Tiidenberg, K. (2014). There's no limit to your love - scripting the polyamorous self. *Journal für Psychologie*, 22(1), 1–27.
- Tiidenberg, K., & Nagel, E. v d. (2020). *Sex and social media*. Emerald Publishing.
- Weber, M. S. (2020). Web archives: A critical method for the future of digital research. WARCNet Papers, 17. https://cc.au.dk/fileadmin/user_upload/WARCnet/Weber_Web_Archives_A_Critical_Method.pdf
- Webster, P. (2017). Users, technologies, organisations: Towards a cultural history of world web archiving. In N. Brügger (Ed.), *Web 25: Histories from the first 25 years of the world wide web* (pp. 170–190). Peter Lang.
- Webster, P. (2019). Existing web archives. In N. Brügger & I. Milligan (Eds.), *The SAGE handbook of web history* (pp. 30–41). SAGE Publications Ltd.
- Winters, J. (2019). Giving with one hand, taking with the other: e-legal deposit, web archives and researcher access. In P. Gooding & M. Terras (Eds.), *Electronic legal deposit: Shaping the library collections of the future*. Facet Publishing.