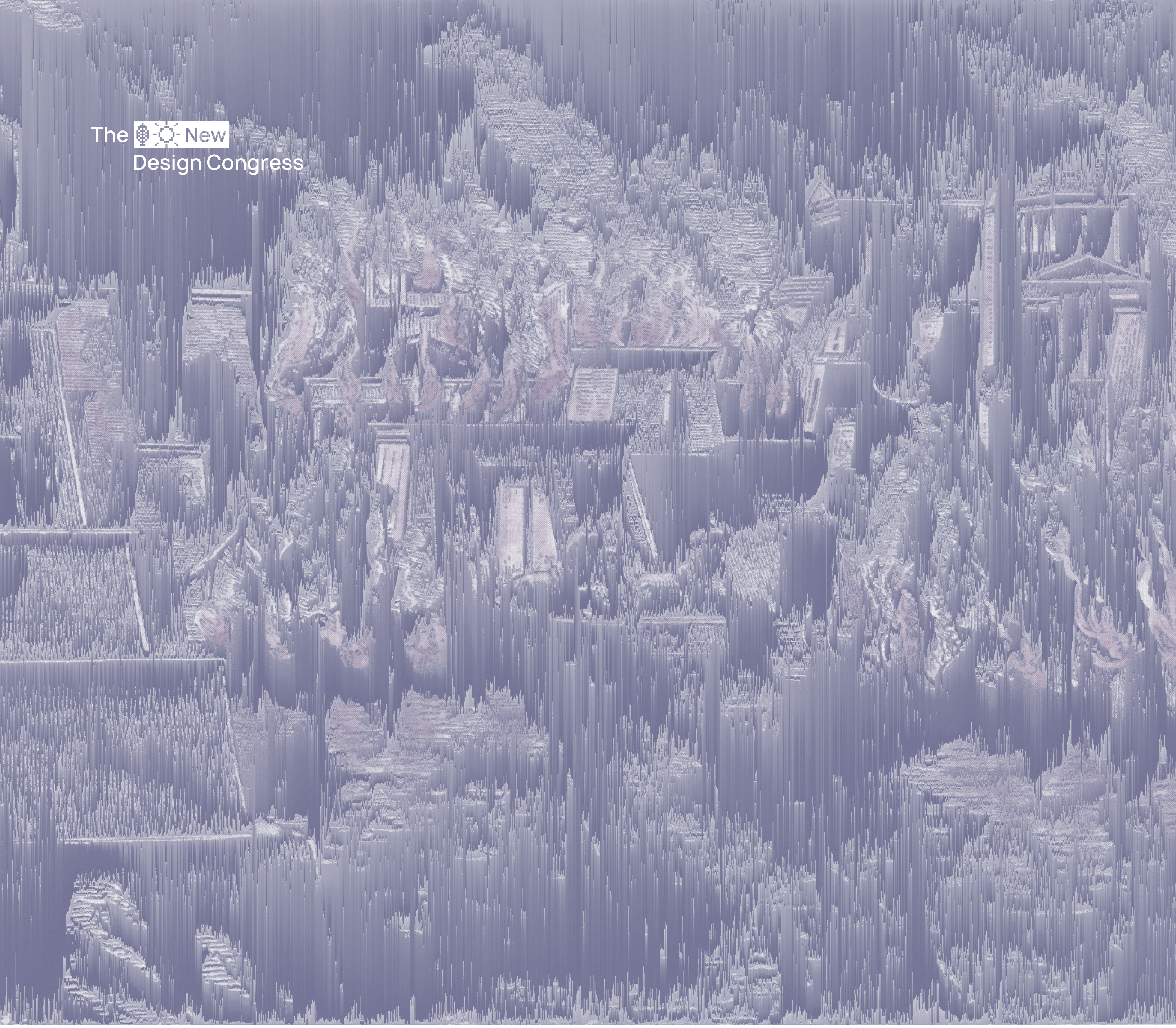
The New Design Congress



MEMORY IN UNCERTAINTY

WEB PRESERVATION IN THE POLYCRISIS

A NEW DESIGN CONGRESS REPORT ✳ NOVEMBER 2022

The New
Design Congress

First edition published 22 November 2022.

Produced in collaboration with Webrecorder, with support from the Filecoin Foundation and Superbloom. Please see the Acknowledgements section of this publication for additional credits and funding declarations.

Also available online at https://newdesigncongress.org

Table of Contents

Executive summary

This research evaluates the design of web archival tools and the broader social and political contexts of web archiving tools and practice, both from the systemic realities of web archiving as a practice, and through the context of a specific emergent tool, the Webrecorder open-source project. Through these lenses, this research uncovers threats to the growth and resilience of web preservation tools and long term data storage, driven by systemic factors — inadequate risk assessments by funders, deteriorating geopolitical and ecological conditions, and a lack of collaborative research and knowledge sharing within the practice of web archiving and its related fields.

Digital and web archiving is the practice of curating, collecting, storing and preserving large collections of material on computer systems and networks. Digital archives can be assembled through the production of a digital replica of a real-world object via photography, scanning or other digitisation processes. Web archives are developed partially or entirely from digital objects, such as websites, historical software programs and other products. Within the bounds of the practice's own definition, digital and web archives are created to preserve historical objects,[1] develop cultural repositories, or maintain social history for items deemed to be of public interest. This is only part of the broader range of applications of digital and web archive practices, and while in theory these two disciplines exhibit their own idiosyncrasies, in practice their frontier remain porous, digital archives employing web archive techniques and vice-versa. This internal friction within the broader field of archiving was highlighted by Trevor Owens, Head of Digital Content Management at the Library of Congress, back in 2014:

> *"One of the tricks to working in an interdisciplinary field like digital preservation is that all too often we can be using the same terms but not actually talking about the same things. In my opinion, the most fraught term in digital preservation discussions is 'archive.' At this point, it has come to mean a lot of different things in different contexts."*[2]

[1] Claunch, Kristina. 'Research Guides: Archive Discovery: A How-To Guide: What Are Archives & Digital Archives?' Accessed 12 November 2022. https://shsulibraryguides.org/c.php?g=86819&p=558330.

[2] Owens, Trevor. 'What Do You Mean by Archive? Genres of Usage for Digital Preservers'. Webpage. The Signal, 27 February 2014. https://blogs.loc.gov/thesignal/2014/02/what-do-you-mean-by-archive-

This research finds that the field of web archiving and its landscape of tools and institutions are out of step with the realities of rising instability and complexity of the 21st century. As societal digitisation accelerates, so too has the belligerence of state and corporate power, the democratisation and intensity of targeted harassment, and the collapse of consent by communities plagued by ongoing (and often unwanted) datafication. Drawing from the climate of the 2020s and informed by an extensive landscape review, as well as a series of qualitative interviews with digital and web archive practitioners conducted between December 2021 and May 2022, this research identifies key intersectional systemic challenges to the fields of digital and web archiving. Many of these issues threaten the quality and resilience of web archives and the integrity of digitally-archived material. They also have profound implications to the physical safety, legal standing and mental health of archive practitioners and communities subjected to archiving. This research identifies and documents issues of ethics, consent, digital security, colonialism, resilience, custodianship and tool complexity.

Despite the challenges identified in the research, there exist compelling reasons to remain optimistic. Funders and projects committed to addressing web and digital archiving in the polycrisis can explore emergent technologies, stronger socio-technical literacy amongst archivists and critical interventions in the colonial structures of digital systems as immediate points of intervention. By acknowledging the shortcomings of cybernetics, resisting the desire to apply software solutionism at scale, and developing a nuanced and informed understanding of the realities of archiving in digitised societies, a broad surface of opportunities can emerge to develop resilient, considered, safe and context-sensitive archival technologies and practice for our uncertain world. ✳

genres-of-usage-for-digital-preservers/.

I. Preserving history in the violent present for an uncertain future

How do we save the past in a violent present for an uncertain future? As the new decade lurches forward, general-purpose networked computing faces significant socio-technical and political trials. Long gone are the days of technological optimism; the benefits and opportunities afforded to societies by digital infrastructure and digital culture are marred by serious socio-technical flaws that make the systems we rely upon fragile and ethically compromised. From the subtle psychological influences of digital interfaces that compress our cognitive potential, to the hidden power structures baked into opaque infrastructures, the challenges posed by the digitisation of society are systemic and complex.

2022 has been a remarkable demonstration of how brittle digitised societies have turned out to be. There are countless demonstrations of this fragility, but to list a few examples:

- In Ukraine, civilians collaborate with the Department of Digital Transformation, using smartphone apps and LTE network towers to crowd-source precise location of Russian ground forces for State air-strikes,[3] while the provision of the Starlink satellite system comes at the cost of unexpected outages over funding issues[4] and the geopolitical whims of its CEO;[5]

- Warring countries suffering enormous and disruptive network attacks that bring down civil services, or result in massive data leaks. Russian sanctions remain brutally effective in unexpected ways, such as the withdrawal of Apple Wallet integration in the Moscow Metro ticketing system leaving commuters unable to legally board public transport.[6] In

[3] Olejnik, Lukasz. ‘Smartphones Blur the Line Between Civilian and Combatant’. *Wired*. Accessed 12 November 2022. https://www.wired.com/story/smartphones-ukraine-civilian-combatant/.

[4] Lyngaas, Sean, and Alex Marquardt. ‘Ukraine suffered a comms outage when 1,300 SpaceX satellite units went offline over funding issues’. *CNN*. Accessed 13 November 2022. https://edition.cnn.com/2022/11/04/politics/spacex-ukraine-elon-musk-starlink-internet-outage/index.html.

[5] France, Anthony. ‘Diplomat’s four-letter blast at Elon Musk over his Russia peace deal’. *Evening Standard.* Accessed 13 November 2022. https://www.standard.co.uk/news/world/elon-musk-twitter-russia-peace-deal-ukraine-andrij-melnyk-b1029923.html.

[6] Sherfinski, D. 2022, March 4. ‘Ukraine crisis highlights Big Tech’s potential to disrupt daily life’. *Reuters*. https://www.reuters.com/legal/litigation/ukraine-crisis-highlights-big-techs-potential-disrupt-daily-life-2022-03-04/.

other situations, ransomware attacks now paralyse entire school districts[7] and healthcare systems;[8]

- In Afghanistan, biometric databases myopically set up by the US occupation forces to prevent aid or financial fraud were seized by the Taliban and will be used to track political opponents;[9]

- In Europe and North America, instability around energy infrastructure threatens rolling blackouts and non-functioning public or corporate services;

- In the United States, ongoing deterioration of the political order leading to the prosecution of teenagers seeking abortion, aided by messages turned over from social media to oppressive law enforcement, as well as the very tangible fear that menstrual cycle apps will be used to track and criminalise sexuality and bodily autonomy in a post-Roe vs. Wade era;[10]

- Internationally and online, the continuing rise of substantial international harassment campaigns, often targeted at public or political figures from under-represented minority demographics.

Each of these examples represent a local or global flashpoint facilitated by the design decisions of a complex, interlocking software and hardware stack now under pressure from external forces. Despite an increased awareness of the manifestation of digital infrastructure, the attitudes of tool builders, policy makers, and infrastructure designers have not kept pace. Many of the unintended consequences of digitisation are the result of weaponised design,[11] a process in which a system or interface harms users while behaving exactly as intended. As

[7] Klein, Alyson. 'Why the Los Angeles Cyberattack Is a Wake-Up Call for Every School District'. *Education Week*, 6 September 2022, sec. Privacy & Security. https://www.edweek.org/technology/why-the-los-angeles-cyberattack-is-a-wake-up-call-for-every-school-district/2022/09.

[8] Drees, Jackie. 'California Clinic to Close after Ransomware Wipes out Patient Records'. Becker's Hospital Review, Accessed 12 November 2022. https://www.beckershospitalreview.com/cybersecurity/california-clinic-to-close-after-ransomware-wipes-out-patient-records.html.

[9] Guo, Eileen, and Hikmat Noori. 'This Is the Real Story of the Afghan Biometric Databases Abandoned to the Taliban'. MIT Technology Review. Accessed 12 November 2022. https://www.technologyreview.com/2021/08/30/1033941/afghanistan-biometric-databases-us-military-40-data-points/.

[10] McCallum, Shiona. 'Period Tracking Apps Warning over Roe v Wade Case in US'. *BBC News*, 7 May 2022, sec. Technology. https://www.bbc.com/news/technology-61347934.

[11] Diehm, Cade. "On Weaponised Design." Tactical Tech, 16 February, 2018. https://newdesigncongress.org/en/pub/on-weaponised-design.

an umbrella term, weaponised design identifies shortcomings of bias, poorly considered trade-offs, and undeserved placement of trust in organisational behaviour as key causes of harm, and points to broader gaps in policy and practice that paralyse attempts to confront the consequences of digital infrastructure design. Key to understanding this is the unwavering belief by tool-makers, security researchers and other technologists of the inherent integrity of their work within a compromised system.

The disconnect of practice from material conditions is accelerating thanks to a generational narrowing of perspective of systems designers caught in the vice-like grip of the doctrines of scale and cybernetics. The a priori belief in the self-sustaining nature of such technical systems, borrowed from the study and purposeful destruction of ecosystems by colonial scientific ventures, informs much of the blinkered understanding powering the fields of cybernetics. This legacy influences ecology and the entirety of contemporary technology.[12] A rationalist systems approach to digital practice cultivates a feedback loop where the solutions offered to solve structural harms create new structural harms. While the intentions of practitioners often come from a sincere desire for intervention, these interventions themselves risk cascading second and third order effects. Those who bear the brunt of future harms are often already marginalised and disenfranchised by the very technology systems now deployed to produce new solutions. Examples of this include:

- The never-ending cat-and-mouse game of personal digital security, where individuals must be trained in unfamiliar rituals and technology to defend themselves against their own personal devices;
- The ethics-led 2010s activist movement that highlighted racial bias in machine vision, resulting in the proliferation of violent surveillance systems targeted at minorities who had previously failed to be recognised by biometric systems;[13]

[12] Royer, Benjamin. ‘The Imperial Sensorium’. The New Design Congress, 21 June 2022. https://newdesigncongress.org/en/pub/the-imperial-sensorium.

[13] Diehm, Cade. *Will Design Ethics Save Software? (Ethereum Foundation Devcon5)*, 2019. https://www.youtube.com/watch?v=Bk-NSADkdrs.

- The ongoing futile attempts to develop ethical digital consent where users are asked to provide informed consent to data sharing in incomprehensibly complex networks of relations and actors.[14]

While much work needs to be done to bring broader awareness of the shortcomings of the practices of tool and infrastructure building, momentum to address these problems is growing. Re-thinking assumptions around infrastructure scale, challenging the storage of user data, increasing awareness of socio-technical security[15] and efforts to embrace – *rather than erase* – the complexity of digital systems all represent tangible threads for which an alternative and more resilient types of digital tooling may emerge.

Digital archiving is the practice of collecting and/or digitising, cataloguing, storing, curating, navigating and retrieving cultural or historical material for preservation, using computers and other digital or electronic systems. Digital archiving is a computing discipline with a number of unique complexities. The practice borrows heavily from its academic equivalent and draws its definitions of collection, curation, preservation and storage from historical practice. At the same time, digital archiving is both enhanced and affected by digital systems. Digital archiving is not always focused on safekeeping digital artefacts, digitisation of physical ephemera is just as important for the practice. The digital aspects of digital archive have particular influences on the practice through methods of storage, implications of privacy, the challenges of digital resilience, material accuracy, fidelity and other factors.

Web archiving is a subsection of digital archiving and preservation, where the focus of archiving is contained to material available on the internet. Web archiving remains a niche discipline but is a profoundly important one. The preservation and curation of web-based material for cultural, legal or historical reasons can be just as crucial as its physical equivalents, although the expectation surrounding the temporality of internet content remains hotly

[14] Diehm, Cade, Kelsey Smith, Ame Elliott, and Georgia Bullen. ‘The Limits to Digital Consent: Understanding the Risks of Ethical Consent and Data Collection for Underrepresented Communities’. *Simply Secure*, 25 October 2021, https://simplysecure.org/resources/The_Limits_to_Digital_Consent_FINAL_Oct2021.pdf.

[15] Goerzen, Matt, Elizabeth Anne Watkins, and Gabrielle Lim. ‘Entanglements and Exploits: Sociotechnical Security as an Analytic Framework’. *9th USENIX Workshop on Free and Open Communications on the Internet*, 13 August 2019.

contested. The United States has produced the majority of attitudes and practices towards digital archiving, driven by state actors (e.g. the Library of Congress), publicly funded institutions (e.g. The Smithsonian) and larger non-profits (e.g. the Internet Archive) and their supporters (e.g. Electronic Frontier Foundation). Also playing a major role is a selection of European counterparts, with major contributors hailing from Germany, Denmark, Holland and the United Kingdom amongst others.

The landscape for web archive tooling is small. Tools are resourced either by a handful of primarily US or Western public or political interests, supported through voluntary or grassroots open-source projects, private endowments, State apparatuses, art institutions, or - more recently - through successful cryptocurrency speculation and capital raising. Each of these sources have distinct clusters of social motivations and expectations of what an archive is and how they are created, curated and maintained. Incredibly, almost all archival efforts rely on just a handful of base tools to capture and maintain their collections.

Commercial digital archiving storage software is almost always inappropriate for their advertised purpose. Open-source archiving tools and systems often require large topologies of infrastructure and specialised expertise to create and maintain, and their commercial counterparts provide products and services to make digital archiving more accessible and cost effective. However, the requirements of archiving and the nature of capitalism present a conflict of interest, where resources committed for preservation are collaterals for profit extraction. If and when a financial agreement ceases between an archivist and a commercial vendor, access to an archive may be restricted or the archive may be destroyed.

This report is part of a material examination of digital archival tools and practice, against the backdrop of a rapidly deteriorating set of global conditions. It is the second publication within the larger research roadmap by New Design Congress, and thus is scoped to web archiving rather than the entire discipline - a framing that provides a necessary focus but also entails limits. Commercial web archival systems are also excluded from this research for scoping reasons, with exceptions made on an individual basis.

At a surface level, web archiving appears to have a simple definition. However, the practice is far more complicated than it seems. Web archives include:

- Static websites, plain HTML and CSS, and optionally page-enhancing javascript;
- Complex live-updating web applications (such as social media platforms) that require special tools and significant effort to collect an accurate representation for archival purposes;
- Audio and video media, including podcasts, streaming services, video sharing sites, etc;
- Video games, including web-based game consoles, web-distributed video games, distribution platforms, etc;
- Binary data, such as niche software applications or even malware;
- Proprietary objects, such as Adobe Flash and Microsoft Silverlight components that require special visualisation in order to reproduce.

Web archives are also often substantially larger than other forms of archives due to the accessibility of web authoring tools, the nature of hyper-linking as a core paradigm of the web, the ubiquity of social media platforms, and the low cost of content creation and distribution combined with the requirement of storing versions of archived material that changes over time. Web archives deal with significant complexity both in their diverse range of preservation objects and the sheer scale of potential archival material inherited from the vastness of the Net.

Web archive tooling is maintained by a small number of entities. The material manifestations of web archiving can be observed through the tools built by these individuals and organisations. One influential web archivist entity is the Internet Archive, a San Francisco-based institution with a significant contribution to the field. Over the past 25 years, the Internet Archive has, in their own words, been *"building a digital library of Internet sites and other cultural artefacts in digital form."*[16] The Internet Archive is a widely known web archiving institution with significant influence, whose output includes tools and formats for archiving and digitisation, public access to curated and

[16] 'Internet Archive: About IA'. Accessed 12 November 2022. https://archive.org/about/.

user-generated collections, and pay-walled archival services that support the organisation beyond endowments and donations.

Because the Internet Archive's philosophies and products are influential, examining its institutional behaviour, its philosophies and priorities, the choices in the design and implementation of its tools, and its history serve as a useful shorthand for illustrating broader strengths and failings within the archiving discipline. This is illuminated in moments where the organisation's ideologies and methodologies clash with the material realities of world events, create uncertainty around collaborators' safety, or cultivate tension with vulnerable online communities subject to archiving.

The Internet Archive and its collaborators (e.g. Archive Team) consider the Internet to be of cultural significance, and their prime directive is to preserve as much of its contents as possible. These teams are motivated by a shared perception and ideological objection to the destruction of online material – defined as digital culture – at the hands of corporate decision-making processes or government interests. The Internet Archive will often archive and preserve a particular platform that will soon cease to exist, collect material subject to copyright disputes or censorship, or develop entire large-scale archives during times of real-world state conflict.

In framing their practice through such an urgent lens, where decisions must be made quickly to preserve fragile or at-risk digital culture, the Internet Archive is able to dismiss objections to digital archiving by framing such criticism as not understanding the permanence of the Internet. While this position can often be justified, this urgency also helps drive the development of tooling and infrastructure using approaches that do not account for the agency of individuals and communities subject to archiving. The most basic example of this is the practice of ignoring a site owner's *robots.txt* file – a ubiquitous way for system administrators to express consent by explicitly asking autonomous programs not to archive a particular site.[17] Archive Team co-founder and Internet Archive software curator Jason Scott describes their work as motivated by a shared sense of powerlessness against digital rot:

[17] Graham, Mark. 'Robots.Txt Meant for Search Engines Don't Work Well for Web Archives'. *Internet Archive Blogs* (blog), 17 April 2017. https://blog.archive.org/2017/04/17/robots-txt-meant-for-search-engines-dont-work-well-for-web-archives/.

> *"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."*[18]

While there is tremendous value in capturing a working copy of the Internet, this is an uncompromising and hardline stance driven by ideological opposition to censorship and corporate platform economics. The simple act of indiscriminately ignoring the directives described in *robots.txt*, regardless of justification, can also be interpreted as disrespecting the wishes expressed by system administrators to withdraw consent and who do not wish their systems to be archived.

An example of the tension between the archiver and archive subject consent can also be seen in Archive Team's rapid archiving of LGBTQ+ communities during the deplatforming of 'adult content' by Tumblr in 2018.[19] Motivated by a sense of urgency in the face of real permanent data loss in the aftermath of Tumblr's policy change dictated by their then new parent company Yahoo, the Archive Team crawled hundreds of thousands of potentially at-risk user blogs, only deploying an opt-out consent system in response to outcry from their practice. When criticised for this indiscriminate practice, both in this context and others, a common justification for this approach to archiving is that a subject of archiving ought to understand that the "Internet never forgets." If a queer person does not want to be absorbed into a permanent institutional archive, they must not participate in sexual or gender politics online.

It is in this divergence of understanding of networks and the response of the archiver that a key example of the tensions of web archiving are best illustrated. The user-perceived impermanence of Tumblr or an overwhelming desire to participate in a niche culture through the network effects of a social platform, combined with a shoot-from-the-hip approach to user consent during archive practice shares historical parallels with organised institutions forcefully archiving disempowered or vulnerable archive targets. The late-stage roll-out of an application process to allow users to opt out of the archive effort saw potentially vulnerable users navigating a system they did not fully understand – if they had any knowledge at all that they had been subjected to

[18] Scott, Jason. *Open Source, Open Hostility, Open Doors*. Open Source Bridge, 26 June 2012. https://www.youtube.com/watch?v=tJqZGRIwtxk.

[19] Electronic Frontier Foundation. 'What Tumblr's Ban on "Adult Content" Actually Did', 20 May 2019. https://www.eff.org/tossedout/tumblr-ban-adult-content.

web archiving from a large institution. In acting radically against a corporate vandal, the treatment of its targeted queer userbase and dismissal of their concerns frame the targets of archiving as complicit collaborators with the archivers' capitalist opponent, rather than a disempowered community caught between conflicting ideologies.

In August 2022, Twitch streamer Clara "Keffals" Sorrenti was 'swatted'[20] by far-right reactionaries for her work as a transgender activist antagonising against the rapidly deteriorating climate faced by sexual and gender minorities in Western societies. This was not a once-off event: the act of leveraging an individual's personal information to coerce hyper-militarised police squads into performing violent raids on harassment targets is a deplorable reality of the internet. For Keffals, this act had been perpetrated by members of KiwiFarms, a far-right forum whose users engage in brutal harassment and surveillance campaigns against outspoken or visibly online minorities. In response to her personal information being posted on the KiwiFarms forum, along with libellous and unsubstantiated claims made against her character, Keffals used her public visibility as a Twitch streamer to launch *#DropKiwiFarms*, a grassroots campaign to pressure infrastructure resilience service provider Cloudflare to drop KiwiFarms as a customer. Despite the latter appearing to violate Cloudflare's Terms of Service, Keffals and her campaign team had to endure weeks of escalating threats to their personal safety before the company withdrew their protection of KiwiFarms.

As the organisation's tools had been used to create an archive of the site and personal information of targets, the Internet Archive faced calls to take down their KiwiFarms archive. In the political effort to dismantle KiwiFarms as a far-right political agent and hate group, the Internet Archive was presented with a situation that contradicted the institution's political convictions. Having provided archival mirrors of the entirety of KiwiFarms that included personal information of political targets, the Internet Archive reacted by removing access to their archive of the KiwiFarms site after Cloudflare's

[20] Swatting is a colloquial term for an act of criminal activity in which an online harasser, armed with the personal information of a victim, makes a false report of a serious crime to invoke a militarised police reaction against the victim or their home. Swatting is often deployed as an intimidation tactic, but perpetrators have also expressed desire for police violence to escalate to the injury or death of their victim.

decision. While this was undeniably the correct decision, in doing so, the institution had unwittingly become the agitator in their own ideology, a destructor of internet culture facing criticism of 'erasing' a historical event.[21]

The claim that the removal of KiwiFarms from the Internet Archive has a chilling effect on censorship or the integrity of internet history is completely baseless, and the Internet Archive was right to move swiftly to remove the archived site. However, the argument opposing the Internet Archive's actions subtly leverages a key ideological distinction found in all colonialist archival systems: that the only valuable memory of record is one that captures an accurate 'apolitical' and unmodified representation of a historical context. As Hong Kong philosopher Yuk Hui points out,

> *"The will to archive turns archives into sites of power. Besides the dominant narratives set up in the archives, we can also observe power in the relationship between institutions and archives. Each institution has its archives that contain its history and discourses. In order to maintain its status quo, each institution needs to give its archive a proper name […]. An archive is also a symbol of authenticity and authority - a monument of modernity."*[22]

The ability to govern, modify, authenticate and destroy archival information raises complex questions around censorship and data resilience, but also user and curator safety. These additional questions are often ignored or downplayed. While the Internet Archive is able to accomplish the takedown of violent material through its governance, other archive and preservation projects cannot. The Inter-Planetary File System (IPFS) is a decentralised protocol that, in their own words, *"preserves and grows humanity's knowledge by making the web upgradeable, resilient, and more open."*[23] First released in 2015, IPFS is a sophisticated protocol for distribution, authenticating and storing data that

[21] This is, of course, nonsense. One doesn't need a historically accurate reproduction of a harassment campaign riddled with a target's personal information to re-tell the history of the site.

[22] Hui, Yuk. 'A Contribution to the Political Economy of Personal Archives'. Edited by Ganaele Langlois, Joanna Redden, and Greg Elmer. *Compromised Data: From Social Media to Big Data*, 2015, 226–46. https://doi.org/10.5040/9781501306549.

[23] 'IPFS Powers the Distributed Web'. Accessed 12 November 2022. https://ipfs.tech/.

can be accessed using methods that, to users, feel familiar to common methods of file or web-browser access.[24]

According to the project's website, one of IPFS's use cases is to provide resilience and permanence to an impermanent Internet.[25] IPFS maintains access to data via decentralisation, where data is replicated across multiple custodians for redundancy and resilience. In order to ensure that the distributed data is not tampered with, IPFS leverages a form of immutability, where data cannot be modified or deleted. All modifications to IPFS content is instead versioned. In an era where democratised access to communication and knowledge clashes daily with censorship and information warfare, the need for data resilience in a generally untrusted and ungoverned distributed network of custodians is obvious. Enforcing transparency and versioning creates a system of accountability for network participants. But this is a hardline stance towards data resilience that has immediate and under-acknowledged real-world implications.

In the case of KiwiFarms, IPFS could have theoretically been deployed to assemble an undeletable archive of the site. This would result in a catastrophic permanent repository of personal information collected with the intent to harass and intimidate. The exploration of resilient protocols like IPFS by far-right actors whose internet presence is in the process of being deplatformed is not new. In 2019, an Australia-born fascist posted a manifesto and livestream link to the far-right 8chan message board before murdering 51 people in Christchurch, New Zealand. In the subsequent blowback to the atrocity, the administrators of 8chan openly explored data permanence options[26] - including IPFS, peer-to-peer protocols and blockchain projects - as they raced to subvert the takedown of the site.

For KiwiFarms, the removal of the site from the Internet Archive spurred the forum members to begin creating their own archive of the site. KiwiFarms members had already published how-to guides for using tools that can be operated by individuals rather than relying on institutional services to produce an archive. The constant risk of deplatforming experienced by far-right operatives has

[24] Trautwein, Dennis, Aravindh Raman, Gareth Tyson, Ignacio Castro, Will Scott, Moritz Schubotz, Bela Gipp, and Yiannis Psaras. 'Design and Evaluation of IPFS: A Storage Layer for the Decentralized Web'. In *Proceedings of the ACM SIGCOMM 2022 Conference*, 739–52. SIGCOMM '22. New York, NY, USA: Association for Computing Machinery, 2022. https://doi.org/10.1145/3544216.3544232.

[25] 'IPFS Powers the Distributed Web'. Accessed 12 November 2022. https://ipfs.tech/.

[26] Kuhn, Daniel. '93 Days Dark: 8chan Coder Explains How Blockchain Saved His Troll Forum'. CoinDesk, 6 November 2019. https://www.coindesk.com/markets/2019/11/06/93-days-dark-8chan-coder-explains-how-blockchain-saved-his-troll-forum/.

driven its proponents to examine alternatives. As KiwiFarms faltered, its members shared advice for operating archival tools to preserve the site. Among the recommended tools was Webrecorder. Webrecorder is an open-source suite of tools that is capable of producing high fidelity portable web archives via user friendly browser plugins and automation software. For KiwiFarms, Webrecorder was one of a suite of emerging tools that provides local-first archiving free from control by institutions. Some users were already familiar with the open-source tool because, in one sense, Webrecorder offers its users the ability to route around the established relationships between institutional service providers on one side and users, communities and other entities that rely on digital infrastructure on the other. This is an ideology that is rising in popularity thanks to increased awareness of the implications of surrendering agency to platform owners in exchange for convenience or access to digital products and services.

For tool makers providing platforms for archive and preservation like Webrecorder, the complexity of responding to users' needs is heightened by the fact that providers cannot look, 'physically' or ethically, into these users' archives. In our research, a participant recounted an unexpected question from a user:

> *“There was one user asking 'How could they sort my collection of porn videos by the age of the actors?' We really had to think, what are we doing now? Do we even reply to this? Or do we look into this person's materials? Legally, if you don't know what someone is doing on your platform you are also not responsible for it. If you have no way of knowing, you're not responsible for it. There was another time when there was a pull requests to our open-source project: a very small pull request about some visual cosmetics. When we did due diligence on that user, we quickly discovered they were an outspoken neo-Nazi.”*

Stories of surprised open-source developers encountering reactionary forks of their livelihoods are also common:

> *“Around 2017-2018, someone forked the DAT project/DAT protocol, the Beaker browser and a whole bunch of other tools and basically made a far-right clone of all of them. And all they did really was just rebrand the entire thing as a singular fascist project… the guy was*

> *bothering all of the DAT people in that in that sort of Pepe style, alt-right kind of way. Just a really uncomfortable, kind of low key scary but not threatening. Just like 4chan trolling."*[27]

Whether driven by desires for self sovereignty, real or perceived privacy benefits, bad or traumatic experiences at the hands of a platform owner, or something else, on-device or self-hosted alternatives are both increasing in popularity[28] and becoming more accessible to non-technical users. Given that the far-right are amongst those who experience acute platform instability, they often act as a canary in the coal-mine, seizing the movement for alternative self-determined products and infrastructure to pursue their agendas uninterrupted.

In an expert interview live-streamed as part of research into infrastructural colonialism by New Design Congress in 2021, digital artist and Black trans activist Danielle Braithwaite-Shirley detailed how privilege constrains the potential of digital infrastructures. To paraphrase her words:

> *"In an alternative history of the development of video game technologies, what would game engines had look like had they been built and controlled by under-represented demographics, whose aesthetics and relationships to computers often differ substantially to the privilege of the male-dominated field of the time? What features, graphic options and interactions were left behind due to the identities and material realities of those who created them?"*[29]

Braithwaite-Shirley is far from the first to condemn the narrow conceptualisation of infrastructure developed from privilege and safety, and despite speaking about the niche of game engines, this provocation transcends

[27] Other examples abound, for instance in the case of the Fediverse, see Diehm, Cade. 'This Is Fine: Optimism & Emergency in the P2P Network'. The New Design Congress, 16 July 2020. https://newdesigncongress.org/en/pub/this-is-fine.

[28] Kehayias, John. 'Meet the Self-Hosters, Taking Back the Internet One Server at a Time'. *Motherboard, Vice Magazine* (blog), 2 September 2021. https://www.vice.com/en/article/pkb4ng/meet-the-self-hosters-taking-back-the-internet-one-server-at-a-time.

[29] Diehm, Cade and Elys Jones. 'The Para-Real Episode 2: We Are Here Because of Those Who Were Not: Claiming The Para-Real with Artist Danielle Braithwaite-Shirley'. The New Design Congress, 7 September 2022. https://tv.undersco.re/w/veXDC5a76MQXkY9NqY6WHr.

disciplines. For web archiving, this should be considered through the Internet Archive's role as *an archival practice facilitator*, Webrecorder's role as a *democratising force for practice and portability* and IPFS's role as a kind of *archival venue or storage warehouse*. The Internet Archive's ideology of digital permanence and belief that free digital culture is under threat, and Webrecorder's belief in democratic, portable tools informs how they build their tools, who archives, how they do it, and what archives these tools then create. For IPFS, it is the uncompromising devotion to enforced distributed ownership and immutable data that informs who can participate in a distributed archive and the safety of who and what is archived in perpetuity. To operate indiscriminately is to operate from a position of power. The negotiation of ideology that becomes baked into tools via rules and assumptions falls to those who have to grapple with the consequences of archiving, not the tool-makers themselves.

As noted earlier in this report, it must be clearly understood that the purpose of examining the Internet Archive, Webrecorder and IPFS is not to single out these projects and institutions for criticism as bad actors. Each of these institutions have made significant, unique contributions to the capability, accessibility and resilience of digital culture preservation at curatorial, technical and infrastructure levels. Each of these examples also exhibits unique constraints due to their respective nature: Internet Archive is an institution, Webrecorder is a small open-source tool project, and IPFS is a large open-source protocol developed by a private entity. Their contributions have shaped social and technological perspectives for data permanence and access, and have helped to democratise web archiving. But as global political stability unravels, these institutions also act as representatives of the manifesting issues of power and privilege, where their ideologies influence tool design, custodianship and curatorial practice. These ideologies are not shifting in line with the broader currents of the world. The criticisms examined here can be applied broadly to the practice of web archiving and act as a shorthand to understand a set of first principles that are common across web archiving practice.

Perhaps unsurprisingly, there is little publicly available research or critical evaluation of the existing beliefs and practices of web archiving and how they manifest consequences for those who are involved with, subjected to, or interact with the web archiving process. In the broad landscape review conducted in late 2021, the majority of existing research focuses on digital security or data integrity, colonialism in digital archiving, the user experience of tools, or the political or philosophical underpinnings of digital archiving practice. This report is a wide inter-disciplinary assessment of the realities of web archiving. This research aims to define and explore the gaps that exist between the principles of tool-design in web archiving, and the shifting political realities that provide existential challenges to these principles and their manifestations. ✳

II. Research methodology

This research was commissioned in December 2021 by the Webrecorder open-source project with the immediate goal to help the project inform the Web Archive Compressed Zip (WACZ) format and the user experience of the Webrecorder collection of tools.

Following a landscape review, the researchers undertook an internal threat modelling exercise with the Webrecorder team to develop *anti-user stories*. In digital software design and development, a user story is *"a very high-level definition of a requirement, containing just enough information so that the developers can produce a reasonable estimate of the effort to implement it."*[30] In most cases, users are grouped through shared socio-economic or biographical data, by their chosen devices, or by their technical or cognitive abilities. An anti-user story is a user story that anticipates unintended or unwanted requirements of attackers – users who intend to weaponise the design of the digital product to inflict harm on others. In information security, the practice of *threat modelling* would parallel an anti-user story. In this case, the user and anti-user stories were developed specifically for Webrecorder tools and the WACZ format, and covered opportunities and threats related to functionality, interface, integrity and privacy. The user stories and anti-user stories were compiled on the Webrecorder specifications GitHub project.[31]

The questions that emerged through the landscape review and user/anti-user stories were divided into key areas of focus:

Identity & structures

1. What are the use cases for web archiving in the 2020s?

2. What kind of individuals are involved in the broader landscape of digital and web archiving in the 2020s?

3. What are the motivations for archiving material via digital methods? Have these motivations changed since the last decade?

[30] AgileModeling.com. 'User Stories: An Agile Introduction'. Accessed 12 November 2022. http://www.agilemodeling.com/artifacts/userStory.htm.

[31] GitHub. 'Issues · Webrecorder/Specs'. Accessed 12 November 2022. https://github.com/webrecorder/specs.

4. What roles do organisations and institutions play in shaping the policies, dynamics and outcomes of archiving?

5. What are the historical criticisms of web archiving?

<u>Socio-technical security</u>

6. What are the overarching threats…

    a. …for archivists?

    b. …for custodians?

    c. …for individuals, communities and organisations subjected to web archiving?

    d. …for viewers, both in their access of and perceptions formed by a web archive?

    e. …for archive tools and formats?

    f. …for the archive itself?

7. How are web archiving tools vulnerable to weaponised design?

8. Are currently understandings of baseline risk acceptable? If not, can baseline risk profiles be developed?

<u>Permissions & connectivity</u>

1. How do networked web archives fit into the use cases identified above?

2. What opportunities and threats emerge as an explicit property of decentralised or resilient networked archives?

3. What permission patterns can be identified from a deeper understanding of the topographies of archive network types or information architectures?

4. How do permission systems dictate the topographies of networked archives and help or hinder collaboration between or within archive teams?

Integrity

1. What opportunities and motivations exist for validating the integrity of an archive?

2. How do these identified cases change when the archive is replicated or decentralised?

3. When is uncertainty in an archive's integrity important?

4. What relationships exist between chain of custody or archive governance and archive integrity?

5. How do archives influence comprehension? How can this be manipulated?

Navigation

1. How do current human-computer interaction patterns and/or user experience paradigms assist or hinder large-scale web archives, from the perspectives of archivist, curator, audience and attacker?

2. Are there opportunities to develop new paradigms for navigating web archives?

3. What opportunities exist to improve accessibility for creating and navigating web archiving tools?

Agency

1. What are the implications of user identity within a collaborative networked web archive?

2. Are there opportunities to implement alternative identity paradigms?

3. What risks exist specific to personal identifying information (PII) captured inside a web archive?

    a. Self-doxxing e.g. by an archivist

    b. Regulatory concerns (e.g. GDPR)

4. Do opportunities exist to design user agency paradigms into archive tools and systems and/or their formats?

In December 2021, the researchers published an open call for volunteers to participate in qualitative research interviews that explored the key questions posed by the research. In order to be selected to participate, applicants needed to fulfil one or more key criteria. Applicants needed to:

- Have participated in the practice of web archiving, either actively or historically;
- Be a member of an institution or collaborated with others in their archival practice;
- Hold an occupation as an archivist, journalist, activist or researcher;
- Be a member of a community that had been subjected to archiving;
- Had familiarity with digital archiving tools of any kind.

The open call for participants was circulated through the researchers' networks, and further circulated in adjacent communities. Significant efforts were made to ensure diversity of gender and sexual identity, race, cultural, and socio-economic status. Efforts were also made to ensure institutional archiving participants were not overrepresented in the study. Specific aspects of the global situation presented significant challenges to ensuring broad representation. These included:

- The COVID-19 pandemic and participant burnout, particularly amongst minorities or in communities experiencing instability or deteriorating safety as a result of the pandemic;
- The February 2022 invasion of Ukraine driving efforts to preserve Ukrainian culture and support the documentation of propaganda and news events;
- An acceleration of information warfare experienced over 2020-2022 leading to decreased availability from archivists;
- Wider global economic instability that had real or feared impacts on practitioner livelihoods.

The majority participants were archivists, researchers and academics, and a minority held roles in the tech industry, the arts and the broader civil society constellation. Similarly, Western viewpoints were over-represented, with well over two third of the participants living in, or nationals of, Western countries. Most - but not all - participants identified as cisgender women and men, with an overall balance in representation between the two. This categorisation remains a simplification for expediency and the security of participants, and does not hope to provide any precise metrics beyond pointing out certain biases in the research. It also doesn’t reflect the individual heritage of each participant and its associated influences. The surfacing of such complex interplay of identities, origins and intersectional interests remains the role of the interviews and the subsequent report.

The research interviews were conducted between January 2022 and May 2022 via platforms selected individually by each research participant and facilitated by two researchers - one acting as the interviewer, and the other supporting and note-taking. The interviews were recorded locally by both researchers using OBS Studio, avoiding cloud-based recording features available in services such as Zoom and Jitsi. Although interviews were conducted via video, only audio was recorded. Participants were asked to consent to the interview in advance via the Research Consent Form (see Appendix A).

The research interviews were 90 minutes in length and structured via a series of key questions that reflected the broader research questions (see Appendix B). Participant responses guided the direction of each interview, and the key questions were not always followed sequentially. Recordings of each interview were transcribed and anonymised, before being synthesised as part of the research findings. As per the Research Consent Form, each participant has been offered the chance to review their contribution and withdraw or affirm their participation consent before publication. In the interest of disclosure, one participant withdrew from the project after their participation and their contribution has been removed. The original audio files were destroyed at the conclusion of the research project. ✳

III. Key findings

The broad but flattened definitions of archiving

The definitions of web archiving are not settled and expectations vary wildly across cultural and professional lines.

Some web archives are political in nature and contribute to the empowerment of social and political minorities as a form of social history. This is often a second-order consequence of digital colonialism, where market control of the digital public square has been captured by a handful of companies, whereupon communities are not familiar with any alternative. The political nature of this is deeply kinetic, manifesting as temporal digital records of fast-moving real-world events, as well as the 'rehydration' of social media, where the posting of content on social media and subsequent reactions recording in an interface can be replayed and interpreted. In a key quote from a large temporal archive project, a participant reflected that *"we were kind of inspired from the whole Black Lives Matter preservation team on Twitter."* Other archives are broader and slower in nature, consisting of multi-decade web histories. Sometimes, the curators of archives are more tolerant to graceful degradation, broken components that relied on obsolete web technologies or long-gone server-side rendering.

No matter the type of archive, the role of integrity, accuracy and fidelity are hotly debated. One school of web archiving holds as paramount the ability to fully replay a website, including every interactions of the original site. This perspective considers not just the *content*, but the *functionality* and *fidelity* of the archive collection as being essential to an archive being *accurate*. A second school of thought sees this philosophy as misguided. Driven by concerns around privacy, material density, technical limitations and the colonial nature of the source of truth, the historical accuracy of the archive becomes dependent on both the capture and the custodianship of its contents.

We encountered no examples where the expectations of an archivist matched with what an archival tool could offer. Archivists described how their tools created limitations to their philosophical understanding of archives. Beyond archive crawl accuracy and replay, these limitations were also conceptual: current interfaces and digital nature make it difficult to picture a web archive in one’s head, compared to a physical archive. In more than one case, participants described the *interface dissociation* they felt in their attempts to conceive of or rationalise their work: *“I know 17TB is a lot, but I don’t know how many books that is, or how linear feet that is? What kind of building is my archive held in?”*

The user experience of digital platforms themselves also plays a role in the flattening of an archive, particularly when content and interface become entangled such as in a web archive. For web archiving, both the web content *and* the interface that services the content must be archived. This is significantly different to other digital archives that display their collection with a clearer distinction between material and medium. In an accelerating era of engagement, rich interactions that certain types of material depend upon are often reproduced in an incomplete or broken way. This is especially true in temporal material, such as Instagram Live videos, Twitch streams or Snapchat messages, where impermanence and editorialisation are key. In these instances, the definitions of web archiving stand in direct opposition to the philosophical nature of this ephemeral material:

> *“Web archiving is predicated on some assumptions that are 10+ years old now. There really needs to be a re-imagination of what web archiving is, should be, can be, will be.”*

Archive complexity is overwhelming

> *“It feels like the whole archive is akin to a lake, we are kind of giving multiple entries into the lake. But while material can flow into and fill the lake, you can’t then draw from it. We don’t have a way to do that.”*

The experiences related to configuring, crawling, curation, taxonomy and validation are universally overwhelming to technical and non-technical archivists alike.

Some of the challenges faced by practitioners are caused by the necessary reliance on multiple tools to crawl and collect material for an archive. Differences between the tools and their outputs mean that archivists have to develop specific methodologies for individual tools and the mixing of different tools. While this might be appropriate for configuring a tool for crawling or contributing to an archive, a lack of output standardisation creates substantial complexity that archivists struggle to overcome. Archivists however must rely on multiple tools to cover their respective shortcomings or the priorities inherent in each tool.

Multiple participants noted that the field of web archiving - especially institutional archiving - classifies the internet via the URL and time. While some participants have experimented with persistent identifiers,[32] the majority of the discipline designs relationships between archived material and institutional governance - or decision-making around archives - through a limited implementation of one brittle identifier (usually a serialised unique identifier), and one temporal identifier (e.g. a timestamp).

This is a practice whose shortcomings Trevor Owens, in describing the relationship between Heritrix's[33] relationship to URLs and time, has highlighted: "*[This approach remains] more in keeping with the computing usage of archive as a back-up copy of information than the disciplinary perspective of archives.*"[34] Although simple, combined at scale the URL and the timestamp create a complexity that, together with a lack of archive standardisation, limits the technical and conceptual pathways for archivists to manage their material.

The complexity experienced through tools by archivists may correlate to their over-reliance on automation and may have reinforced the practice of mass-harvesting/broad-crawling. Thoughtful archiving is difficult within a complex archive system, and in response practitioners err on large-in-scope automated over-collection. The output of this process then compounds the problem of archive complexity. Multiple use cases of Heritrix were given as examples, with a common sentiment being that "*the Heritrix model is solely based on automation, and we set and forget. Heritrix really works at scale, but is really sloppy.*"

[32] See for instance, *Persistent identifiers* in Neil, Beagrie. 'Preservation, Trust and Continuing Access for e-Journals'. Digital Preservation Coalition, 1 September 2013. https://doi.org/10.7207/twr13-04.

[33] "Heritrix is a web crawler designed for web archiving. It was written by the Internet Archive. It is available under a free software license and written in Java. The main interface is accessible using a web browser, and there is a command-line tool that can optionally be used to initiate crawls," Wikipedia. See also: https://github.com/internetarchive/heritrix3/wiki.

[34] Owens, Trevor. 'What Do You Mean by Archive? Genres of Usage for Digital Preservers'. Webpage. The Signal, 27 February 2014. https://blogs.loc.gov/thesignal/2014/02/what-do-you-mean-by-archive-genres-of-usage-for-digital-preservers/.

Decentralisation as a pharmakon[35]

Despite broad levels of experience around web archiving, there is a serious lack of shared understanding of how material risk affects different individuals or institutions involved in collaborative archiving. The deficit of shared awareness of risks around participation, material custodianship and asymmetrical risk based on socio-political factors is endemic. Introducing decentralised archival storage systems - which has a second order consequence of also decentralising the legal and physical risk to individual network participants - would be a catastrophic action without significant cultural shift and education for the field. Decentralisation itself is rarely accurately defined by its proponents, and it can take many different forms, such as institutions sharing joint custody on a private IPFS cluster, federated public archives, corporate ventures, etc.

Across the field, there is an acute understanding of how participating in the archiving of a social movements or political events can draw attention to archivists and expose them to potential online security threats or harassment campaigns. Participants expressed concerns about the potential to be identified through accidental doxxing - leaking personal identifiable information - during their archival practice, such as one’s user profile visible in the archived web content. In cases of decentralisation, the inability to modify or retract published material adds an unforgiving permanence to mistakes made in an archival process of overwhelming complexity.

Beyond the rules around deletion within a decentralised archive, material deletion within digital and web archives themselves is also rare. Many institutional archival projects govern their archives with permanence. Problematic or illegal content is instead marked as inaccessible, creating a ‘dark archive.’ This can be institutional policy but also dependent on laws regulating the archive that forbid the deletion of material. Combined with dragnet-style automated archiving - a common activity conducted by institutional archival efforts - many potential candidates for decentralisation contain controversial or highly illegal material. At the same time, the field of institutional archiving is both acutely aware of its own precarity in participating or holding risky archives. These efforts may have legislative protection with regards to illegal content, but at the same time, their policies

[35] In philosophy and critical theory, a *pharmakon* designates that which is at the same time a cure, a poison, and a scapegoat.

under-appreciate the power they wield when they work with those who do not have the same protections: communities, contractors, website owners, activists, or participating institutional and individual partners. This is a dormant disaster in waiting.

Thankfully, there is a level of understanding of the potential consequences of decentralised archive storage, particularly amongst archivists who represent themselves or community-led efforts to produce archives. Participants who were unable to rely on institutional safety were able to articulate their concerns to varying degrees, with many concluding that *"the existing decentralisation protocols are not appropriate for archiving."* To quote one participant:

> *"A big thing about controlling an archive is that you can also remove things. Custodianship means that you have the authority to delete. You have to take responsibility for something and its destruction. None of the decentralised protocols support that."*

Tool difficulties affect archive curation and quality

Alongside the complexity of tool outputs, the technical and user experience difficulties in configuring and operating archival tools directly influence archives - from the curatorial decisions made by teams of what to collect, to the fidelity of an archive and the completeness of what is collected. This is especially true in under-resourced or time sensitive situations.

This means that archives are often culturally influenced by these external providers. Web archives are wholly dependent on a constant tensions between the whims of internet technologies and the abilities of tools. Unlike physical archives, where non-technical users can form their own solutions to challenging archival problems, the practice of digital archiving is inherently one-sided through the technical barriers of writing software. This creates a user/vendor relation that is not reflected in other forms of archival practice. In web archives, technical capability of the team and the available tools shape the archive and narrow the team's potential. This was systemic but did not manifest as common patterns. Each team had their own individual difficulties:

> *"Video is difficult for us to sometimes to capture or to make available. The technological constraints make it difficult."*

> *“The expectation is that web archives are fully interactive replication of the live website. So you can look at a web archive and interact with it as though it was the website on the live web in real time. But we find that very tricky to achieve.”*

Crawls are often supplemented with data bought from the larger institutional archival efforts or directly from social media platforms. In many cases, participants had accepted the role of external data sourcing via third parties into their practice. Participants were broadly appreciative of these relationships with archival institutions, but concerned about institutional change over time: *“What happens after our institutional partner’s leadership retires?”*

Archive navigation is a key area for future research

There are no common practices for developing usable navigation for web archives, either in the practitioner or public user context. This unsolved problem contributes to significant overhead in labelling, authenticating and working with web archive material, limits the adoption of web archiving tools, and harms the portability and quality of their curation.

Participants described highly personalised workflows for navigating large archives, developed specifically for their team or individual circumstances. An unexpected common technique was the deployment of out-of-band navigation strategies: manual systems such as spreadsheets, paper tagging or other documentation that exist beyond archive tools, operating systems, file systems and web browsers. In cases where navigation was considered a priority, teams had dedicated efforts to maintain an up-to-date taxonomy and navigation structure that was manually updated alongside reviews or additions to the archive. Common assumptions of possible solutions – such as full text search – were rejected by participants as contributing to additional complexity:

> *“We’re looking at search access and how we display information. We can do full text search, but it’s not refined. A search term returns hundreds of thousands of results and we still can’t navigate them.”*

Web archive navigation faces additional complexities from the user experience interfaces of the material collected. Websites are often optimised for the business objectives set by the website or platform owner, but these goals are completely irrelevant in an archive context. For example, a social media platform's interface may use user experience techniques to priorities certain navigation pathways or promote specific platform functionality, but these will not work in the archive. In light of this, a web archive has multiple competing navigation systems, one within the archived material (especially with high fidelity, accurate material) and the other made up of the archive's navigation interface itself. This represents a significant and under-researched cognitive load when browsing an archive. Further challenges are introduced from inconsistent functionality due to incomplete or broken interfaces within collected archived material:

> “*There are only two options available to a user. They can enter the site and exit it via a curator-provided navigation structure, or they could enter via any one individual archive point and then navigate or browse freely on the site. The latter is only possible if the site has been captured in full, and if all the various interactions are available.*”

To combat this, archivists described many methods. Internally, practitioners often use URL schemas as entry-points into an archive, tracked via their aforementioned out-of-band workflows. To augment public facing navigation, practitioners described processes of modifying or annotating archive material itself, leaving recognisable signposts embedded in archive material to help the browsing public navigate the material or confirm the non-functionality of part of the archived interface. The act of modifying a web archive for the purposes of ad-hoc user experience design has significant implications for efforts to develop cryptographic authentication strategies for archive integrity.

Archive tools and processes reconfigure trauma response

> *“We don’t collect moments of joy really, we collect bad moments. The trauma of working on that kind of material is very, very difficult.”*

> *“My workload during COVID tripled. I wasn’t even working on a COVID project.”*

This research was undertaken during a deteriorating period in the early 21st century. The COVID-19 pandemic was well under way, having torn through entire communities and overwhelmed hospitals worldwide in its first pre-vaccine year. Decades of climate inaction had begun to manifest as frequent weather-based disasters. But beyond this, participants had worked on projects that either focused on or had intersected with far-right violence, suppression of Indigenous,[36] diaspora, non-Western and minority voices and protest, as well as other major traumatic events. Alongside this, the act of large scale dragnet collection of social media often collects material that has not been subject to content moderation.

Participants who chose to share their experiences with traumatic material spoke with a surprising level of awareness around the nature of the relationship between the archival process, ingestion, taxonomy, and display of material. While the experience of and response to trauma is unique to each individual, tool interfaces played an unexpected role in amplifying or mitigating the potential for harm.

Participants described how automated collection of traumatic material either helped protect them from harm, and conversely overwhelmed others with its volume and how the sheer magnitude of harmful material was presented in the archive interface. Participants described the ‘roulette wheel’ of suggested taxonomies for material, and the potential for sudden exposure to harm when correcting a material classification or authenticating the accuracy of a piece of archived content.

[36] Throughout this report, the terms ‘Indigenous’ and ‘Natives’ are employed as overgeneralising shorthand to describe very different peoples, experiences and situations, such as *“displaced, urban, on reserve, African, Asian, of Turtle Island, of Abya Yala, Amazonian, co-managing, frontline, Arctic, autonomous, criminalised, disenfranchised, self-governed, uncontacted, on their ancestral territory, confronting or collaborating with external stakeholder, etc.”* The reader should always keep in mind the fundamental heterogeneity that these terms can describe. Reinforcing pan-indigeneity had, and still has, a colonial utility. We would like to thank the participant who brought this to our attention.

In all cases where participants spoke of trauma in archival practice, participants referred to user experience flows that either helped them mitigate harm, or – unfortunately more often – they felt contributed to a personal traumatic outcome. In determining what recommendations should be considered as part of this research, there was a surprising lack of applicable prior research into the potential outcomes of interface design optimisations with regards to traumatic harm amplification or reduction.

Beyond the potential of trauma within the archive, participants also described the trauma of understanding the potential or realised weaponised design of an archive, for example in cases where an archive was repurposed to prosecute Indigenous activists. In this case, the framing of the archive as a historical memory had entirely different conceptual meaning depending on the background of the individuals subjected to archiving, many of whom had never needed to consider shared cultural memory as a tool for policing and surveillance:

> *“I worry that the people that we work with don’t fully grasp the idea of data having a potential to be permanent.”*

One participant described a choice made by an institution to archive the personal website of victim of suicide. The participant described an internal conflict between the institutional mandate to provide ‘archival accuracy’ against the obvious trauma and deeply personal material. While the tooling made the process of preservation effortless, the larger archival services and tool makers do not fully consider the role of trauma in their practices. The gap in policy, combined with the potential fidelity of archive tools, creates a sort of socio-technical purgatory that can haunt practitioners:

> *“This person who wanted it out in the world has now passed. This is a document of trauma. What’s the right thing to do here? Should it be publicly available, or even be preserved? Who do I ask? Would it be different if a grassroots organisation contacted a site owner, and requested this document be saved? Would they think differently about people accessing this sensitive content they currently hold? I think about this a lot.”*

Colonial methodology and language narrows archive potential

The archival frameworks of digital and web archiving are cogent with colonial epistemology, who sees itself as the sole judge of what knowledge should be, how it should be done, and for what purpose. The practice of web archiving is descendant from Eurocentric academic practice and as a result, the language, philosophies and set of behaviours have carried over to the internet age unchallenged. This is evident in even the lexicon of web archiving itself, the 'capture' of a web page is an uncritical utilisation of generations of colonialist fascination with the Other, capturing and storing objects far from their context and without the consent of those subjected to the capture:

> *"Museums are in the capital cities, museums are in the colonial centres, artefacts were gathered and taken away, researchers have gathered data and taken away the concept of archives living in a community."*

Participant perspective of archiving was either completely outside of decolonialist criticism or deeply informed by it. Participants that offered decolonisation criticism of archival practice described issues of accuracy, agency, flexibility, resilience, digital safety and curatorial richness. Many of these participants could identify opportunities to improve, iterate or, in some cases, completely re-imagine the practice. No one could provide examples of decolonised archival tools:

> *"The mental models of the Western archiving discipline is alien to Indigenous communities, a lot of care needs to be expounded in order to not expose them to danger generated by digital media."*

Participants who did not reflect on colonialist or power structures inherent in the practice of web archiving were more likely to be involved in archival efforts that had concerning policies or qualities, such as cavalier attitudes to dragnet archival practice, lack of consideration towards archivist and community privacy, or a lack of proactive policies regarding illegal and sensitive content held within an archive.

The colonialist approach to web archiving is an active barrier to producing a new generation of politically-aware tools. Without a direct interrogation led and informed by decolonialist voices, it is likely the next generation of tools will, through the defence structures inherent to colonialism, uphold a set of implementations that broadly influence the practice and continue to narrow the potential for web archives. As one participants warns:

> *“The granularity of the obsession of finding the most minute details for all things is just part of the colonial life. We’re trying to conquer the whole concept of all global knowledge across humanity.”*

The outcomes of this motivation are already being felt.

Digital archiving is vulnerable to political and ecological threats

The digital archive is a surface of broad political threats, targeted both at those subjected to archiving, and the archive itself. Practitioners and other participants alike described cases of weaponised design, where the implementation and availability of an archive harm communities while performing exactly as intended. Digital archives are vulnerable to attack and weaponisation from harassment operations, where information about an individual or community is collected and re-purposed to fit a false narrative. An example of this is detailed in Part I of this report, where the weaponisation and distortion of personal information of individuals by members of a harassment forum was an active strategy to generate false narratives to discredit political opponents.

The development of material for a political web archive also has potential consequences for archivists themselves. The surveillance-riddled modern internet contains interfaces that can leak personal information and identify the archivist responsible for crawling a site. This can be as simple as being logged into a social media platform and having that user state collected along with the intended material for archive, or an archivist’s identity can be derived via careful analysis of collected targeted advertising, social graphs and other metadata.

The weaponisation and political danger of archives must be considered carefully as a key factor in the development of authenticity or integrity systems, as well as efforts to provide deletion-resistant archival storage. A dominant tonality expressed during the interviews was that *“it would be so much easier to go other* *places to find any kind of content you wanted, than to try to use this archive."* On the contrary, some voices questioned this diagnostic and raised the very real threats pertaining to the digitisation of archives:

> *"One person's archive is another person's police dossier. It has to be understood that this is going into an archive, it is just as accessible to a researcher with good intent as someone with bad intent."*

Beyond the political vulnerabilities of archives, current approaches to digital and web archiving are vulnerable to ecological and other physical threats to data resilience or network access. This can manifest as policy directive to destroy material unexpectedly during deteriorating political situations, during a change in company or government leadership, or as belligerence between nations and the weaponisation of infrastructure like data centres, companies or backbone networks. An illustration of such vulnerabilities occurred when *"the Trump administration removed climate crisis research from the government website, and in some cases destroyed the data."*

Threats of this nature often have historical allegories. For example, archives of the colonisation of Kenya, the massacre of the Mau Mau as well as their internment in concentration and extermination camps were burned and dissimulated by the British government.[37] [38] The reality of institutionally-driven destruction of preserved material - both in the present and their historical allegories - is not widely considered. The trend amongst larger institutions and tool makers in this space is to consider their positions as stable, as though digitisation and a surface-level democratic process offer resilience and fortification:

> *"Do we store the contents on a university server, thereby being under their power and control of what gets added? Do we rely on the good faith of myself or other librarians who could be gone the next day?"*

Finally, the rapidly deteriorating climate situation has immediate effects for the resilience of digital archives. An over-reliance on climate-controlled data

[37] Cobain, Ian, and Richard Norton-Taylor. 'Mau Mau Massacre Cover-up Detailed in Newly-Opened Secret Files'. *The Guardian*, 30 November 2012, sec. World news. https://www.theguardian.com/world/2012/nov/30/maumau-massacre-secret-files.

[38] Cobain, Ian, Owen Bowcott, and Richard Norton-Taylor. 'Britain Destroyed Records of Colonial Crimes'. *The Guardian*, 17 April 2012, sec. UK news. https://www.theguardian.com/uk/2012/apr/18/britain-destroyed-records-colonial-crimes.

centres, network disruptions from major climate events, and a lack of portability that makes very large web archives difficult to physically relocate, poses a significant risk that has not been fully considered. One participant, whose practice centred around anti-colonialist resistance, highlighted how geography and politics were intertwined with the resilience of their work:

> *"Our region will be greatly affected by climate change. We were greatly affected by COVID. Being in such an isolated place puts us in a unique position to consider those challenges."*

The threats to digital archives remain underestimated. From the deteriorating political situation in the United States and in other flashpoints across the globe, to the deployment of surveillance infrastructure worldwide disguised as pandemic response, the unjustified war on Ukraine or the deteriorating resilience of data centres to climate events, assumptions around political and physical resilience baked into the design of today's digital archival systems are being tested in real time.

Regardless of their own assessment of personal or institutional risk, participants were acutely aware of the global political climate. From archive resilience, to concerns around documentation and preservation of protest, questions around supply chains, and institutional risk, a recurring observation by participants was a sense that archive systems were vulnerable to foundational assumptions made by practitioners and tool makers. To varying degrees, participants felt that circumstances were changing faster than the practice could respond to:

> *"The type of malicious actions I think of are less technology-based, but more using social-legal mechanisms. I'm thinking of things like SESTA/FOSTA. I'm constantly thinking about the chilling effect that can be created on people through legal mechanisms towards technology."*

Archive integrity systems have unintended consequences

The archiving of a constellation of accounts bypasses the traditional individualist framework upon which most laws in liberal democracies rest upon.[39] In many examples, the fidelity and ease of access of an archive compared to a physical equivalent transformed the material into an extra-judicial surveillance apparatus, where incentivised adversaries (such as police investigators) could account for complexity with human resources and military-grade technology to build cases against targets.[40] [41]

As discussed earlier, a common assumption is that large, unwieldy archives are inherently safer due to their inefficiency. Some participants described situations that show that this is a dangerous and unfounded position. Dedicated, well-resourced state actors do not see obscurity or complexity as an obstacle. The potential of useful material for a case is enough to request data dumps that can then be sorted by an archivist via legal coercion, or processed by law enforcement. Archives run the risk of misuse in programs such as the US Government's Immigration and Customs Enforcement Agency dragnet, bypassing protections afforded by "sanctuary" cities.[42] Participants also described the misuse of Indigenous, diaspora or social movement archives to identify and police legitimate democratic protest. For a number of participants, these are not theoretical concerns:

> *"We slowly lean closer and closer to fascism. And that means more and more of the materials in the collection that I have been working on incriminates more and more people."*

Archive integrity is often considered essential to fight dis- and mis-information, in open-source intelligence and investigations, and to document criminal activity. Participants were split between recognising the need for

[39] Royer, Benjamin. 'The Imperial Sensorium'. The New Design Congress, 21 June 2022. https://newdesigncongress.org/en/pub/the-imperial-sensorium.

[40] Cox, Joseph. 'Here Is the Manual for the Mass Surveillance Tool Cops Use to Track Phones'. Motherboard, *Vice Magazine*, 1 September 2022. https://www.vice.com/en/article/v7v34a/fog-reveal-local-cops-phone-location-data-manual.

[41] Chabria, Anita , Leila Miller, and Nicole Santa Cruz. 'LAPD Scandal Opens Window into California's Secret Gang Database as Reforms Debated'. *Los Angeles Times*, 3 February 2020. https://www.latimes.com/california/story/2020-02-03/california-attorney-general-xavier-becerra-changes-course-on-revamping-the-states-gang-database.

[42] Pilkington, Ed. 'US Immigration Agency Operates Vast Surveillance Dragnet, Study Finds'. *The Guardian*, 10 May 2022, sec. World news. https://www.theguardian.com/world/2022/may/10/us-immigration-agency-ice-domestic-surveillance-study.

authentication systems within archives, and the multitudes of legitimate use cases for modifying an archive after the fact. While the manipulation of archive to re-contextualise material in bad faith remains a significant problem, participants broadly felt that strong integrity or anti-tampering systems within archives could have negative effects on their curatorial responsibilities. Many participants described the question of archive integrity as a social problem and considered the institution as the primary arbitrator of archive integrity.

The interface required for an integrity system would need to cultivate within a user a degree of authority and trust that material is not tampered with. From an information security perspective, the introduction of archive integrity and authentication systems adds an additional attack surface that can be gamed and defeated. Once defeated, the manipulation of a digital archive that is then misclassified as authentic may amplify the effectiveness of the falsification, by decreasing out-of-band checks by users and weaponising material trust cultivated through the integrity interface to produce additional impact.

WACZ is well regarded but remains hermetic

> *“Webrecorder was a real gift.”*

> *“I’m always grateful to the Webrecorder team for their help.”*

When asked direct questions about their experiences with Webrecorder and the WACZ format, participant responses were overall positive. Many participants understood the role and potential of the WACZ file format, and held excellent opinions of the tools developed by Webrecorder, but overall sensitivity to technical failure or user experience frustrations were amplified by the complexities of web archiving as a practice.

Participants who had actively used Webrecorder are eager to collaborate, finance and support the project. Participants who fell within independent practice or non-institutional demographics described how initiatives like Webrecorder allow them to generate income. By decoupling from cloud or service-based archive tooling models, Webrecorder allows participants to create archives outside the structures of institutions. In parallel, Webrecorder also assists autonomous and community archiving initiatives, especially in instances where participants may distrust an institutional partner, or need to create community-led archives. For example, the Webrecorder tools help with urgent archiving initiatives, and a major effort to archive the digital culture of Ukraine was ongoing for the duration of the research project.

Participants were asked to describe the difficulties they encountered when using Webrecorder software. When recounting technical and user-experience issues, participants were often unexpectedly passionate and frustrated. One major cause of this may be Webrecorder’s reputation for accuracy and performance which also relies on manual processes. Archivists use Webrecorder in a more manual fashion to fine-tune an archive where other tools and automated processes have failed, and users may begin using Webrecorder in a more involved way whilst in a more frustrated state.

Another core trigger with user frustration may be related to overarching issues of archive complexity and a growing sense of fragility within the landscape of archive practice. Broader research efforts to understand the complexities of archiving and community discussion, tool modernisation, risk analysis and a degree of political consensus will all improve the practice. ✳

IV. Analysis of findings

The basic definition of archiving is the “selection and classification of accumulated documents and objects,”[43] the practice of collecting, saving and making sure that materials that have been selected for preservation are accessible and protected in the long term. Archiving involves the identification, assessment and collection of those elements that are deemed of historical interest for future generations. The definition of archiving also encompasses the development of ways in which a trusted repository can be established so that people can access and find the information that they need. This results in custodianship, a complex set ofinteractions and negotiations between different individuals and entities aligned with similar goals, but often with broad interests and expectations of archival practice.

This analysis of findings is a synthesis of the landscape review and qualitative research interviews with archive practitioners. Although originally scoped to the Webrecorder project, this research demonstrates that the issues of ethics, consent, digital security, colonialism, resilience, custodianship and tool complexity are systemic and can be interrogated from an epistemological perspective. All participant contributions are quoted and can be differentiated from other references through an absence of a corresponding footnote.

Philosophies and motivations

To understand how to best design and develop digital tools for a new generation of archival practice, we start with an analysis of the variations in landscape of definitions, philosophies and motivations of practice.

Archive practice is motivated by a variety of philosophical, political and cultural motivations. Some of these include:

- Cultural documentation of public or state actors,
- Historical preservation for cultural or public interest reasons,
- Investigative journalism,

[43] Hui, Yuk. ‘A Contribution to the Political Economy of Personal Archives’. Edited by Ganaele Langlois, Joanna Redden, and Greg Elmer. *Compromised Data: From Social Media to Big Data*, 2015, 226–46. https://doi.org/10.5040/9781501306549.

- Documentation of war crimes or human rights abuses,
- Indigenous[44] or minority visibility and advocacy,
- Preservation of material placed at risk from external circumstances that can range from platform closure to destruction during warfare.

Many participants reflected on the lack of public understanding of the importance of their work. Within now-ubiquitous digital systems, the term ‘archive’ has been redefined as a ‘not-quite deletion’ across popular messaging apps, email clients and social media platforms. Despite its niche status, web archiving is an important philosophical and metaphysical endeavour that preserves the online Zeitgeist and broadens the digital horizon back in time, both in the immediate and the far future. Unlike the utilitarian use of the word ‘archive’ as a UX synonym for ‘remove and forget,’ archivists described the practice as a public good that enables societies to learn from history and *“maybe even get better as a species.”* Although participants were from a variety of civil society and investigative fields, this research did not engage with practitioners involved in more violent use cases of archiving, such as surveillance or targeted harassment.

The field of archiving is similar to other epistemological and scientific endeavours. It acts as a *creator* – in the design and generation of an archive – and a *protector* – via custodianship, cataloguing and preservation. What these roles mean vary wildly, informed by practitioner education, cultural background, threat analysis, personal or institutional objectives, and political beliefs. While the common definitional core of archiving as a practice of collection and curation surfaced throughout the interviews, many participants had their own idiosyncratic approach to the discipline. Participant approaches often pull the discipline in different directions, and while some are merely the expression of individual or institution specialisation or perspective, due to the peculiar nature of digital archiving differing perspectives sometimes reach irreconcilable contradiction.

The process of archival ‘capture’ – whether digital or analogue – is the *“snapshotting of a moment in time, through the use of information.”* Archives are composite objects created through human actions, and in that sense similar to

[44] We would like to remind the reader that the terms ‘Indigenous’ and ‘Natives’ are employed here as overgeneralising shorthand to describe very different peoples, experiences and situations.

historiography. As one participant described it: *"The best example I was told was that museums buy, acquire and show art. Archivists collect all the stuff around the art."* Archives are a sociological record of particular points in time that involves technical ability and the philosophy that drives curation, both exerting influence on each other. *"I found that my decisions and my point of view were necessarily bound up with the object that I collected."*

While there some consensus amongst practitioners as to the underlying philosophies inherent to archiving, digital archiving and web archiving, there are also significant divergence. As mentioned earlier, one philosophy prioritises the ability to fully replay a website - a level of fidelity that includes every interactions of the original source - as a key necessity to a web archive's integrity. There are multiple justifications for this. One argument draws from contemporary media theory inspired by Marshall McLuhan, whereby the content being archived cannot be separated from its user experience, and that the medium and the message are to some extent isomorphic. In other cases, the broader content being collected is essential to the wider Zeitgeist that the archived material exists within - such as algorithmic advertising presented alongside the main content, like/retweet counts, or the ability to explore related content through the information architecture of the interface for which the archived material was originally published on.

To others, the emphasis on accurate reproduction is considered misguided. The most common argument focuses on the complexity of high-fidelity collection and storage/access processes, all of which are vulnerable to digital rot, broad changes in the technology landscape that render material inaccessible, digital rights management, and the evolving frameworks of web development that do not account for accessibility and archival practice. Practitioners who shared this perspective believed that digital archiving will never be a lossless process, both in the functional fidelity of captured material and the epistemological nature of digital archives as digitised representations of culture. These perspectives drew parallels between technically incomplete archives and non-Western forms of archiving through stories, illustration, poetry and oral culture.

This tension highlights the tense political nature of web archiving. The participants involved in institutional or investigative archiving, or who described their work as part of efforts to hold government and power structures accountable, or to document the online histories of marginalised communities during major events, tended to consider the archive as an authoritative source of historical truth, where fidelity and accuracy were paramount. The philosophical relationship between the archivist and the communities subject to archiving can be collaborative, but in effect often considered communities as archival objects. If, in some isolated cases, community involvement was described as voluntary, there was still little opportunity of genuine ownership over the process.

The breadth of archiving goes beyond institutionalised record-keeping. This was illustrated by examples described by other archivists - particularly activists or grassroots practitioners developing cultural or political web archives via autonomous tools or community crowd-sourced labour. Participants who engaged in this work described how web archiving actively contributed to the identity and empowerment of social and political minorities, be it from their documentation of fast-moving political movements, to the cultivation of Indigenous web archives in collaboration with institutions. One participant said:

> *“We created a tool repository that can point future generations to an important moment in our history, kind of like a second revival. We think the protests have galvanised renewed interest in our culture. We want to preserve and capture that for future generations and just deal with the complications of their utilisation of multiple formats that are difficult to preserve.”*

There is a direct relationship between the philosophical definition of a web archive and the resulting participation, delegation of labour and expectations of fidelity within the web archive. This philosophical difference has crucial outcomes for the health, viability and applicability of an archive. In cases where practitioners were able to embrace the limitations of web archiving, practitioners described a fault-tolerant philosophy that cultivated distributed and cooperative contribution to and curation of the project, where archiving is seen as an endogenous rather than an exogenous facet of life:

> *“I think when we first started, the term archive wasn’t in our dialogue. But we found that people really respond to photography paired with stories. So in our archive we highlight cultural events,*

> *document liberation movements and archive the realities of people's lives. We archive trauma as well, how our identities can cause trauma to us. Even though digital archives sound very boring and heavy, I feel like ours is very entertaining. It's very fun – our contributors have fun with it!"*

Practice

Archive practice involves independent professionals, journalists, lawyers, artists, civil society workers, government employees, civilians and volunteers. They archive a wide variety of subjects, such as internet art, government communications, material for investigations and prosecution, commercial properties and platforms, community and cultural material, collections of digital products and literature, or in some cases as much of a nation's internet output as possible. Digital and web archiving as practices adopt very different forms depending on their scale – from small initiatives with a very limited budget to a giant enterprise undertaken by governmental institutions.

The digitisation of societies growing, the pivot to digital archiving began as organisations faced increasing digitisation of their subjects of archiving. This was often not driven by convenience, but a reaction to the fading away of analogue materials: *"the print magazines, newspapers and grassroots publication that were heavily used for research about the 20th century do not exist today. We don't know what people are going to use for the 21st century but we might as well try and collect some of it because maybe they'll use it."*

The digitisation of archival practice is still in an adolescent stage, stuck between traditional methods inherited from the archiving of analogue materials, and the technological alternatives that may potentially be unleashed by the information economy. As participants discussed their practice, a set of key concepts emerged as good heuristics for digital archiving: *significant properties*, *graceful degradation*, and *original order*. One institutional participant, with a background in academic archiving practice, summarised these heuristics:

> *"Are the main components present? The focus of an archive is identifying the most important parts of the page for archival. This is a significant property. Sometimes, you can identify significant*

> *properties from the structure of a page. I may see that images are missing in an image-heavy page and determine the images mattered as much or more than any other element of the page - they are classified as a 'significant property.' I may consider their absence as a failed attempt at archival."*

Going further, the relationship between graceful degradation and significant properties is equally important, and not necessarily controlled by the archivist:

> *"If someone is looking at a website and they only care about the text, then they might not care that a significant property is missing if, for example, the archive has a description of what the missing images were, such as alt text. What does graceful degradation look like? It depends on the classification of significant properties by an archivist, and also on how the user of the archive is defining significant properties."*

A major challenge to web archive practice is that an archivist needs to anticipate the present and future of an archive, use this prediction to inform the significant properties, and make decisions on what is acceptable as graceful degradation. Alongside this process of curatorial speculation, archivists must plan their practice around the original order of the collected material:

> *"You're not trying to introduce your values and judgements into the things that you do. You are trying to keep the material in its original order."*

> *"Web archiving is a science. It is defined by the way you create series and collections in a conceptually meaningful way, where things go together. And also where things reflect the original purpose of the materials."*

As a necessary precondition to satisfy these requirements, participants described different ways to scope archival projects. Sometimes archives are produced in reaction to emergency or urgency, such as when a website is going to soon be put offline: *“We were doing kind of rapid-response collecting galleries that were going out of business.”* Some establish a list of domains that are at risk of decaying, e.g. closing, based on old technologies, etc., or rely on a tool to track changes or updates on an intended archive site over time. A list of pages, domains and websites to crawl can then be drawn.

Following this scoping, archival practice involves the process of capture. The majority of this is automated via crawling, in which automated scripts are run on a practitioner’s computer or remote server within their control. Participants commonly cited Archive-It, Heritrix and the Webrecorder suite of tools as amongst their primary tools for archiving, switching between different tools depending on speed, known limitations, and to offset failures or difficulties with certain parts of the crawl. Webrecorder is commonly used to collect dynamic or highly interactive material, which often fails to be fully captured by other, more automated tools.

Participants described using automated crawls before or during scoping to test the waters of an archive’s proposed scope:

> *“We run test crawls, we change things about the the rules and what’s in scope, what’s out of scope.”*

Combined with the common overarching belief in the value of storing immutable web archives addressed earlier, it was often unclear whether the process of scoping is truly designed to define and restrain the crawling or archival process.

Automated crawls require a lot of human care and intervention due to their limitations. They rely on extensive manual QA, visibility around statistics, key performance indicators, and error reporting to assist the QA process. Participants described how current implementations of automatic crawlers make it very difficult to detect and correct errors, or change course in response to issues with the crawling process:

> *“It’s incredibly hard to correct the crawling processes. The crawl takes two weeks, and then you see that you’ve made a mistake in your scope, and you have to start all over again. That turnaround means that the material that you want to capture might be gone already, or*

> *it might not be accessible to the crawler at all. There are also financial implications of this arduous process."*

Crawls are often supplemented with data bought from the Internet Archive or directly from social media platforms. Some rely on submissions from donors, either through select individuals or public participation. Large donations tend to be from private collections, and often encompass other material common to web archives, such as obsolete formats or binaries:

> *"There was a collection from a conference, a weird CD-ROM that would run an executable file on your computer. It would auto-play an application with sound bites of different speakers. We couldn't make it accessible. We could put it in our preservation repository, we could do a disk-image of the disk, we could keep the individual executable files as a preservation master file. But there wasn't any way to meaningfully serve that through the digital viewer. So that was where you had this conflict of wanting to preserve this item, complicated very much by certain factors like [the limitations of] our digital viewer."*

Major national institutions and the Wayback Machine employ other forms of public participation, such as a call-out to the public to submit URLs for archiving. These can be backed by manual approval of submitted requests for archive, or automatic - as is the case for the Wayback Machine. In Indigenous or grassroots contexts, this participatory approach to archiving is often taken further in an anti-bureaucratic way, where archives emerge from improvised touch-points: people submitting their work via social media hashtags, tagging or messaging the archive curators, or contributing via email.

The process of documentation follows the capture and gathering of materials, with the creation of finding aids, labels, and contextualisation of the archived material. Documentation is often completed over a period of years after the capture. Digital and web archives are prone to exponential growth over time, assembled from thousands of web pages, from websites built up over many years and through multiple initiatives, and necessitate indexes, guides and instructions. The delivery of the archive in a usable, accessible, annotated, interoperable structure and format is of utmost importance for the long-term survival of an archive. Documentation indicating how to use the archive, what can be navigated, is also a key part of this part of practice.

Like the process of QA, documentation is difficult due to the fact that practitioners still speculate over who the users of the archive are, what they need, what their use cases are, and whether an archive has been successful and is capable of satisfying archive user needs. Participants linked the reporting and logging processes to scoping and documentation, a necessary bureaucratic task that remains under-appreciated by tool-makers.

Archives are stored and accessed using opinionated solutions. For institutions working alone, or smaller archives without the access to capital that allows them to invest in infrastructure, practitioners rely on Amazon Web Services (AWS), other large cloud service providers, public institutions and archival organisations such as the Internet Archive. Larger institutions often collaborate with their peers to create wide-area network mirrors of their archives. In almost all cases where participants reflected on storage and custodianship, the costs and resilience of archives were raised as major concerns. Participants - particularly those engaged in large-scale dragnet archival programmes - described an ever-increasing cost in hardware and software associated with bandwidth, storage and administration.

For more grassroots participants, or in cases where participants were engaged in unusual archive practice – such as using social media as the archive – the resilience or potential for data loss via hardware failure, malicious intent or deplatforming were cited as significant concerns. For non-Western participants, concerns around the deteriorating geopolitical situation resulting in trade sanctions or loss of access via soft-power weaponisation of digital infrastructure by opposing Nation States.

Archive custodianship has an additional role: user access. Some archives permit access to everyone, only needing identification at the physical location of the archive, and none needed online. Some restrict the access to researchers and member of official/recognised institutions.

Archive deletion is rare. Instead, material that is flagged is ‘unpublished’ and placed in a ‘closed archive.’ This material still resides within the archive in a sort of ‘dark archive,’ inaccessible by users but still considered part of the collection. When pressed on the reason for this approach to content moderation, participants pointed to laws and by-laws regulating the archive, where deleting materials were forbidden or only authorised due to a lack of space. Archiving implies recording content that are - depending on hegemonic norms of a society - morally and ethically dubious, illegal or taboo, i.e. pornography, sex work, illegal content, materials with malware, hate speech, etc.

Almost all participants expressed an interest in greater custodian autonomy and many understood the potential of decentralised technologies as a means of achieving resilience, lowering access and storage costs, and ensuring the broader safety of an archive. However, the unmanageable size of archives, the presence of ‘dark archives’ arising from non-deletion policies, and the reality of unequal protections for custodians such as legislative exemptions for institutions versus individuals or geographic differences, mean that the risks associated with decentralised solutions is significant.

Complexity

Across all facets of the practice, complexity remains one of the most significant challenge for web archiving. It has substantial consequences for archive quality, tool development costs, user experience, navigation, and safety. The research participants highlighted specific complexities at various stages of their interviews, including:

- Difficulty in visualising or otherwise ‘materially’ conceptualising a web archive, especially compared to a physical archive.

- Obsolescence of tools and technologies used to publish web material or store, display and execute material in an archive. This is true for both hardware and software and applies equally to both the archive and the subject of archiving. Content online can disappear, become impossible to archive without tool modifications, rely on a technology that isn’t supported anymore (e.g. Flash or digital rights management), be subjected to rate limiting or other anti-bot protections, or the subjects of a web archive can simply migrate to another platform, creating a break in continuity for the archive.

- Difficulty in archiving non-standard or proprietary formats, such as Adobe Flash: *“If you go to the back-end of a typical Flash-based website, you’ll have some HTML, some JavaScript and maybe a little bit of CSS, but the rest is embedded inside a Flash container. The structure is locked away inside and this is a hierarchy that doesn’t exist when you analyse the back-end to help plan or complete QA on the archive.”*

- Ever-evolving difficulties of archiving the cutting-edge outputs of the web design industry, compounded by the millions invested in new frameworks and front-end development practices compared to the limited investment in web archiving. This creates a situation of perpetual catching-up as archivists and tool-makers encounter new front-end solutions that break their practice.

- Difficulty in archiving ephemeral platforms, such as streaming video with audience chat, private chat platforms and published viral metrics alongside social media content. For example, *“Instagram Live has comments and likes coming in real time, but if the person who creates the Instagram Live Video saves their stream, the comment history isn’t reproduced. You don’t really see the reactions that are coming in. There are screen recording tools which can save that information to some degree. But that’s still a flattening of this context: you can replay the live comment stream overlaid on the saved video, but you can’t actually scroll through them the way you would if you were watching that Instagram Live video in real time.”*

- Loss of access of key parts to an archive, requiring a form of ‘digital social archaeology’ to identify, trace and approach a website’s owner to ask for copies of the desired material, and adapting the supplied material to the archive.

- Mass-harvest/broad crawls with multiple tools create archives that are so large and with inconsistent metadata, meaning archivists cannot systematically determine the sensitivity, toxicity or illegality of content within an archive.

- Difficulty in navigating and cataloguing large archives, driven by the complete inadequacy of existing navigation paradigms for meaningful browsing and curation solutions. Multiple participants noted the use of out-of-band archive navigation documented via Excel spreadsheets or other highly manual processes. Participants also noted that institutions *"treat the web as one soup of URL and time that is the same for everyone,"* meaning that the URL and the time-stamp are over-relied upon for archivist navigation. Many participants shared parallel experiences of maintaining handwritten or highly manual metadata structures for their archival systems: *"Why does a book have all this metadata, and it's infinitely more searchable, despite the fact that it's not a full text search?"*

- Difficulty in navigating and utilising large archives as a user. This is one area where web archives diverge from the practices of organisation and labelling inherent to physical archives. The user of the archive has to know a system more than what the system itself can present the archivist at the moment of the capture, and often user navigation requires advanced knowledge of desired material URL and capture date.

- Inability for donors and archivists to fully assess privacy and confidentiality risks in archival processes and potential exposure of personal information during user access. For example, participants described how donors expressed concerns over whether their personal data or social media accounts will be archived if they donate website files. This again signifies a lack of clear understanding of the boundaries of these systems, and how consent cannot be presupposed in any act of mass-crawl.

- Inability to fully understand or observe user access for purposes of maintaining archive safety or performance. Many participants admitted that they had very limited information - or in some cases no information whatsoever - about the use of the archive beyond user self-reporting. This was especially true in archives that relied on large scale third parties for support: *"We get limited statistics on how many people clicked on our material by our provider. We don't know a lot and we don't know how they're getting it. We're flying blind. We don't understand what our users are searching for."*

- In examples where archives are built upon third-party platforms (such as re-purposed social media as the container for archives), participants recounted the additional layer of complexity around external content moderation policy, which is often opaque and arbitrary, and vulnerable to abuse via targeted campaigns of harassment. One participant - a curator of a large cultural archive maintained in public on social media - described an example:

> *"For us, curation is a delicate balancing act. A volunteer sent us 18+ R-rated imagery of herself, and the images were really beautiful. We had a discussion amongst ourselves, to determine the risk it would get taken down, and in the worst case, take our entire archive with it via account termination. Obviously maintaining our own website would allow us to have a bit more freedom with what we can post, but at the expense of community participation."*

In each case, complexity itself is not the primary threat to archive practice and access. Instead, the combination of complexities amplified deep uncertainty felt by participants, donors, administrators, users and subjects of archiving. Attempts to remove complexity from archival practice risks a narrowing of capability and erasure of archive nuance via opinionated design decisions in tooling. Participants expressed deep curiosity and fascination alongside (or in spite of) the complexity they encounter in their practice, but opaque tools, underdeveloped reporting systems, lack of standardisation and deeply inferior navigation and taxonomical systems make complexity insurmountable for many practitioners.

Anxieties and threats

Participants were asked to describe real or theoretical threats to their personal safety as the result of their archival practice, and to assess threats to their archives. It is not the purpose of this chapter to propose solutions, the nuance of safety and the potential kinetics of a rapidly changing situation involving personal safety deserve further careful consideration before a response is proposed or implemented.

Amongst participants, some common anxieties included:

- Archivist burnout caused through resourcing, complexity and high levels of emotional labour;
- The perceived broader public interpretation of archiving as feminised work;
- Increase of workload due to a combination of declining global stability and accelerating levels of digitisation in societies;
- Feelings of precarity due to limited financial resources in the practice, combined with archive complexity generating unpaid busywork - smaller institutions in particular face intense financial issues to operate web archiving initiatives: *"some national libraries, they're really struggling;"*
- Traumatic material crawled and inserted into archives either intentionally or accidentally (for example unflagged Terms of Service-breaking content captured from social media);
- Misinformation and the psychological effects of propaganda;
- Alienation due to the solitary condition of archiving, amplified by the digital process;
- Ongoing and unaddressed tension cultivated by an understanding of the potential of weaponising an archive: *"one person's archive is another person's police dossier;"*
- Anxieties around accidental leaking of personal information through login requirements, or tools 'archiving' credentials and login details, leading to being targeted or otherwise exposed as contributing to an archive;
- Local legality or criminal culpability with regards to archived material or the participation in custodianship of content;
- Hierarchical pressures on employees within archival institutions around the archiving of unfolding political struggles.

The position of practitioners plays a major role in how they perceive threats to themselves and their archives. Institutions - such as well-funded initiatives within government and academia - tend to manage and relate to archives from the top down, relying on large-scale automated crawling. In doing so, these practices de-prioritise on-the-ground inquiry. When combined with context collapse[45] and the flattening nature of digital interfaces, these practices are at risk of being severed from the subjects they archive. The same is true for tool-makers.

Problematic assumptions around institutional trust are also a source for real unmitigated risk. When offered a provocation around archive weaponisation via data breach, participants - particularly individual practitioners and other representatives from minority communities - counter-proposed that hacking by nefarious actors could be less dangerous that the law itself:

> *"The type of malicious actions I worry about aren't technology based. I'm worried about social legal mechanisms like SESTA/FOSTA or the targeting of queer communities. I constantly think about the chilling effect that can be created on people through legal mechanisms directed at technology."*

Hostile legislation leading to the targeted deplatforming of minorities is a prime use case for digital archiving, yet the fact that material has been removed due to legal demands, or weaponised as part of a targeted investigation, highlights the real harm that has been now replicated into an archive.

When considering the physical safety of an archive, participants often described a preference to *"preserve incorrectly that lose it all."* Preserving incorrectly is a useful euphemism that can imply some corrupted data at best, mortal danger for targeted communities at worst. The hegemony around definitions of preservation, accuracy and loss belong to a certain mental model that is alien to Indigenous or non-Western societies. Within this context, mass-harvest automated archive initiatives can be considered as driven by the anxiety and the paranoia that something could escape the Western epistemological gaze. Amongst

[45] boyd, danah. 'How "Context Collapse" Was Coined: My Recollection'. *Apophenia* (blog), 12 August 2013. https://www.zephoria.org/thoughts/archives/2013/12/08/coining-context-collapse.html.

some Western-identifying participants, the anxiety of data loss and the desire for internet permanence was strong: *"I think there is this anxiety because of how much digital material has been lost."*

For Indigenous, diaspora and minority communities and practitioners, the design of archival systems and the way material and experiences are saved and encoded will affect they way they are treated, perceived, used and exploited. The threats are existential and cannot be overemphasised. The leveraging of archives to target, disempower or dehumanise communities is a well understood issue with physical archives, but the digital equivalent makes it easier to automatically map relations, presences, and political orientations. Web archiving must grapple with its problematic ties to the technology, approaches, and infrastructure supplied by big tech, which has a long history of partnership with oppressors.

Archivists and archive tools also bear responsibility. One participant expressed frustration at the expectation that Indigenous communities participate in frameworks and policies developed by researchers, archivists, and toolmakers that facilitate existential danger. During the research, other practitioners expressed problematic perspectives that supported these frustrations and confirmed these concerns.

Within tools, the dangers to vulnerable communities exist in redaction, custodianship, the weaponisation of archive integrity, and the re-purposing of archived material for violence. Redaction isn't just a matter of scrubbing sensitive material: data de-anonymisation is a widely practised technique, and targeted redactions themselves can provide clues to identities and details.

Community archivists were more likely to use novel means of producing archives, such as re-purposing social media as archive platforms and cultivating collective archive contribution. This leads to a sense of data precarity bound to arbitrary and well-documented inconsistent Content Moderation policies:

> *"For us, there are loads of consequences of getting reported. Our archive can get shadow banned without notice from the platform. Other content can be removed, either through algorithms or mass campaigns of reporting. It's something that we think about now that we didn't when we started. Because our platform is so big, the last thing you want is for our archive to be shut down."*

It is incumbent on web archivists to protect the people featured in the archive. Archivists must carefully think about restricting access to law enforcement, deciding what to publish or not, what to redact, when to destroy and what the implications of cryptographic integrity mean when archives are re-purposed for violence. Archivists and tool builders must acknowledge their role as a mediation between coercive powers and the people they rely on for their archives. To do this responsibly, practitioners must engage in ongoing threat assessment and proactive harm mitigation in conversation with the people featured in the archive.

Ethics & colonialism

Woven through both the landscape review and participant interviews were themes around ethics. No doubt the broader political realities of current events remains a major driver of increased sensitivity around these themes for participants. Ethics broadly covered issues around consent, legislation, trauma, morality, colonialism and liability. Participants were invited to reflect on these issues within the context of their experiences and consider the manifestation of ethics in their practice. Participants were also invited to reflect on the ethics of the field more broadly.

Many participants from a variety of demographics found consensus in the importance of archiving with a degree of ethical professionalism and sensitivity. One example is the cross-discipline ongoing discussion around the issue of consent, and the complexities tied to archival and material take downs. As described later in this section, there are many examples where the discussions of consent frustratingly do not manifest in practice, or are overridden by competing desires. That it was discussed broadly by almost all participants indicates an ongoing shift in ethical debate within the practice.

Participants also broadly agreed that archives should be editorialised as little as possible beyond the scoping of the archive project. Whether archiving political events, or performing large scale dragnet collections of an entire country’s posting history, this frequently leads to practitioners grappling with the underbelly of societies: racism, violence, death, sexual assault, pedophilia, gore, harassment, malware, etc. Despite a shared intersectional resistance by archivists to immediately and dogmatically destroy instances in which such content is encountered, there are differing opinions on the gravity of this material from the perspective of an archivist. This has consequences for potential exposure, and new risk for decentralised proposals for archival custodianship.

When tools are designed, they materialise with a baseline set of ethics “baked” into them. Yet, the automation of the crawling process itself does not stop the ethical urge to manually, visually confront the materials. Speaking with participants that were exposed to traumatic content at various points in user-experience flows within common automation tools, a theory of interface trauma has emerged: the design decisions of tool makers as to how and when to display traumatic content as part of an automated archival process may have an influence on the levels of harm experienced by a practitioner. Further research is needed to understand this phenomenon.

One key issue that emerged was the rejection of the impermanent internet and related paranoia of the potential disappearance of material from the web. For some, the belief of data permanence was an ideology that trumped issues of consent, invasion of privacy or other ethical considerations. In these cases, the urgent need to archive was the justification to acting bluntly in ethically grey circumstances. We consider this a ‘*collect now, justify later*’ approach to archival practice, a philosophy that has significant second and third order cultural and legal implications for custodianship, decentralisation, and the material safety of the subjects of archival practice.

This philosophy was expressed more frequently in well-funded initiatives within a cultural, state or academic structures and strong mandates to archive. It represents a flattened yet uncompromising understanding that was widespread. Some manifestations of this ethical framework included:

- Collapsing the definition of privilege and power by assuming the archive was operating in an egalitarian society: an example given by a participant explained that the archival treatment of a politician should be handled no differently than a private citizen on social media.

- Justification of potentially or knowingly disregarding consent or privacy for a subject of archiving, by describing their use of the internet as a form of social contract towards data permanence, and minimising objections to this as naïveté or bad faith on behalf of the user: *“our policy is that if it’s on the web, it’s public.”*

- The belief that anti-censorship or pursuit of truth should be upheld above other concerns. This was especially true in cases where subjects of archive were considered powerful and potentially unaccountable - such as politicians protected by political structures and laws. In these cases, additional trust was placed in media structures to correct the record or participate in accountability.

- A desire to archive indiscriminately with the goal of capturing profound material for analysis or posterity - particularly during urgent, real time events: *“We had no guarantee that Twitter would still exist in 20 years, so we felt the need to grab it right now. The utterances of today are so different, to collect the doodles that people drew in the 1960s and 1970s at political meetings, an archivist would physically have to approach people for that material. Nowadays, this can be done remotely.”*

- Total faith in the institution to steward real highly personal or sensitive archived material and maintain governance in an ‘impartial’ or ‘moral’ way, without accounting for change in organisational leadership.

We believe that the positioning of practitioners in these institutions, combined with user detachment cultivated through abstractions in digital interfaces and Western hegemonic ideals about archiving leads to a risk of practitioners being severed from the people and things they archive. This manifests as a lack of inquiry on the ground and little effort to establish contact or dialogue with subjects of archiving. This is a sort of numbness for practitioners that is further cultivated by the abstractions and complexities of automated and mass-harvest archival tools.

Whether archiving a government’s web presence or the entire internet output of a nation, this institutional numbness helps to create archives that are potentially very dangerous - conceptually through narrow Eurocentric taxonomies that do not get challenged; legally and psychologically through the large-scale collection and downplaying of the harms in encountering illegal or harmful content in an archive; or politically and physically through disempowerment of subjects of archiving, blowback from powerful archive targets or threats to physical safety. Even in cases where practitioners understand that the universities and other governmental institutions were part of the problem, participants pointed to curation and custodianship policy as a directive for grappling with ethics and threats.

Conversely, the testimonies of practitioners from Indigenous communities or minority groups provided significant decolonialist criticism of archival practice, as well as promising threads for recommendations for tool builders and modification of practice. The concept of ownership or control over a cultural artefact is foreign for many targeted communities, which makes on-the-ground negotiation with institutions problematic. This combines with the very lexicon of archiving (such as the term ‘capture’ being derived from the literal physical removal of people, wildlife and artefacts by colonisers) through to descriptions, labels, and taxonomy being defined according to the Western gaze.[46] All of this is baked into all areas of web archiving, and narrows available perspective whilst creating ethical concerns.

“Come correct or don’t come at all”,[47] recounted a participant, reflecting on the tensions formed from ethical misalignment and power structures between vulnerable communities and institutions, and an inability practitioners to take care when engaging subjects of archiving:

> *“Community organisations tend to be leery and very wary of academic researchers coming in and collecting experience and expertise. We call it parachuting, it’s very transactional, taking up resources, without putting anything back into the community.”*

The ongoing interrogation of ethics remains a crucially important part of the critical study and refinement of archival practice, and the designers of archival tools exist in a unique position where the ethics of the tool-building team are the baseline from which the practitioner can operate. The designers inject their framework of the world, and the archivist user extends these ethics. The ethical and colonialist implications of what teams choose to build

[46] Matsuda, Shavonn-Haevyn. “Toward A Hawaiian Knowledge Organization System: A Survey On Access To Hawaiian Knowledge In Libraries And Archives”, August 2015. https://scholarspace.manoa.hawaii.edu/items/495d3950-3bb6-4b9f-9b8e-0314a94b584e

[47] Caswell, Michelle, Jennifer Douglas, June Chow, Rachael Bradshaw, Samip Mallick, Nivetha Karthikeyan, Bergis Jules, et al. ‘“Come Correct or Don’t Come at All:” Building More Equitable Relationships Between Archival Studies Scholars and Community Archives’, 2 December 2021. https://escholarship.org/uc/item/7v00k2qz.

will have direct consequences - both in influencing archival practice, and outcomes for integrity, safety and perspective within archives:

> *"An archivist does have an ethical responsibility to understand political realities on the ground, especially with regards to the communities or individuals that are creating this material that they want to archive. The West has this very individualistic mindset of 'I have rights, and I have self-determination, I have my responsibilities.' We see the individual self's speech and social interaction that can make the self accountable. This runs counter to many Indigenous philosophies that define responsibilities and obligations beyond the self, to the people in your community, to the land, to the world that you live on. I think archivists have an ethical obligation to understand that they are not passive. They are in an active role with regards to interacting with the communities that they're trying to document and help preserve in some way, shape or form, they have a responsibility to understand that community's needs, to listen to its concerns, and not to exploit it."*

Resilience, custodianship & integrity

Maintaining access to and the safekeeping of an archive is deeply interlinked. Resilience refers to data safety, archive rot, and likelihood of destruction. Custodianship refers to the governance of the archive, and its social and legal relationship between the material and its guardians. Integrity refers to the ability for an archive to resist tampering. By definition, all archivists are concerned with both archive resilience, custodianship and integrity, regardless of whether they are personally responsible for either of these areas of archive practice.

To solve for archive destruction, the most common approach is a reliance on third parties, particularly AWS Cloud technology or large archival institutions that provide hosting or duplication services for smaller archival groups. These are chosen for their proven history in resilience, and because they offer bandwidth and storage very cheaply. Due to the size of archives, some institutions have no local copies, only relying on these third-parties. These relationships create a complex chain of custody that is heavily reliant on hegemonic private actors and generates a sense of precarity in some participants.

Newer archival projects are beginning to express concerns about the use of third party services, opting instead for their own grassroots implementations of replication. There is a lot of discussion about the use of decentralisation protocols to simplify the process of distributed archiving. Drawing on broader New Design Congress research into climate threats to data and network connectivity, and the potential for loss of access due to geopolitical events, it is unequivocally clear that alternative distribution models are developed and deployed to provide actual, usable alternatives that do not rely on the status quo of a client/vendor relationship.

Yet, as discussed extensively in this report, the current proposals for decentralisation do not account for the political realities of users.[48] Dark archives, illegal and sensitive content, network analysis by bad actors, and data permanence are all inadequately accounted for in the decentralised alternatives. Given the desire expressed by participants to maintain decentralised, resilient archive networks, if these issues are not solved, then the outcomes to archivists and subjects of archiving will be catastrophic. The prosecution, for example, of an archivist who inadvertently hosts a ‘dark archive’ seeded to them by an institution who is legally protected is an almost certain future event. Similarly, the use of decentralised, immutable archives on blockchains or other immutable or versioned systems will absolutely result in the non-consensual posting of personal information, or sensitive or illegal imagery, that is hard to take down - or functionally impossible in the case of blockchains - and disastrous for the victim.

Digital rot was most often encountered by participants through tool failures - especially resulting from changes to a platform targeted for archiving - and shifting obsolescence. Practitioners described the web browser itself as an evolving surface for digital rot, where bugs affect accurate reproduction of archived material, or capabilities are added or lost. Participants described the need for better reporting tools to help contend with data rot, and participants who were more comfortable with digital systems suggested alternatives such as virtual machines to further reduce data rot by bundling a known working environment with the archived material.

[48] Diehm, Cade. ‘This Is Fine: Optimism & Emergency in the P2P Network’. The New Design Congress, 16 July 2020. https://newdesigncongress.org/en/pub/this-is-fine.

For participants, custodianship was usually maintained by an institution but informed by practitioner scoping, curation and QA on the archive. Participants described how working within or with an institution provided a certain amount of logistical support and safety to varying degrees. However, archive custodians are more visible both internally and externally in institutions. The employment status of a practitioner often has a relationship to their safety in the role as a custodian.

Institutions often provide varying levels of support, but custodians must make decisions that affect the institution, and often the institutions first and foremost aim to preserve themselves and their abstraction - sometimes at the expense of employees and communities. The liability inherent in custodianship leads to a precarious relationship between practitioner and institution. Sometimes this scales depending on the size of the institution and the reputation of the practitioner. Because of archive complexity and issues around QA, navigation and reporting, the issues of liability are more serious for archivists than they should be.

At the same time, practitioners sometimes weren’t able to fully see the institutional power they wield when they confront communities, website owners, activists, etc. For participants who found themselves in these situations, the feeling was difficult to reconcile with:

> *“I wonder if it would be different if maybe a grassroots organisation contacted a site. I wonder if a site owner would feel differently either way about whether to allow that kind of access and copy being made.”*

Given that custodianship carries both a degree of power and legal responsibility, it is clear that small open-source/free software initiatives shouldn’t have to fully bear the responsibility of answering to the pressures of custodianship. In reflecting on this, one participant pondered:

> *“I don’t think this should be the responsibility of the small open-source tools, or small cultural institutions, or even bigger museums to fix this. But there should be concerted action. For example, the software preservation network is a very interesting initiative that also has a legal arm. And that was able, with a lot of lobbying work, to get software into the fair-use regulation in the United States.”*

The broader tool landscape

Archivists use a variety of tools for their work, and this is driven by strengths, limitation, speed, technical ability and user familiarity. Common tools listed by participants include:

- The Webrecorder suite of tools;
- Automated crawling tools such as Heritrix and Archive-it;
- Targeted archival tools, such as Wayback Machine;
- Command-line tools, like cURL;
- Legislative tools, such as Freedom of Information requests;
- Data takeout services from commercial digital platforms.

Although the landscape for tools is small, the differences between tools, the lack of standardisation of outputs, and the size of archives, all generate complexities for curation and the provision of access. Overall, tools are somewhat interoperable, but frequently fail at specific critical points.

Participants broadly described a preference for automatic crawling tools, a fact that is covered extensively throughout these findings. Participants overwhelmingly expressed a desire for tools that could generate metadata across different outputs from different tools for reporting and QA purposes.

Tools must cater for a small but booming practice, and accommodate multiple and diverse roles or use cases across the flow of practice for archiving – such as scoping, harvesting, curation and custodianship. The economic aspect of this issue is a complex one. Participants were very aware of the costs associated with tool development. It is clear that ongoing financial support is needed not just to maintain the current functionality of tools, but also for practices and use cases that aren't an immediate priority for the biggest actors and institutions, despite being of vital importance for the field of archiving. These tools – particularly the satellite open-source libraries they depend on – need to be financially supported so as to provide technical solutions for these issues, which otherwise remain critically unaddressed. In many cases, the tool landscape is seen by participants as underfunded, which creates anxiety around longevity and the fear of potential mismanagement of a tool project leading to the tool’s end-of-life.

Emulation also plays a role in the broader landscape of tools, particularly in replayability. A participant for instance highlighted the importance of emulation as a shortcut for displaying all sorts of archived material – especially obsolete material. However, emulation requires an even larger expenditure of labour and money for development and maintenance.

Some participants described novel uses of tools for archiving, such as social media platforms. They offered an easy-to-use interface that solved the issue of access while providing granular metrics on the audience and community, two issues that plague many other archival initiatives. Curators can easily interact with their community through polls, live streams, comments and direct messages, as well as curate the content along different categories through platform features. This repurposing of social media platforms creates a virtuous circle between the archivists and the community, and bypasses the linear and temporal form imposed by the algorithmic timeline of a social media platform.

This is not a call for migrating archives *en masse* to mainstream social platforms, or for blindly imitating their features. However, we believe that these examples offer a significant divergence from more mainstream archive designs, and shows what opportunity remains to develop tools that facility thoroughly different, vibrant approaches to archiving.

Webrecorder & WACZ

When asked specific questions about Webrecorder and the WACZ file format, participants expressed an overall excellent opinion of the tools developed by Webrecorder. Almost all participants were using Webrecorder tools regularly, had in the past, or planned to in a future archive effort. They are eager to participate, collaborate, finance and support, seeing the project as a vital component of web archival practice.

Participants listed the tooling portability, community cultivation and accuracy of output as main key strengths of the Webrecorder project. At the same time, the manual nature of some of the tools in the Webrecorder suite set the expectations of the overall Webrecorder project. When struggling with the stability or complexity of an archive effort, participants were particularly sensitive to the quirks or user-experience pain points of the project.

Participant comments included:

> *“Webrecorder was a real gift to my work.”*

> *“I’m always grateful to the Webrecorder team for their help.”*

> *“Webrecorder is excellent at recording social media content, like Facebook, in a very user-friendly way.”*

> *“Navigating the WACZ file feels like browsing the web.”*

> *“Webrecorder captures audio materials very well.”*

> *“Webrecorder is good at recording dynamic and complex pages compared to other tools.”*

> *“Independent web archivists owe a lot to initiatives like Webrecorder that allows them to generate an income. Webrecorder offers the possibility to create archives outside the structures of institutions, and helps autonomous and community archiving initiatives.”*

Participants described cases of Webrecorder being deployed quickly to archive fast-moving events. This was made possible in part because of Webrecorder’s local-first browser extension offerings, that allow practitioners to quickly recruit and orchestrate a team of archivists with relatively little preparation and training:

> *“The Webrecorder tools help us with urgent archiving initiative, and when working under this pressure with volunteers, we found it to be very user friendly. You can use it as a browser extension! I was very pleased when our team found it.”*

Many participants highlighted the importance of the archive as a deliverable or package, and the ability of Webrecorder and WACZ to deliver portability was a primary driver for its deployment in practice.

When evaluating participant criticisms of Webrecorder, common experiences emerged around key parts of the overall design of Webrecorder and WACZ, and the project's user experience *in parallel or comparison* with other tools. For example, some participants described their difficulties via a comparison framing of the user experience of Webrecorder against the user experience of Archive-It. User experience criticism tended to extend from Webrecorder's focus on manual over automated processes, and it was clear from feedback that many practitioners familiar with Webrecorder use it in a manual fashion, fine-tuning archives where other tools and automated processes have failed.

Participants described WACZ as a format with a lot of potential, with some noticeable limitations. Many practitioners described issues of interoperability. Practitioners described barriers or potential metadata discarding when converting WACZ files between different tools or absorbing them into broader archives. Some described situations where, when a WACZ file was transferred manually between computers, the transfer rendered the archive as incomplete. It was not clear from the interviews whether these experiences were due to the WACZ format, or external factors. Participants recognised that interoperability and collaborations between open-source tools would likely solve these issues, and that many of the challenges they were facing were due to fragmentation of vision within the broader tool landscape: *"Everybody's making tools that are like, maybe going to 10%-90% workable stuff."*

Multiple participants explained that they felt that some of the components of the format were *"not readily accessible,"* and they struggled to interpret the format. To practitioners, despite solid user experience in the Webrecorder tools, WACZ remains somewhat a black box:

> *"If we could just parse out individual pages, that would be so useful. I think for future researchers, making it easily interoperable, like being able to define it in rows on a spreadsheet, where we can identify capture date, a file size, a specific URL for a work file, that a user could click on, and say 'Hey, I want this act, I want this file,' and then we could just easily transfer that specific file, that*

> *way we preserve privacy of the individual, or there's some steps so that it's not just completely open to the public."*

Participants expressed a desire for greater management of permissions within an archive, allowing archivists to design settings around granular access, curation and broader control over the collection, processing and display of content and individual files within the WACZ file. Practitioners considered format-level instructions that could standardise an archive, allowing the technical curation of what is archived, omitted, or redacted. After a crawl, participants expressed a desire for additional tools around reporting, navigation, and visualisation of a WACZ archive's information structure. The ability to plan and deploy detailed reporting systems within the Webrecorder ecosystem would allow archivists keep track of their progress and recognise potential partial failures of an archive.

Synthesis

Beyond this standard template that most archivists follow, the research surfaced important contradictions within the act of archiving, or rather, as we have alluded to throughout this report, the various facets of the Foucauldian discourse[49] surrounding the act of archiving. Sometimes the value of the materials is defined after the act of archiving, a vision which necessarily implicates the act of curation and custodianship, engaging the responsibility of the archivists. Sometimes, there's an *a priori* assumption that some materials will be of potential value. This drives a need to capture them immediately, to be sorted out later. Materials are thus imbued with value at opposed steps of the process, often by the same individuals, a situation which seems to arise from three distinct challenges:

1. The chaotic nature of the surfacing of important materials which can be available at the moment of their creation (official government declaration) or rediscovered (under a desk, in a basement, on a random hard drive).

[49] The analysis of discourse in Michel Foucault's work focuses on the power relationships expressed through, and the truth-value imparted by, language and behaviour. Such an analysis encompasses the mesh generated by the legal, the epistemological, the vernacular, the literary and other forms besides, particularly the way they blend into a conceptual grid meant to render the world intelligible.

2. The difference between active preservation (technical examination and selection, conservation, methods of storage in correct environments, housekeeping and collection control procedures, technical restoration, rejuvenation, duplication and quality control) and passive preservation (synonymous with ‘storage,’ keeping the material in an ideal environment and not subjecting it to any mechanical risk through use).

3. The friction between the traditional framework of analogue archives and the technological potentials unleashed by digital technologies.

While this is not constitutive of a fatal mistake from the field, it nevertheless seems clear that digital archiving, as a unified practice, doesn’t seem fully aware about what is motivating its actions. As documented earlier, some web archive initiatives see as vital the recording of comments, likes and other such metrics associated to social media posts, while others only save the posts themselves. In this case, the difference with a traditional analogue materials - the fact that now archived material come with a sort of popularity poll - is acknowledged, but the consequences aren’t clearly framed. While a few archivists have ascertained the potential risks of capturing such extra content, the field appears split on the question. As more and more content and expression move to social media, eg. histories of struggle, working-class and targeted communities organisation, etc., archivists perceive as vital the preservation and capture of this history, while the social media model exposes followers, their identities, personal information, leading to context collapse and outright weaponised design.

What we’ve so far termed *mass-harvest*, the act of focusing the practice of archiving around technological mass-crawls, is justified by practitioners in the case of un- and re-scoped projects, projects that are too massive to be properly defined and given strict boundaries. Such projects seem expressly designed - consciously or unconsciously - to be so massive as to disregard any actual scoping. This tendency projects its gaze as far as the legal limits allow, be they GDPR, an official government remit, copyrights or the common law. Forgetting Montesquieu’s famous maxim that “*a thing is not just because it is law,*”[50] a lot of institutions have developed extremely elaborate legal access

[50] Montesquieu, Charles de Secondat, Anne M. Cohler, Basia Carolyn Miller, and Harold Samuel Stone.

and display frameworks that surround the act of mass-harvest, with a para-legal discourse meant to, in the last analysis, justify the practice:

> *“Our permissions are based on crawling and on display. One of them is based on the copyright laws of the country that the site originates from. And another is based on copyright laws of the US. So it’s based on the country and the category of the website, like journalism, or government or non profit. And we either can crawl with no notice and display with no notice. Or we might need to notify in one way or another situation, or we might have to ask for explicit permission. Across the board, all creative expression, websites or exhibition websites, we have to ask permission to crawl and to display.”*

> *“We have a one year embargo on all content that’s crawled, so it just won’t display content to the public, or to whoever has access to that access point or that material. And then the way that it functions is kind of on a rolling basis, each day, a little more content will pop up if you know what to look for technically.”*

This legalist ethics dissolves for institutions like universities when they have to deal with their students’ data. In this case, the utmost care is taken, a mindfulness that is not always extended to the general public. This intricate legal sophistication thus stands in stark contrast with the cavalier disregard for any ethical concern over the archiving of people’s expression. The commonly held impression that social media are private spaces, even on public mode, is disregarded.[51]

The issue is exacerbated by the fact that, as has been made abundantly clear throughout this report, the mental models of the information economy, already little understood by even its proponents, are profoundly alien for some Native and Indigenous communities, such as the permanence of data. In some case, this permanence enshrines discords between groups that are the results of the manipulations, policies and interfaces social media platforms deploy:

The Spirit of the Laws. Cambridge Texts in the History of Political Thought. Cambridge ; New York: Cambridge University Press, 1989.

[51] Or that social media aren’t even on the internet, as millions of Facebook users seem to believe. See: Mirani, Leo. ‘Millions of Facebook Users Have No Idea They’re Using the Internet’. Quartz, 9 February 2015. https://qz.com/333313/milliions-of-facebook-users-have-no-idea-theyre-using-the-internet/.

> *"For a lot of people, the first screens that they've seen have actually been smartphones. Some haven't even seen a TV and they've never used a laptop. Facebook is really great for connecting with people far away, and really terrible for communities because it's causing really, really bad gossip and conflict. It creates distance within villages, which never used to exist."*

Mass-harvest seems then, for some, to also be predicated upon the belief that it is possible to perfectly simulate an older version of the web. Many participants have criticised this approach: *"[within] this enormous mass of data, […] you can traverse the web as though you were there in 1997,"* despite the fact that archiving, *"isn't a neutral process, but it is always driven by an actor, the archivist who is acting on certain ideas or a mission or is making mistakes or all these kinds of things."*

Oddly reminiscent to the logic of Big Data, mass-harvest remains a new tendency arising from its technical possibility, yet born from a long ideological heritage. We can uncover the roots of Big Data in the evolution of the statistical discipline throughout the 18th and 19th century, developed to 'rationally' manage whole populations. In his analysis of new technologies of power and the relationship between national statistics and the act of governing, French philosopher Michel Foucault argues that statistics went on to reveal *"that population has its own regularities, its own rate of deaths and diseases, its cycles of scarcity, etc.; statistics shows also that phenomena that are irreducible to those of the family, such as epidemics, endemic levels of mortality, ascending spirals of labour and wealth; lastly it shows that, through its shifts, customs, activities, etc., population has specific economic effects: statistics [make] it possible to quantify these specific phenomena of population […]"* and thus:

> *"[…] it is the population itself on which government will act either directly through large-scale campaigns, or indirectly through techniques that will make possible, without the full awareness of the people, the stimulation of birth rates, the directing of the flow of population into certain regions or activities, etc."*[52]

[52] Burchell, Graham, and Michel Foucault, eds. *The Foucault Effect: Studies in Governmentality; with*

Yet where statistics used to be employed in a predefined context as part of the scientific method, their subsequent rabid deployment has led to the massive harvests of anonymised data that purports to make patterns emerge through the use of algorithms.[53] In the context of web archiving, this creates, as one participant critically dubbed it, *“fantastic visions of what could be possible,”* where the focus on mass-harvesting over custodianship and curation can have the potential of moving the goal posts of the so-called neutrality of archives. A participant noted: *“My professional opinion is that the context around which that information is being collected is very important,”* and the Big Data approach to archiving purposefully decontextualises the data harvested.

The justifications for mass-harvesting are often self-contradictory, either with themselves or the broader science of archiving. Scientific terms become floating signifiers: *“Our attitude is that we need a representative sample. We try and get as much as we can.”* In this case, the concept of representative sample is posited upon the methodical selection of a subset of a population designed to accurately reflect the characteristics of a larger group. As a result, the representative sample is turned on its head and seen as instead arising from the indiscriminate capture straight from the larger group. It is difficult to discern how this approach respects the key concepts of significant properties, graceful degradation, and original order, which are supposed to establish archiving as a science. It is also incompatible with an important facet of archiving that comes up again and again - that of the multiplicity of archives that, autonomously and network-like, can cover each others’ epistemological gaps by their very number and diversity:

> *“In the world of web archiving for this organisation, even if we miss a really important picture, like Gloria Steinem talking to Jane Fonda or the head of the CIA, it’s okay. It will probably come out somewhere else.”*

Two Lectures and an Interwiew with Michel Foucault. London [u.a]: Harvester Wheatsheaf, 1991. See also: Foucault, Michel, François Ewald, and Alessandro Fontana. *Security, Territory, Population: Lectures at the Collège de France, 1977–1978*. Edited by Michel Senellart. Translated by Graham Burchell. 1. Picador ed. Lectures at the Collège de France. New York, NY: Picador, 2009. and Foucault, Michel, Michel Senellart, and Michel Foucault. *The Birth of Biopolitics: Lectures at the Collège de France, 1978–79*. 1st pbk ed., [Repr.]. Lectures at the Collège de France. New York: Picador, 2010.

[53] Rouvroy, Antoinette, and Thomas Berns. ‘Gouvernementalité Algorithmique et Perspectives d’émancipation: Le Disparate Comme Condition d’individuation Par La Relation?’ *Réseaux* n° 177, no. 1 (1 April 2013): 163–96. https://doi.org/10.3917/res.177.0163.

> *“We’d like to get more conservative women materials, because most of the women we collect were radicals or trying to work against the system, and they weren’t really conservative. We had a whole group of women in the United States whose papers are not really collected. [Some are] suspicious of Harvard and suspicious of our library, and that’s fine. There are other archives in the country and the world.”*

This vision of representative sample and accuracy also challenges the important notions of curation and custodianship as an act of multiplicity:

> *“You’re of course capturing maybe different things from the web; or depending on what is your idea of the web, you’re also capturing probably other materials; then someone has another idea. These parts contribute to the creation of the archival material, making it accessible and maintaining it over time. That’s the role of a custodian.”*

The epistemological weight of an indiscriminate archiving and thus its socio-political implications are at worst ignored, at best rely on a fantasied vision of the press to correct errors that could have been captured:

> *“If it’s public, I don’t have any issue. We don’t have really the time to do corrections. If [the archive material] is of enough public interest, it will come up in the media. Then we will have [the correction from the press] in the selective harvest.”*

Beyond a flagrant naïveté regarding the economical dynamics that propel so much of the contemporary news industry (attention economy, quality of journalism quantified by visits and clicks, etc), where a correction often attracts a lot less attention than tabloid headlines, this approach to accuracy could promote a vicious circle where the errors captured during a mass-harvest will be corrected by a further increase in the power of harvesting. This justification, and the fact that mass-harvest troves are so unwieldy that nobody can know for sure whether laws are respected or not, cohere in a monolithic, *scientistic* view of archiving where the sum total of knowledge must necessarily be centralised[54] within a single repository.

[54] Whose shape can also be decentralised and distributed. It doesn’t affect the centralised nature of the control exerted.

Other issues involved the expected longevity of a web archive. While some voiced the temporality of the archive in precise terms, the interviews surfaced a conceptual fog surrounding the length of time the archive should be kept. This combines awkwardly with the indeterminacy of the *“fantastic visions of what could be possible”* with mass-harvest, where the archive ends up projected ad infinitum without any understanding of the consequences, reverting to undefined metaphysical archival virtues. Some participants were quick to address this issue:

> *“I think the idea of forever is totally silly. And no one who uses the word forever in an archive in context can be taken seriously. The question is: what do you mean by forever? That it’s always exactly as you see today? You can’t even predict what’s happening. Next, Apple is making laptop with circular screens or whatever, and then everyone is basically screwed. There has to be a strategy for next steps that could happen. And these changes are also sometimes small, like an update that prevents auto-playing sound.”*

It appeared that frequently, the position of keeping something forever was cogent with the idea of control over that thing. Conversely, the individuals, groups or communities subjected to that archiving are not granted control. This is a very visceral example of a power structure that has existed for a lot longer than the internet has been around for: the idea of archives, captured material, and their collection within a centralised space, where agency for the subject of epistemological power is lost.[55]

The conversation and the discourse is slowly shifting. From a focus on technical and staffing matters, more and more are trying to address how to better support and treat communities that archives depends on:

[55] On the role of museums in the colonial project, see for instance Anderson, Benedict. *Imagined Communities, Reflections on the Origin and Spread of Nationalism. Verso Books*. August 2016.

> *"I feel like a lot of the thinking up until the past few years has been around technical and logistical considerations, or considerations of staffing and roles. Now we're finally having a conversation around doing this work while supporting and adding rather than taking from these communities that we're trying to document. Actually as a benefit rather than subtraction?"*

Web archivists have this desire to be more inclusive, which often is expressed by asking for contributions from Indigenous, diaspora, minority and non-Western communities:

> *"We wanted to develop [this framework]: we'll define the scope, then have a forum where people start submitting URLs to a form. We can then assess if we need to archive your work. That is the hope, but we're not there yet."*

This well-meaning gesture, while contributing to at least a conversation between institutions and socio-political minorities, could also lead to what Yuk Hui has dubbed the crowd-sourcing of central archives:

> *"Even though today some of these institutions implement 'open policies,' these are more or less always strategies of crowd sourcing under the name of the humanities or digital humanities in order to reinforce centralization. Public participation is still restricted to a minimum level, akin the relation between tourists and monuments."*[56]

This results in externalising costs on already economically deprived group, while the decision over what is worth archiving is kept within the institutional core, what historian Partha Chatterjee described as the *"instituted knowledge of society, as it exists in recorded history,"* that is to say *"the knowledge obtained by the dominant classes in their exercise of power. The dominated, by virtue of their very powerlessness, have no means of recording their knowledge within those instituted processes, except as an object of the exercise of power."*[57] Despite the genuine well-meaning intentions of the individual

[56] Hui, Yuk. 'A Contribution to the Political Economy of Personal Archives'. Edited by Ganaele Langlois, Joanna Redden, and Greg Elmer. *Compromised Data: From Social Media to Big Data*, 2015, 226–46. https://doi.org/10.5040/9781501306549.

[57] Chatterjee, Partha. *The Nation and Its Fragments: Colonial and Postcolonial Histories*. Princeton Studies in Culture/Power/History. Princeton, N.J: Princeton University Press, 1993.

employees of institutions, this consolidates into a non-committal institutional attitude that derives status and capital from the exploitation of socio-political minorities - even if these gains are not redirected to the archival department of the institutions! This in turn creates a feeling of a zero-sum game between archivists and communities:

> *“It would be easy to say institutional archives owe it to community archives to share resources or share tools. But those institutional archives still feel stretched and overtaxed, and they don’t have enough resources. It feels like a zero-sum game, where archives are undervalued. If archival practice was more respected, maybe there would be more resources.”*

It remains difficult to not perceive how institutions, as employers, end up playing their employees against communities, so as to further the exploitation of both. It nevertheless strikes us that being an archivist with a stretched budget is still a much more enviable position than, for instance, a Native Hawaiian whose water source has been poisoned by US military occupation.[58] It is incumbent upon archivists to discard this false consciousness, and articulate an intersectional class-power with the communities they capture. ✳

[58] Treisman, Rachel. ‘Thousands displaced from Oahu military base due to contamination in Navy water system’. *NPR*. Accessed 14 November 2022. https://www.npr.org/2021/12/15/1064514935/water-contamination-hawaii

V. Recommendations

Practical and technical recommendations

1. **Archiving must be reimagined by taking into account the new potentials, the challenges and risks of the digital:** new ways of analysing, viewing, and navigating archives, based on imagescapes and soundscapes, must be explored, to cater for future ways of navigating the web. Conversely, digital paradigms must be duly interrogated, to make sure that they do not enshrine exploitative structures of thoughts and practices.

2. **The place and influence of the archivist within the archive must be critically assessed:** some participants suggested that archivists should be identified in some ways (not necessarily individually or by name, ie. the team, the broader structure, the institution), in order to understand the context of the archival project and the potential biases. However, the utmost care must be taken in curtailing the disclosure of personal information.

3. **New models for permissions and access need to be developed to further public access while restraining bad actors and surveillance policing:** creating D-spaces and virtual reading rooms to allow and control the access can prevent bad actors from accessing the archive. This could also prevent the communities who are archived to easily access their materials, should the policy of the institutions change.

4. **Archival tooling projects should review and research preservation benchmarking tools, and consider designing new shared benchmarking standards:** *“such as the Open Archival Information System (OAIS protocol), the Trustworthy Digital Repository (TDR) certification, the Data Seal of Approval.”*

5. **Reassess the role of scoping in digital archive projects:** stuck between the rock of intangible scoping - so lenient as to allow mass-harvesting - and the hard place of legalist remits - backed by entrenched power structures -, digital and web archives have an understandably hard time laying down a discipline-wide scientific and ethical method of curtailing the grasp of their data capture. We however remain optimistic that,

through the genuine ethical concerns and the scientific rigour exhibited by a majority in the discipline, digital and web archiving can become a bona-fide autonomous epistemological endeavour as free as possible from the sins of the past. This can be achieved through the simple use of the archival concepts of *graceful degradation*, *significant properties* and *original order*. The first will force the creation of smaller, more focused and resilient archive collections. The second will highlight the focus and the bias of an archive. The third, together with the subjective caveats of the two first concepts, generates a drive for objectivity tampered with the signposting of inescapable subjectivity.

6. **Do not deploy decentralisation until protocol designers have grappled with information and network risk. Deploy decentralisation while always keeping in mind its Janus-faced nature:** the LOCKSS approach of selective collection with permission from the publisher, distributed storage, and restricted dissemination contrasts with, for example, the Internet Archive’s approach of omnivorous collection without permission from the publisher, centralised storage, and unrestricted dissemination. The LOCKSS system is far smaller, but it can preserve subscription materials to which the Internet Archive has no access. As one participant recounts: *“The Lots Of Copies Keep Stuff Safe (LOCKSS) model, for some archive organisations, is affordable. It’s about decentralisation with institutions that are part of a programme, and you’re storing other folks data encrypted on your system, and everyone else has your data encrypted on their system. It’s like a kind of feed-on net archiving. And if you delete your stuff, it’s also getting deleted everywhere else. And if not, people cannot read it. Because you are holding the key. I think that’s an interesting approach.”*

Decolonising archives as an ethical imperative

It cannot be overemphasised that non-Western communities subject to colonialism are in a state of severe destabilisation, having already bore the brunt of political violence and capitalist-wrought ecological decline far before the Western powers’ dim awareness of the climate crisis. These communities need complete controls over their maps, their representations, their information and their narratives. **Communities and social minority groups must own their own data**

and archives, which includes infrastructural control such as servers, interfaces and capital.

As has been made clear throughout this report, **the mental models of the Western archiving discipline do not account for Indigenous perspectives**. These mental models are baked into digital and web archiving through the very engineering and design of digital systems. **Alongside having non-Western contributors lead collaborations**, research should be conducted by design teams on how to cultivate greater awareness of the potential unintended consequences of the practice of web archiving. There is a significant opportunity to continue developing alternative first principles for digital systems, and many of these should be Indigenous led. **Redefining ownership, connectivity and permanence beyond their Eurocentric status-quo should be a priority for new and evolving archive tools.** A lot of care needs to be taken by tool designers to cultivate archive practices that minimise the potential threats inherent in the power imbalance of Indigenous user bases.

This report demonstrates that there is a clear realisation among practitioners that something should be done outside the realm of universities and established organisations. How to fund such a movement is a key question. **Yet if institutions are genuinely committed to the process of decolonisation, they must financially and technically help, and, alongside these resources, cede power and autonomy to Indigenous archive practices.** Any other course runs the risks of recuperation, exploitation and whitewashing – especially when institutions are directly linked to predatory investigation, exploitation and financial speculation targeting commons and Indigenous lands.[59]

[59] *“Over the past decade, Harvard University has become a major global investor in farmland and particularly invested in land in western Bahia and southern Piauí in the Cerrado region [displacing] landless peasants and traditional communities who considered the area public land that became irregularly usurped,”* See: https://www.ejatlas.org/conflict/harvards-land-speculation-and-displacement-of-landless-peasants-in-west-bahia-brazil. Accessed 14 November 2022;

“Ikea bought its land from an unlikely source: the Harvard University endowment, which snatched up Romanian property after a post-communist land restitution law left an antic privatization system in its wake, handing over half of the country’s public forestland to private interests. Starting in 2004, the university, using various shells and non-profit formations, began buying big with the help of a Romanian businessman, Dragos Lipan. A number of these holdings were fire sales of dubious restitution claims, and Harvard soon found itself in legal hot water.” ‘Sammon, Alexander. ‘Ikea’s Race for the Last of Europe’s Old-Growth Forest’. *The New Republic*. Accessed 14 November 2022. https://newrepublic.com/article/165245/ikea-romania-europe-old-growth-forest

For Webrecorder & WACZ

Webrecorder and its supporting funders should widen the scope of their near-term roadmap to account for the challenges raised in this report. Although much of the research applies to the entire field of web archiving, the issues raised will create obstacles for growth and adoption, while leaving the project vulnerable to the shortcomings and vulnerabilities in the practices of digital archiving, decentralisation and personal computing.

1. **Webrecorder must prioritise the visualisation of the successes or failures of a WACZ archive practice**. Overwhelmingly, participants requested tools and design models that allow for a more granular access and control over the content and individual files within the WACZ file, so as to select what is actually archived, what should be omitted, redacted, etc.

2. **Webrecorder should experiment with new methods of visually representing the topology of large-scaled WACZ archives**. The ability to navigate the archive in some sort of visual way, such as a slit-scan or other high-level visualisations, could be beneficial: *"And I kind of like, loop through that data visually and try to find out: where do I look for problems? And then I try to figure out: can we scope in the CSS page hosted on another domain, or the images hosted on another domain. And I tried to figure this out, but it's like I said, from 15,000 feet."*

3. **Consider the amplified cost to user trust when deciding upon feature removal.** The precarity of archive practice combined with the challenges around archive size creates a situation in which user trust may be more fragile than other projects. For example, many participants independently highlighted the disappearance of the list feature, as it helped build an index and was *"used to organise subsections of the site."* As Webrecorder continues to develop its roadmap, the project should take care with regards to the heightened stake that its users may feel towards the project.

4. **Multiple participant have raised the need for validation tools, or validation features.** Tools to validate both the accuracy of a collected archive and the integrity of material both featured highly in participant reflections: *"We need to validate, there is no good validation tools right now […] there actually needs to be some kind of validation tools*

for the newer technology, like Webrecorder, but nobody’s doing it at the moment. But we do the validation.” However, these features and workflows have significant potential for weaponised design and may impact existing archivist workflows.

5. **Develop workflows for archivists to monitor websites for updates.** Given the temporal nature of the web, participants reflected on the manual processes of returning to websites to check for updates. An assisted workflow for users would lower the load on practitioners, but risks the introduction of new complexities with regards to managing alerts, etc.

6. **Extend the supported materials that can be held in a WACZ file.** Materials can come in various formats, such as donated CDs, HTML drives on hard drives, objects found in basements, and significantly outdated code bases. Towards the end of this research, our team theoretically explored the potential for expanding WACZ files to support these adjacent materials, eg. by providing emulators or portable VMs that can help practitioners maintain obsolete material without high overhead.

7. **WACZ and Webrecorder interfaces should store and display additional metadata.** While practitioners deal with information overload, they equally struggle with context collapse, where the range of archive material becomes difficult to curate. At the same time, practitioners universally demonstrated an ability to build out workflows for navigation when given the tools to do so. One way that Webrecorder can support better archive practice is to increase the amount and scope of metadata as proposed in the WACZ specification,[60] and provide interfaces and endpoints for archivists to extend and integrate into their workflows.

8. **Engage with the digital archiving community to collectively define and develop new standards and new approaches for persistent identifier schemes**, such as Persistent Web Identifiers (PWID), Digital Object Identifier (DOI), Handle, Archival Resource Key (ARK), etc.

9. **Design workflows and interfaces for capturing and playing back streaming content.** With the increasing importance of streaming content, the need to archive streaming platforms is on the short-term horizon. Playback in

[60] ‘Webrecorder Specifications’, 8 November 2022. https://github.com/webrecorder/specs.

this context should include the synchronisation of other user interface elements, such as chat streams, audience reactions, etc.

For funders and future research opportunities

Funders working within the digital archival space must understand that the practice sits at an important crossroad. Through this research, we have identified numerous threats to the growth of alternative technologies and projects related to web and digital archiving, data storage and decentralisation. These are risks to the success and resilience of all archival projects and institutions, driven by the splintering political and ecological conditions and institutional-wide inadequate response to the shifting material conditions.

1. **Redefine Digitisation as a Material**: In order to meet its non-negotiable obligations to the climate and social justice, digitisation of the built environment must be considered not as a service, but as a material, similar to concrete or steel. We see digital systems in our cities as ephemeral because their services and interactions are mostly invisible, but the digital revolution depends on an exponential collection of computers, wires and purpose-built devices, a brittle network furnished by a bloody and unsustainable global supply chain whose carbon footprint, environmental impact and human cost dwarf all other materials of the built environment.

2. **Rethink digital identity paradigms**: Digital identity plays a significant role in modern computing. In almost all examples of socially-driven computing interactions between two or more users, digital identities rely on user profiles and unique identifiers to help users cognitively recognise and validate their intended recipient or collaborator.

   These digital identity paradigms carry significant risks. Attackers leverage digital identity systems, compromising accounts in order to impersonate trusted users, and manipulating targets into completing tasks or disclosing information. In these cases, a well designed digital identity that features a strong presentational layer is used as a disguise by the attacker. This disguise plays a critical role in the success of the attack, and example of weaponised design in which the curatorial user interface is used as a tool to convince a target of an attacker’s legitimacy, regardless of the strength of any cryptography.

This rationalist implementation digital identity systems – *I curate, therefore I am* – also assists attackers in the goal of building network graphs of interactions and relationships between individuals. In these instances, the self-curating nature of these identity systems by definition facilitate an ability to perform look-ups of individuals – a requirement for collaboration within a digital identity system but also an effective tool for the dragnet unmasking of users within a network and their observable relationships. When combined with web of trust techniques, the effectiveness of these systems as surveillance and mapping tools accelerates exponentially, where declaration of recognition between individuals serves as forensically sound data points for social network analysis for an attacker.

3. **Understand the brittle digital society**: As societies rush to embrace the promise of efficiency and capability of governance and commerce through digitisation, the effects of digital fragility are routinely ignored. As a result, societies with higher degrees of digitisation are beginning to demonstrate a set of brittle side effects that are complex, potentially paralysing and deeply profound. From the information and infrastructure warfare in Ukraine and Russia, to the over-reliance on algorithms and AI for decision-making, to the loss of data from floods or other catastrophes, we must urgently grapple with the reality that digitisation compromises the foundations of a society’s ability to function in ways we don’t fully understand.

4. **Research and develop applied and reusable navigation interfaces that support cataloguing, exploration and research**: Machine-augmented thinking remains the most compelling use case for personal computers. Between the 1970s and 1980s, a number of breakthrough interfaces were demoed publicly and considered to be the future of digital networked computing. Projects like Ted Nelson’s Xanadu, Hypercard’s 200 Points of Light, the Leap interface, and even the early vision for the web positioned desktop computers as tools that accentuated a person’s ability to understand the world around them. This vision remains completely unfulfilled because dominant design paradigms – now three generations old – have created

stagnant thinking around how to navigate complex topologies of information within a context of heightened personal and network risk.

We believe new approaches to software design are required to develop clarity in complexity and combat information saturation and context collapse. These would be new and novel interfaces and information architecture that allow individuals and small groups to seize the information firehose, and engage with its contents on their own terms in order to curate and preserve large collections of multi-dimensional and cross-format web archives.

5. **Fully understand the threat of decentralisation:** The last fifteen years have seen a surge of interest in decentralised technology. From well-funded projects like IPFS to the emergence of large scale peer-to-peer or federated information networks such as Dat, Secure Scuttlebutt and ActivityPub, there is renewed life in peer-to-peer technologies; a renaissance that enjoys widespread growth, driven by the desire for platform commons and the democratisation of power and agency. As we enter the 2020s, centralised power and decentralised communities are on the verge of outright conflict for the control of the digital public space. The resilience of centralised networks and the political organisation of their owners remains significantly underestimated by protocol activists. At the same time, the decentralised networks and the communities they serve have never been more vulnerable. The peer-to-peer community is dangerously unprepared for a crisis-fuelled future that has very suddenly arrived at their door. ✳

VI. Acknowledgements & appendices

This report was produced by Cade Diehm and Benjamin Royer from the New Design Congress, and was commissioned by Webrecorder through a grant from the Filecoin Foundation. We received feedback from Elys Jones, Colter Wehmeier, Henry Wilkinson and two others who wished to remain anonymous.

The authors wish to acknowledge Ilya Kreymer, Ed Summers, Lorena Ramírez-López and the Webrecorder team, Dietrich Ayala from Protocol Labs and Danny O'Brien from the Filecoin Foundation, and the research participants.

The New Design Congress looks to the future by confronting the gap between society's understanding of what appears to be happening and what is actually happening in today's digital systems. We work with universities, internet subcultures, at-risk communities, companies, non-profits, environmentalists, policy makers and technologists to produce ambitious alternative forks: new paradigms for digital identity, digital information and integrity, digital economic self-sovereignty and digital climate intervention. In looking to the past and identifying – as well as including – voices of those who have thus far been excluded, our work will produce robust, real-world methods and tools for intervening in societal hotspots before they become raging wildfires. Our work speaks for itself: investment in New Design Congress is an investment in daring solutions built on solid foundations that embrace complexity and resilience to inform positive change. New Design Congress is a fiscally-sponsored project of Superbloom.

Webrecorder builds tools specialising in a 'user-driven' form of web archiving, where the user is able to direct the archiving process through their browser. From the beginning, the goal of Webrecorder has been to build quality open-source tools that enable 'web archiving for all,' to allow anyone with a browser to create their own web archives, and to accurately replay them at a later time. The goal of Webrecorder tools is to provide highly accurate capture and replay of websites, working with a variety of existing storage options and services. In addition to advancing web archive capture and replay, the Webrecorder project is focused on advancing open-source software development and research.

Filecoin Foundation is the steward of the Filecoin community. We aspire to put the power of humanity’s most important information back into the hands of everyone. We exist to help people build their vision on Filecoin and to support the growth of the decentralised web.

Superbloom is a non-profit leveraging design as a transformative practice to shift power in the tech ecosystem. We apply a holistic approach and view design as an intervention opportunity to centre people and their needs. Our vision is a world where everyone has the knowledge, network, and digital tools needed to enrich their lives. Founded in 2014 as Simply Secure, our organisation has operated at the intersection of digital design and human rights, steadily evolving alongside the needs of our growing community. We’ve expanded our original focus on usable design practice for secure technologies to include programmatic interventions and research on socio-technical issues like surveillance capitalism, disinformation, and digital consent. Today we work to overcome the alarming lack of available resources in the public-interest technology space by openly sharing our knowledge and partnering with organisations to identify unmet needs, design for access, and extend the impact of newly built tools and interventions. ✳

APPENDIX A:

New Design Congress x Webrecorder Archive Research Project
Participant Consent Form // February - April 2022

You are invited to take part in an interview with personnel from The New Design Congress and Webrecorder.

Please read this form carefully, or have someone read it to you. Ask any questions you may have before agreeing to take part in this interview.

Questions may be directed to the personnel who sent this consent form to you, or by contacting **consent@newdesigncongress.org**. If the consent form is read to the interviewee, either in English or another language, please also have a witness who was present for the reading sign below. Should you have questions after agreeing to take part in this interview, please contact the personnel who sent this consent form to you or email **consent@newdesigncongress.org**.

What is this about: The purpose of this research is to learn more about how and why people use archive tools, and to understand the challenges they encounter during their archival work.

Notes and recordings: With your permission, we would like to take handwritten notes alongside an audio recording of the interview. These will be used for review and analysis purposes. We will not share raw notes or recordings made with anyone outside of New Design Congress or Webrecorder personnel. Any excerpted information or quotations that are used in presentations or publications will be made anonymous. These notes will be stored within New Design Congress' infrastructure.

Please note that New Design Congress is a fiscally sponsored project of Simply Secure, and as such has additional responsibilities to its supporting organization. More information regarding the storage of research data, along with details about the relationship and obligations of New Design Congress to Simply Secure is available at: https://newdesigncongress.org/en/privacy.

More information about ~~Simply Secure~~ [now Superbloom] is available at https://simplysecure.org.

More information about Webrecorder is available at https://webrecorder.net.

Risks and benefits: There will be no invasions of privacy as a result of this research. Any transcriptions that are made of an audio or video recording will have all identifying information removed. Calendar invites and other data the research will be destroyed within 30 days of the completion of the interview. We will take all necessary and appropriate precautions to limit any risk of your participation.

Taking part is voluntary: Taking part in an interview with us is completely voluntary. You do not have to answer any questions that you do not feel comfortable answering. You may instruct the interviewer to stop the interview at any time, in which case no subsequent actions performed by you will be included in our project or publications. You may also withdraw your consent and instruct us to destroy all record of your participation at any time.

Confidentiality: Anything that we make public about our research will not include any information that will make it possible to identify you. Your name, address, and other personal information will not appear in any transcriptions of this interview, and they will not be released to anyone without your written permission. Research records will be kept in a secure location, and only we will have access to them.

Please read and sign the Statement of Consent on the following page.

Webrecorder x New Design Congress Archive Research Project
Participant Bill of Rights

1. I can ask questions about the interview, the organization, or the interviewer at any time.

2. I do not have to answer any question that I do not want to.

3. I can refuse to be video or audio recorded and still participate in this interview.

4. I can leave at any time and withdraw my consent before, during or after the interview.

5. I can provide confidential feedback on my interview experience to the interviewer’s manager.

6. I must approve the use of any photos, audio, videos or anonymized quotes that are used publicly, whether on a website, on a blog, or in the press.

7. Once a photo, video or quote has been published, I have the right to request it be taken down at any point in the future.

Statement of Consent

- I have read this form or it has been read to me.

- I have had the opportunity to ask questions and any questions that I have asked have been answered to my satisfaction.

- I know my rights as a participant.

- I consent voluntarily to participate in this interview and/or usability test and to have any information I provide or audio or video recordings that are made be used in the manner described above.

Name: Date:

______________________________ ______________________

Signature:

____

APPENDIX B: Interview Structure

Each interview is conducted via a platform selected by the participant, and facilitated by two researchers. Interviews are recorded by both facilitators using OBS Studio to record and save locally. Participants are asked to consent to the interview in advance via the Research Consent Form.

Interviews are unstructured, and should follow the top level numbering where possible. Given the time frame for each interview, it is likely that not all questions documented here will be covered.

Interview guide

1. Introductions & technical check
    a. Facilitator introductions
    b. Explain the purpose of the interview
        i. Recap the research goals
        ii. Cover the themes (practice of archiving, institutions, custodianship, etc.)
    c. Cover privacy and consent
        i. Review the Research Consent Form between participant and facilitators
        ii. Confirm with participant that the interview can be terminated at any point
    d. Ask if the participant has any questions
    e. Ask for secondary verbal consent from participant

2. Setting the stage – Participant introduction
    a. *Inform the participant that the recording has started*
    b. Can you tell us a little about yourself?
    c. Can you tell us about your professional experiences of web archiving?
        i. How long? / if not archiving, when do you plan to start?

ii. Why do you / do you want to archive?

iii. How central to your work?

iv. How central to your institution?

3. Definitions of archiving

a. What is your personal definition of archiving?

b. What archiving tools do you use / do you plan to use?

c. Does your personal definition of archiving match the capabilities of the tools you use?

4. Contextualizing practice

a. Please describe an archive project that you have worked on / observed

i. What were its goals?

ii. What tools were used?

iii. Who was the archive for?

iv. What were its successes?

v. What were its failures?

vi. Were there opportunities that existed that were talked about amongst the collaborators?

5. Examining archiving practice

a. Let’s explore your archival practice.

i. How do the tools you use support your work?

ii. How do the tools you use hinder your work?

iii. How do you currently share archived material?

iv. What motivates you to share archive material?

v. How do you organize and categorize your archive material?

vi. What institutional structures does your archived material flow through? (eg, are there specific roles for different parts of the archiving practice)

vii. How important is the chain of custody / authorship to your archival practice, the policies of your institution, or the audience for your archives?

6. Uncovering anxieties

a. *Offer a trigger warning, reiterate participant control*

b. Have you ever felt unsafe / experienced a threat during archiving?

c. Have you ever felt unsafe / experienced a threat while holding archive material?

d. Have you ever felt concerned for the resilience, safety or integrity of the archive material itself?

7. Imagining scenarios

a. Thinking about your most recent work:

i. How could that archived material be weaponized to target you?

ii. How could that archived material be weaponized to target your organization?

iii. How could that archive material be weaponized in a disinformation campaign, or to target someone else?

iv. What would be the easiest way to destroy that archive?

v. What would the implications be if the archive were to be accessible for 100 years?

b. There is a significant push to examine colonialism and consent with regards to institutional documentation of societies.

i. What challenges do you think current archiving practices and tools present…

- … for issues of consent?

- … for issues of power imbalances?

- … for issues of digital ethics?

ii. If your organization has policies that address these issues, can you describe them?

- Can you reflect on your opinion on these policies? What works, and what doesn't?

8. <u>Looking outwards and forward</u>

   a. Thinking about what we have covered today, are there any institutions whose practices you admire and would like to adopt?

   b. Are there any broader social issues now or on the horizon (eg pandemic, climate, etc) that you feel are not being considered by archival tool makers?

   c. Philosophically, what do you believe the broader purpose of archiving practice is?

9. <u>Final thoughts</u>

   a. Anything we haven't covered today?

10. <u>Wrap up</u>

    a. Inform the participant that the recording has stopped.

    b. Debrief, describe next steps.

    c. Ask for off-the-record questions.