A Highly transformative age for web archives

Nicola Bingham, Valérie Schafer, Jane Winters, Anat Ben-David

Abstract: This chapter explores the evolving landscape of web archiving. It considers how web archives document challenging times, may help to analyse them, and respond to events, disruptions, social demands, and crises. It examines emergency response practices and research trends. The chapter also addresses current and forthcoming challenges such as adapting to platformization, AI, the closure of APIs, and evolving legal frameworks. It highlights how web archives are dynamic entities intertwined with contemporary socio-technical contexts, continually adapting to navigate the complexities of a highly transformative age.

Keywords: web studies, web archiving, collections, crisis, emergency responses.

Over nearly three decades of web archiving, we have witnessed profound transformations in the landscape: the emergence of different players, practices, and processes; the evolution of aims and goals; the recognition of new challenges; and the blossoming of collaborations between archival institutions and researchers. Web archive studies has begun to take shape as a discrete field of research, related to but distinct from digital humanities (Brügger 2021), data science and platform studies, and is itself beginning to be interrogated critically. Ben-David (2021, 181–82) notes that "The field of web archiving and web archive research is maturing" and consequently, there is now "room for thinking critically about web archives and for rethinking some of their premises".

Web archives have long been viewed as a rich and "important primary source for humanities researchers" among others (Winters 2017, 245), amenable to qualitative and quantitative research alike. There is scope both for a detailed study of the French presence in London (Huc-Hepher 2021) and for a survey of the entire Danish web domain (Nielsen 2021). The organization of many web archives is both shaped by and shapes these micro- and macro-level approaches, with carefully curated special collections complementing the vast, heterogeneous domain crawls undertaken by many national libraries. The lens through which we view web archives is, however, changing. They remain invaluable repositories of information but are increasingly being considered as important cultural heritage artifacts in their own right. The value of even online memes as cultural heritage is evident from a project like the Meme Wall, which was

Nicola Bingham, British Library, United Kingdom, Nicola.Bingham@bl.uk, 0000-0002-5510-9869
Valérie Schafer, University of Luxembourg, Luxembourg, valerie.schafer@uni.lu, 0000-0002-8204-1265
Jane Winters, University College London, United Kingdom, jane.winters@sas.ac.uk, 0000-0001-5502-5887
Anat Ben-David, University of Israel, Israel, anatbd@gmail.com, 0000-0003-4510-5634

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Nicola Bingham, Valérie Schafer, Jane Winters, Anat Ben-David, *Conclusion: A Highly transformative age for web archives* , © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.29, in Sophie Gebeil, Jean-Christophe Peyssard (edited by), *Exploring the Archived Web during a Highly Transformative Age. Proceedings of the 5th international RESAW conference, Marseille, June 2024*, pp. 343-362, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

created by the Saving Ukrainian Cultural Heritage Online (SUCHO)[1] initiative. The securing of information in the face of conflict was important, but so too was the preservation of digital culture.

Web archiving has also had to respond to and try to keep pace with increasingly rapid change in the broader digital landscape. New web and social media platforms have required the development and combination of different tools and approaches. Social media, in particular, is something of a "moving target", requiring agility and innovation from those who would archive and study it. More sophisticated approaches to documenting and sharing data have also been influential for research and practice in web archive studies, although they are not always easily accommodated. The growing emphasis on data that is Findable, Accessible, Interoperable and Reusable (FAIR) and the influence of open science initiatives have changed the field significantly, but the relative lack of change in the interlocking legal frameworks that govern the harvesting and archiving of the web creates friction.

In this final chapter, and as we are coming to the end of this collective reflection, we would argue that web archives and web archive studies are reaching an inflection point. After a "long process resulting in the consolidation of standards, best practices, shared methods, tools, and knowledge" (Ben David 2021, 182), there is an opportunity to reevaluate web archiving research and practice and to reconsider the relationship between web archives and their contemporary socio-technical contexts. Web archives are not merely static repositories; they are dynamic entities closely entangled with contemporary challenges. These include ethical approaches to the archiving, preservation, and reuse of personal and public data; the environmental impact of digital preservation in the face of a climate crisis; and the requirement to respond swiftly to unforeseen events and crises.

All of these transformations unfold simultaneously within web archiving institutions and in the broader context of changing digital and scientific practice. They affect the objects and subjects of study; methods of data collection and preservation; and the demands and expectations placed on web archives by society. In this chapter, we explore the extent to which web archives are both active participants in and influenced by this highly transformative age, and how they are responding to it. It begins by considering web archives as 'archives of crisis', which play a vital role in recording the traces of natural and man-made crises as they play out in online spaces. It then discusses how web archives are (and are not) in tune with these challenging and febrile times, while also exploring the processes of constant renewal, adaptation, and ultimately transformation that have

[1] https://www.sucho.org. For the Meme Wall, see https://memes.sucho.org

allowed web archives to weather the digital and social storms of the early 21st century. Finally, it identifies current and future challenges that will necessitate continuing adaptation and innovation in web archiving and web archive studies.

1. Web archives responding to events, disruptions, social demands and crises

Web archives serve as snapshots of our online world and attempt to mirror the challenges and crises that unfold. However, they are always filtered through the lens of human decisions, technological limitations, and resource constraints, embedding an inherent subjectivity. In acknowledging disruptions and crises, web archives exhibit a long-standing vigilance and responsiveness, reacting by creating special collections that highlight these pivotal moments.

1.1 Societal expectations and institutional roles

Societal expectations place a significant burden on heritage institutions in addressing crises and collecting information about them. Libraries and archives are seen as custodians of history, responsible for preserving the collective memory of society. Therefore, their role in archiving digital content during crises is crucial in ensuring that these moments are not lost or distorted with time. The public expects these institutions to capture a diverse range of perspectives, ensuring that the archive is representative and reflective of the entirety of an event. This situation was, for instance, heightened during the COVID-19 crisis, when institutions preserved a born-digital record of the pandemic (see for instance "Mapping the archival horizon: A Comprehensive survey of COVID-19 web collections in European GLAM institutions" by Nicola Bingham).

Web archives have always shown sensitivity to disruption and crisis, and web archiving institutions have demonstrated an awareness, and responsiveness to traumatic events that predates the archived web. Recently, however, there has been a noticeable professionalization of emergency response practices within the web archiving community. While the British Library, through the UK Web Archive (UKWA), had already reflected the 2005 terrorist attacks in London through a small, curated collection, the web archiving responses to the Paris terrorist attacks in 2015 were quite different. Ten years after events in London, the technical means, the digital landscape, and the skills and resources available to web archivists had changed, leading the French institutions that preserve the web—BnF (Bibliothèque nationale de France) and Ina (Institut National de l'Audiovisuel)—to collect a vast amount of websites and social media content (i.e., 20 million tweets preserved by Ina on the 13 November terrorist attacks, see Schafer et al. 2019).

This shift towards "living archives" (Rollason-Cass and Reed 2015) is characterized by increased experience, better guidelines, international collaboration, and heightened reactivity. The COVID-19 crisis expedited this professionalization, as evidenced in the next section, emphasizing the need for swift and comprehensive archiving of digital content during unforeseen events. Cultural heritage institutions have also fostered widening participation in relation to web archiving, either through special collections curated by researchers or research teams (as seen in UKWA or the BnF, for example), through calls for participation to the general public, or by involving colleagues during the COVID crisis. For instance, the BnF enriched its COVID collection with more local perspectives thanks to its local correspondents or integrated BnF staff, some of whom were isolated during the lockdown and unable to perform their usual tasks (Gebeil et al. 2020).

In the realm of professionals, international participation and increasingly refined processes within the global context, particularly thanks to the IIPC (International Internet Preservation Consortium), also demonstrate international solidarity coupled with a community of practices.

1.2 Professionalization of emergency response practices

The period encompassing the pandemic from 2020 to 2022 expedited the professionalization trajectory of emergency response practices within the sphere of web archiving. This swift evolution was evidenced by enhanced expertise, heightened responsiveness, the formulation of guiding frameworks, and concerted collaboration among cultural heritage institutions.

The importance placed on preserving online materials related to the pandemic may be illustrated by the increase in data allocation for the IIPC COVID-19 collection[2] from the Internet Archive/Archive-It. An extra two terabytes of data were allocated, free of charge, bringing the total data budget to five terabytes specifically for archiving COVID-19-related web content. This gesture signifies a proactive response to the exceptional volume and critical nature of online information surrounding the global health crisis and underscores a collective effort to ensure comprehensive and robust preservation of international web content for future historical, research, and public informational purposes (Geeraert and Bingham 2020).

Libraries and archives demonstrated multifaceted approaches to fostering collaboration in COVID-19 web archiving activities. Notably, cultural institutions actively forged collaborative alliances, recognizing the inherent complexity of documenting an issue of such magnitude and acknowledging

[2] See the collection at https://archive-it.org/collections/13529

the necessity for collective effort in this endeavor. The Royal Danish Library, for example, was part of a general project documenting coronavirus lockdowns in Denmark in 2020. This effort was a cooperation between several cultural institutions, including the National Archives (Rigsarkivet), the National Museum (Nationalmuseet), the Workers Museum (Arbejdermuseet), and local archives, which resulted in a more comprehensive collection than could be achieved by any one institution on its own (Schostag 2020).

COVID-19 archiving projects accelerated a trend for collecting a diverse range of voices in web archives:

> It seems as though a threatening present and the urge to collect as many voices as possible on the COVID-19 crisis in web archives encourage us to question history as it has been written and discussed in the modern era (Priem and Grosvenor 2022).

The urgency of capturing numerous perspectives amid the COVID-19 crisis prompted a re-examination of historical narratives. The authors suggest that both the digital age and the pandemic fostered an augmented awareness of establishing new dimensions for engaging with and reflecting upon the past.

Chiara Zuanni (2022) makes the point that the pandemic led to a surge in born-digital material collected by museums and memory organizations, posing challenges in collection, preservation, and display. The volume and hybrid nature of these collections (mixing physical and digital objects) necessitated new approaches and technical capabilities, pushing institutions to develop innovative methods.

As mentioned above, institutions also adopted participatory approaches, aiming for more comprehensive and diverse collections, while at the same time trends towards grassroots and bottom-up initiatives are also becoming apparent in the field of web archiving.

1.3 Participation and grassroots efforts

During the Arab Spring, collections were curated by institutions such as the National Library of Tunisia (BnT), the Internet Archive and the Library of Congress. In the case of the BnT, the institution underwent a transformative shift in its collecting approach to web archiving, as detailed by Raja Ben Slama (2023 and her chapter "Web archiving in Tunisia post-2011: The National Library of Tunisia's experience"). The BnT recognized the need for a collective effort involving public institutions, associations, and volunteer researchers to collect and archive scattered digital documents from the web and citizens' mobile phones. This collaboration aimed to capture firsthand accounts and materials from those who participated in the uprisings. It highlighted the necessity of adapting to the changing landscape by creating a web archiving unit outside the traditional legal deposit service

to address the lack of provision for archiving documents from social networks and various creative forms that emerged post-revolution. These challenges were outside the framework of existing legal regulations and required innovative solutions. This suggests an ongoing process of adaptation and growth in response to the dynamic nature of digital content and the challenges of preserving and organizing it effectively, exacerbated by the temporal urgency of responding to crisis events.

Preserving online content during crisis events like terrorist attacks and movements for social justice has spurred advancements in both web archiving technologies and policies guiding collection development. The Documenting the Now project[3] played a significant role in archiving digital content related to the shooting of Michael Brown in Ferguson, Missouri, in 2014. This initiative aimed to capture and preserve the digital traces of social media conversations, images, videos, and online content that emerged in the aftermath of this pivotal event. The project recognized the importance of archiving these digital conversations in real-time. It developed tools and methodologies to capture and preserve Twitter feeds, Facebook posts, Instagram images, and other social media content.

The project also addressed ethical considerations regarding the archiving of sensitive and potentially traumatic material. It engaged with issues of consent, privacy, and the responsible preservation of digital records in a sensitive societal context. In its second phase, Documenting the Now underlined the need for local participation and "digital community-based archives from the perspectives of local activists and in equitable partnership with them" [4].

This trend of non-institutional actors playing a pivotal role in archiving critical events appears to be strengthening. Taking on the challenge of community engagement in web archiving, SUCHO exemplified success by mobilizing volunteers to safeguard Ukraine's online cultural heritage. As the war unfolded, the response was swift, and the project launched on March 1, 2022, successfully bringing together over 1,500 volunteers from more than 38 countries. Coordinating such an initiative required establishing procedures to guide volunteer efforts effectively. According to the 2022 report[5], a metadata team created, for example, "metadata guidelines and video tutorials for direct upload of individual items with metadata to Internet Archive [...]", while also launching a test aimed at volunteers to better adapt these guidelines. It differs somewhat from the guidelines and target audience of the Archive Team, which also issued calls for volunteers and saved numerous online platforms in jeopardy, such as

[3] http://www.docnow.io/
[4] https://archive.mith.umd.edu/mith-2020/documenting-the-now-phase-2/
[5] https://www.sucho.org/assets/Mar-Dec-2022_End_of_Year_Updates.pdf

Geocities or Mobileme, but with a more technical approach[6]. Non-institutional, grassroots efforts in web archiving play a crucial role in capturing diverse perspectives, filling gaps in institutional collections, and ensuring the preservation of digital heritage in a more inclusive and comprehensive manner. As noted by Cui Cui et al. (2023), such efforts often target content that might not be on the radar of larger institutions. They focus on niche topics, local events, or marginalized communities that might not receive adequate attention from traditional archives. This engagement fosters a sense of ownership and empowerment within the community, allowing them to contribute to preserving their history and narratives.

However, challenges persist and not all crises and conflicts receive equal attention or coverage within web archives due to limitations in resources and differing priorities among institutions. Bridging these gaps is crucial to ensure more comprehensive web studies and inclusive representation of our digital history.

2. Web studies in tune with the challenging times

While our initial focus has centered on the dedicated efforts and practices of web archiving, this digital heritage also serves as a valuable resource for researchers. A comprehensive approach to crises and to our highly transformative age necessitates an intrinsic connection to related research endeavors. This interest is exemplified in various chapters of this book, showcasing a sensitivity to controversies, crises, and related memories. The chapters delve into diverse topics, such as "The Words of online hospitality" by Dana Diminescu and Quentin Lobbé; "Mapping the archival horizon: A Comprehensive survey of COVID-19 web collections in European GLAM institutions" by Nicola Bingham; and "Archiving public health discourse in the UK Web Archive" by Alice Austin. Researchers have embraced the challenge of using web archives to enhance the understanding of the first decades of the 21st century.

2.1 Web studies to analyze challenging times

Research programs have emerged as proactive initiatives addressing critical issues. Noteworthy examples include ASAP[7] (Archives Sauvegarde Attentats Paris, funded by the French CNRS), launched in the aftermath of terrorist attacks in 2016, WARCnet's dedicated focus on COVID studies within working group 2[8], and the recent project by Anat Ben-David on the

[6] https://wiki.archiveteam.org/index.php/ArchiveTeam_Warrior

[7] https://asap.hypotheses.org/author/web90

[8] https://cc.au.dk/en/warcnet

history of climate news images using web archives. These research programs, among others, reflect a sensitivity to crises and social transformations. Web archiving collections support these current trends, whether they are constituted by researchers or by web archivists, as demonstrated by those accessible in Archive-It that, for example, capture social movements. From #Metoo and Black Lives Matter to the social upheavals following Fukushima in 2011 and the 'Arab Spring', these collections rely on a collaborative and crowdsourcing model that was established in 2007 after the Virginia Tech campus shooting.

Recognizing the complex nature of crises, there is a growing need for interdisciplinary approaches. While technical and cognitive skills are crucial, recent global challenges demand a broader perspective. Geopolitical, health-related, and social issues associated with recent crises transcend digital expertise, necessitating collaboration among diverse professionals. Initiatives like WARCnet exemplify this interdisciplinary as well as interprofessional model, gathering together web archivists and researchers. It also calls more and more for international research teams. Despite the efforts of a reactive and committed community, challenges persist. The limitations of restricted time and funding pose hurdles to comprehensive archiving and research. Balancing the urgency of capturing the present with sustainable, long-term strategies remains a critical consideration.

2.2 Memories as a key research theme

Web archives are also at the core of some memory studies (see, for example, Clavert 2018 and Gebeil 2016; 2021). This field is increasingly growing in web studies, as evidenced by various chapters in this book, such as "Websites as historical sources? The benefits and limitations of using the websites of former repatriates for the history of schooling in colonial Algeria?' by Christine Mussard, and "A Social media archive for digital memory research" by Costis Dallas and Ingrida Kelpšienė. Studies of socio-digital networks also contribute significantly to this field. Moreover, they intentionally act as memory producers, as noted by Jacobsen and Beer (2022). In the study of social networks, it is worth highlighting the significance of distant reading and the technical and scientific challenges involved, as identified early on by Frédéric Clavert and actively addressed in the CONNECTIVE social media digital archive project presented by Dallas and Kelpšienė in this book. This project accounts for user interactions through posts, comments, and reactions on social media, representing these interactions in a knowledge graph of deposits, agents, threads, and messages representing these social media interactions.

Mentioning knowledge graphs and distant reading, advancements in tools and processes, including but not limited to distant reading and metadata analysis, have to be underlined.

A WARCnet report on *Skills, tools, and knowledge ecologies in web archive research* (Healy et al. 2022) highlighted the sheer range of skills and professional knowledge required to work with web archives, from software and tools, through digital curation processes and workflows to data analysis and web design methods. Researchers and practitioners are required both to acquire a broad base of technical and archival skills and to develop their knowledge and expertise in order to keep pace with the rapidly changing web, digital preservation, and cultural heritage landscapes. Web archiving in national libraries and other memory institutions is generally undertaken by small teams who are facing increasing volumes of work against a background of systemic underinvestment in the cultural heritage sector. In this context, it is difficult to find either the time or the resourcing for effective and sustained programs of training and professional development. With a few notable exceptions, the digital skills training offered to humanities researchers does not encompass born-digital archives. Consequently, they are frequently left to develop their skills independently, learning through trial and error.

Like many of the other challenges identified in this chapter, collaboration and partnership offer an effective means of addressing the skills deficit (see "A network to develop the use of web archives: Three outcomes of the ResPaDon project" by Emmanuelle Bermès et al.). Both WARCnet and RESAW have shown sustained commitment to skills training for web archiving and web archive studies. The IIPC is similarly committed to upskilling web archivists and sharing knowledge between established and new entrants in web archiving. The development of stronger connections with professional bodies in the fields of Digital Humanities, Research Software Engineering and Cultural Heritage, for example the Society of Research Software Engineering, would usefully strengthen these existing networks.

Web researchers and archivists face significant hurdles when creating live web corpora, notably due to the complexity of web scraping tools. Michael Black (2016) emphasizes the lack of a universal solution due to diverse content hosting platforms and the rapid evolution of web language standards. This diversity necessitates tailored tools, posing a challenge for archivists seeking options aligned with their needs amidst commercially oriented or larger research-based offerings. Despite the availability of resources like Voyant and robust programming libraries for text mining novices, effectively using these tools requires additional skills and training

for web researchers and archivists (see for instance the studies "Web archives and hyperlink analyses: The case of videnskab.dk 2009–2022" by Niels Brügger and Katharina Sølling Dahlman, and "Multi-level structure of the First Tuesday communities after the 2000 dot-com crash: A social network analysis of economic actors based on web archives" by Quentin Lobbé).

The development of appropriate tools and infrastructure are essential if researchers and practitioners are to have access to environments in which they can hone and refresh their skills. Initiatives like ARCH, a platform developed by the Archives Unleashed team[9] and then adapted by the Internet Archive "for building research collections, analyzing them computationally, and generating datasets from terabytes and even petabytes of data[10]" have enormous potential to open up web archives for researchers who have limited or no access to technical support in their own institutions. Crucially, training programs, tools, and infrastructure will be required to keep ahead of developments on the live web and help answer issues underlined in chapters of this book, like "Making social media archives: Limitations and archiving practices in the development of representative social media collections" by Beatrice Cannelli and "Challenges in archiving the personalized web" by Erwan Le Merrer, Camilla Penzo, Gilles Tredan, and Lucas Verney. Of course, these challenges go beyond technical issues to encompass ethical and legal considerations.

2.4 Ethical considerations in challenging times

Researchers and web archivists navigate both legal and ethical dimensions in their work. The impact of the GDPR in Europe looms large, particularly concerning the proliferation of personal data in web archives. Research guidelines and policies, exemplified by the shift towards FAIR data, add further layers of complexity. Compliance is not always straightforward, given the limited shareability of web archives content, often constrained by legal deposit frameworks in various countries. Issues related to author rights further complicate matters. However, there is a notable shift towards the shareability of seed lists and metadata, as seen in a project like AWAC2 (Aasman et al. 2021), based on the COVID IIPC collections in Archive-It, which facilitated a comprehensive study of COVID data through distant reading. The use of permalinks also supports more widespread citation. Yet, many challenges persist, especially with vast collections of web archives whose accessibility is limited to library reading rooms. Internet Archive, Arquivo.pt, and projects like SUCHO provide

[9] https://archivesunleashed.org/arch/
[10] https://webservices.archive.org/pages/arch

online access, but the question of reuse remains quite unresolved. Ethical dimensions, explored in the context of Geocities by Ian Milligan, are increasingly pertinent too. He emphasized the ethical dilemma of studying personal pages, asserting that:

> Leaving people out isn't ethical either. Moving to a full opt-in process would likely lead to the historical record being dominated by corporations, celebrities and other powerful people, tech males, and those [who] wanted their public face and history to be seen a particular way (Milligan 2018).

Meghan Dougherty's analysis already highlighted the evolving ethical landscape, shifting from concerns about copyright permissions to a focus on privacy:

> The debate is not simply a matter of whether or not it is ethical to preserve what some users consider to be ephemeral artifacts in permanent and accessible storage. The debate is far more complicated, involving various information behaviors, conflicting expectations, and different interpretations of how our information online represents our most intimate selves (Dougherty 2013).

Ethical debates extend beyond privacy, encompassing issues of inclusiveness, power relations, and potential invisibilization. In a white paper, Documenting the Now highlights tensions such as user awareness, the potential fraudulent use of social media content, and the

> heightened potential for harm to members of marginalized communities using the web and social media, especially when those individuals participate in activities such as protests and other forms of civil disobedience that are traditionally heavily monitored by law enforcement (Jules et al. 2018, 9).

The National Forum on Ethics and Archiving the Web, organized by Rhizome in 2018, further underscored the multifaceted nature of ethical challenges, with panel discussions, like "Archiving Trauma" and "Documenting Hate". Pamela M. Graham (2017) emphasizes the intertwining topics of ethics and diversity. Transparency in the archiving process is considered crucial for creating collections ethically. This article also challenges the black box of web search engine algorithms, referring to Safiya Noble's *Algorithms of oppression*, and pointing out that in web archives, biases should be mitigated rather than perpetuated: by developing "effective search functions, we have the opportunity to offer a very different use experience than what the live web affords".

3. Trends and future challenges

While some challenges have already been underlined, for example the need for constant renewal and adaptation of skills, the requirement to face the rapid changes implemented by platforms and social media networks, the

issue of inclusiveness and public participation, and the need for co-shaping and co-sharing web archives, there are still others that promise to be highly transformative for web archives and practices. Artificial intelligence, misinformation, platformization, and asymmetries are key topics to be addressed to better adapt or consider web archiving in the near future and may become strong drivers of transformation.

3.1 Anticipating the future with artificial intelligence (AI) and machine learning (ML)

AI and ML are reshaping the landscape of web archiving, expanding the types of data captured, diversifying the methods of analysis, and highlighting the challenges of ethical considerations, compliance, and the preservation of diverse perspectives. AI has undoubtedly had a positive impact on web archiving in several areas, for example by enabling the capture and organization of events that might typically fall outside the purview of institutional archives' defined collection policies, or within conventional archival collection policies. This expanded scope enhances the inclusivity and depth of archived content (Sönmez et al. 2016).

In their chapter, Emmanuelle Bermès et al. discuss the ResPaDon project which aims to provide the research community and archivists with methods and tools for the building, analysis, and dissemination of web corpora. To this end, Sciences Po médialab and the BnF organized, ran, and evaluated an experiment based on the use of the Hyphe web crawler on web archives.

The expansion of web archivist and researcher-friendly tools integrating AI and ML are also to be noted, while there arise significant ethical and legal considerations that archivists and web researchers must confront. Black (2016) points out, for example, that while web scraping might not inherently violate intellectual property laws, recent US and EU court cases have scrutinized whether scraping affects data value or causes economic harm to data hosts.

Guidelines are beginning to be developed, for example the OCLC publication, "Guidelines for Libraries' Responsible Use of AI: Responsible Operations", a guide developed in collaboration with professionals from various sectors, presents a framework to address technical, organizational, and social challenges related to the operationalization of data science, ML, and AI in libraries. This agenda highlights seven areas of investigation, providing recommendations to guide discussions and actions toward responsible engagement with these technologies (Padilla 2019).

Lynch (2017) raises concerns about stewardship for algorithms, highlighting that current archival techniques and training might not be sufficient to preserve the perspective of the "age of algorithms" for future understanding. There is a need to adapt archivists' education to capture the impact and implications of algorithms on data capture and processing.

Jaillant and Caputo (2022) highlight significant challenges faced by web archives in utilizing AI. They emphasize the risks associated with algorithmic errors, citing an instance where open-source software flagged innocuous terms incorrectly, leading to false positives. They also address the ethical and social implications of AI-driven decision-making due to the opacity of AI processes and the potential biases within training data. To mitigate these challenges, they advocate for "Explainable AI", which facilitates human comprehension of machine-generated outcomes and stresses the need for interdisciplinary collaborations among archivists, Digital Humanists, and Computer Scientists to navigate these ethical complexities. Recognizing the interdisciplinary nature of these challenges, various networks and initiatives have emerged, such as AURA and AEOLIAN, fostering exchanges among scholars, archivists, librarians, and museum professionals. While AI has gained prominence in various sectors, its application in libraries and archival institutions remains at an experimental stage, necessitating more robust and compelling case studies to drive advancements in this domain.

3.2 Misinformation in a post-truth era

A second challenge, which is already very much with us—"post-truth" was the Oxford Dictionaries international word of the year in 2016—is that of dealing with mis- and disinformation as it enters web and social media archives. Archives have always included misinformation, for example medieval charters that are only identified as forgeries centuries after their creation, but the scale of the challenge when dealing with the archived web is unprecedented. The Archive of Tomorrow project, which ran for 14 months from February 2022, was set up to identify and preserve both online information and misinformation related to public health in the UK, and in particular to the COVID-19 pandemic of 2020 onwards. The project's final report notes that "The need to identify and archive both accurate information as well as the inaccurate is now a pressing societal need" (Archive of Tomorrow 2023), but this is not an easy thing to do. The project identified several practical challenges, from the "question of how to name and describe a collection which was explicitly open to capturing misinformation as well as information" to "Who is responsible for identifying and labeling misinformation in research collections?". The main challenge, however, remains one of how to present and contextualize misinformation such that it can be distinguished from other archived data. This is feasible, although labor intensive, for smaller special collections, but becomes difficult, if not impossible, at the scale of a national web domain. How can what Acker and Chaiet (2020) describe as "The weaponization of web archives", contributing to an online "misinfodemic", be combated?

Metadata and documentation are important tools for web archives, allowing users to investigate provenance and make informed decisions about the accuracy, currency and even reliability of the information they are looking at. Dense metadata and descriptive text run the risk of being ignored or misunderstood, especially when the archived web is itself such a complex source, so the accessible presentation of key contextual information will be crucial in helping researchers and readers to navigate around (mis)information effectively. Expert human curation will remain important, but the use of artificial intelligence to flag problematic content and assist with the creation of metadata is likely to make the task of contextualizing material in large born-digital archives easier. This will not, however, address every kind of misuse, like the "screensampling" identified by Acker and Chaiet (2020), which involves posting screenshots of archived URLs to remove the ability to click on or track these static images of archived online sources.

3.3 Closure of APIs and platformization

Within the framework of IIPC WAC24[11], Frédéric Clavert invited scholars to discuss "Archiving social media in an age of APIcalypse". This discussion arose after two major platforms, Twitter (now X) and Reddit, placed access to their APIs behind a paywall in the early part of 2023. As noted by Clavert, Application Programming Interfaces (APIs) play a crucial role in accessing data and harvesting for archived collections and various research projects. This closure echoes past incidents, such as LinkedIn restricting data access in 2015 and Facebook reducing API functionalities post the Cambridge Analytica scandal, and it is referred to as an "APIcalypse" by Axel Bruns (2019). The closure of APIs has multifaceted repercussions. Notably, it adversely affects collections of the kind that have previously been allowed, for example Clavert's analysis of World War I memories (2018) and Nick Ruest's immediate collection during the Charlie Hebdo attacks[12]. Institutions also leverage APIs, as demonstrated by the INA's creation of extensive collections, some deeply tied to highly transformative events like the Charlie Hebdo attacks and social movements like the protests of the Gilets Jaunes.

Terms of use or access can regularly evolve, necessitating a reconfiguration of collection methods, while consideration must also be given to frequent updates and sometimes unsatisfactory crawls. As noted by Ben Els (National Library of Luxembourg) Facebook actively blocks crawler robots, necessitating the collection of more data for reliable results,

[11] https://netpreserve.org/ga2024/
[12] https://ruebot.net/tags/charliehebdo/

while the cost of capturing Facebook may exceed that of archiving a regular website, with issues such as non-functional videos (Els and Schafer 2020). "The Telegram Archive of the War in Ukraine" serves as a testament to these challenges[13], emphasizing the need for human curation and the constant issues at stake (Holownia and Socha 2022).

The ongoing platformization and the challenges and limits posed by giant web companies to web archiving, are concerning. Loss of functionality and a strong dependence on proprietary platforms are crucial considerations, especially as their influence and usage continue to strengthen, while also touching upon asymmetries and gaps in web archiving.

3.4 Asymmetries and gaps

Although web archives have been characterized by ingenuity, innovation, and responsiveness to both technological and societal change, there remain significant asymmetries in the collectivity of the archived web—an overrepresentation of some voices and experiences and an underrepresentation of others.

Web archiving continues to be an activity primarily in and of the Global North, and this necessarily distorts the digital historical record. The work of the IIPC in organizing global collections, for example the Novel Coronavirus (COVID-19) special collection[14], goes some way to addressing this imbalance. At the time of writing, the country most represented in the COVID-19 collection is the US (1,988 websites), but it is followed by Brazil, Argentina, Peru, Uruguay, Chile, and Bolivia. There are, however, only 85 Chinese websites included in the collection, the first African country to be mentioned is Angola (59 websites) and there are only 15 archived resources for India. The Internet Archive's Whole Earth Web Archive (WEWA), launched in October 2019, was designed as "a proof-of-concept to explore ways to improve access to the archived websites of underrepresented nations around the world". The Internet Archive is committed to "undertaking active outreach to national and heritage institutions in these nations, and to related international organizations, to ensure this work is guided by broader community input" (Bailey 2019), but the WEWA remains a Global North initiative on behalf of nations without their own web archiving infrastructure.

Lor and Britz (2004) have highlighted the complex morality of what they describe as "South-North web archiving", and the need for clarity both about the motivations of archiving institutions and about the balance between the right of access to information and the right to own and control

[13] https://storymaps.arcgis.com/stories/0af72de4b008461bb441fc62fffb9f8d
[14] https://archive-it.org/collections/13529

it. There is much that could be learned from the CARE Principles for Indigenous Data Governance—Collective Benefit, Authority to control, Responsibility, and Ethics—which are not yet part of the mainstream of web archiving research and practice. In the field of Digital Humanities, Quintanilla and Horcasitas (2013) have called for "transnational solidarity", based on "relationships and networks of care that exceed the logic of national boundaries" and which can "lay the groundwork for decolonial and sustainable futures". These are calls to action that web archiving and web archive studies can take up on their own terms.

A particular challenge is the large-scale, top-down, and primarily automated nature of national domain crawls. They are designed to be as comprehensive as possible, but they can be neither complete nor consultative; nor do they systematically include most social media platforms. Consequently, it is at the level of special collections, where careful curation is possible, that gaps and asymmetries in web archives can more easily be addressed. Schafer and Winters (2021) note that "With regards to inclusiveness, there have been some notable efforts to diversify special collections, so that web archives become visibly more inclusive". Ensuring that this is true for the vast web archives of national domains or major global events will, however, remain difficult.

Finally, we cannot ignore a current and future challenge related to sustainability, environmental impacts, resource allocation, and long-term preservation strategies. The environmental cost of AI is increasingly being discussed and there are calls to embrace "digital frugality", and web archives must be involved in these conversations. The experience of collecting born-digital records during the pandemic acted as a catalyst for recognizing the need for more refined and sustainable digital preservation practices. This necessity stimulated demand for the development of appropriate workflows and strategies tailored specifically for born-digital objects. For instance, Blair et al. (2021) covered how to pitch web archiving to funding bodies, how to make appraisal decisions when gathering URL seeds, how to manage crawling within a limited data budget, and tools and techniques for managing this work between several people working remotely. As highlighted by Pendergrass et al. (2019):

> [...] it is time for all cultural heritage professionals who work with digital content to engage with this urgent issue and to critically evaluate current practices in appraisal, permanence, and availability of digital content to create environmentally sustainable digital preservation.

To conclude, as demonstrated during the whole book and in this final chapter, web archives and the efforts of web archivists are vital in documenting and preserving digital material during times of crisis and

disruption. However, inherent biases, resource constraints, and challenges create imperfections and asymmetries in the representation of these events. While increased collaboration within the field is a positive step forward, addressing issues of participation, openness, and resource allocation are crucial in creating more balanced and inclusive web archive collections. Continuities in debates persist alongside obstacles and paradoxes, notably in the realm of author rights, re-use, and shareability. Despite ongoing efforts and notable progress, these issues linger, challenging access and dissemination.

Amidst the call for some change, there is an equally pressing need for stability. Balancing the rapid pace of digital and social changes with the necessity for established practices and research frameworks requires a delicate equilibrium. Shared frames and moments of respite are essential, while the risk of presentism and constant emergency in the face of the numerous crises that arise is to be avoided. The urgency of capturing and documenting the present must be counterbalanced with a nuanced understanding of heritagization and historical context to avoid distortions and oversights in the archived narrative. The allocation of means and resources also emerges as a critical consideration in sustaining effective web archiving initiatives. Striking a balance between social expectations, the demands for innovation and the realities and practicalities of crawling and resource management remains an ongoing challenge.

References

Aasman, Susan, Niels Brügger, Frédéric Clavert, Karin de Wild, Sophie Gebeil, and Valérie Schafer. 2021. "Analysing Web Archives of the Covid-19 Crisis through the IIPC collaborative collection: early findings and further research questions." *International Internet Preservation Consortium Blog*, November 2, 2021. https://netpreserveblog.wordpress.com/2021/11/02/analysing-web-archives-of-the-covid-19-crisis-through-the-iipc-collaborative-collection-early-findings-and-further-research-questions/

Acker, Amelia, and Mitch Chaiet. 2020. "The weaponization of web archives: data craft and Covid-19 publics." *Harvard Kennedy School Misinformation Review* 1. https://doi.org/10.37016/mr-2020-41

Archive of Tomorrow. 2023. *Archive of Tomorrow: Capturing Public Health Discourse in the UK Web Archive*. Edinburgh: National Library of Scotland.

Bailey, Jefferson. 2019. "The Whole Earth Web Archive." *Internet Archive Blogs*, October 30, 2019. https://blog.archive.org/2019/10/30/the-whole-earth-web-archive/.

Ben-David, Anat. 2021. "Critical web archive research." In *The Past Web: Exploring Web Archives*, edited by Daniel Gomes, Elena Demidova, Jane Winters, and Thomas Risse, 181– 88. Cham: Springer Nature.

Ben Slama, Raja. 2023. "Web archiving in Tunisia after 2011. Experience of the National Library of Tunisia." Abstract for the RESAW 2023 Conference. https://resaw2023.sciencesconf.org/resource/page/id/14

Black, Michael L. 2016. "The World Wide Web as complex data set: Expanding the Digital Humanities into the twentieth century and beyond through Internet research." *International Journal of Humanities and Arts Computing* 10, 1: 95–109. 10.3366/ijhac.2016.0162

Blair, Lindsey, Claire Drone-Silvers, Denise Rayman, and Rhys Weber. 2021. "Building a COVID-19 Web Archive with Grant Funding." *Midwest Archives Conference Annual Meeting Presentations*. Iowa State University Digital Press. https://www.iastatedigitalpress.com/macmeetings/article/id/12582/

Brügger, Niels. 2021. "Digital humanities and web archives: possible new paths for combining datasets." *International Journal of Digital Humanities* 2, 145–68.

Bruns, Axel. 2019. "After the 'APIcalypse': social media platforms and their fight against critical scholarly research." *Information, Communication & Society* 22, 11: 1544–66. 10.1080/1369118X.2019.1637447

Clavert, Frédéric. 2018. "Temporalités du Centenaire de la Grande Guerre sur Twitter." In

Temps et temporalités du web, edited by Valérie Schafer, 113–34. Nanterre: Presses universitaires de Paris Nanterre.

Cui, Cui, Stephen Pienfield, Andrew Cox, Frank Hopfgartner. 2023. "Participatory Web Archiving: The path towards more inclusive web archives?" Abstract for the RESAW Conference 2023. https://resaw2023.sciencesconf.org/433545

Dougherty, Meghan. 2013. "Property or Privacy? Reconfiguring Ethical Concerns Around Web Archival Research Methods." AOIR Selected Paper, Denver, USA. https://spir.aoir.org/ojs/index.php/spir/article/view/8804/pdf

Els, Ben, and Valérie Schafer. 2020. "Exploring special web archive collections related to COVID-19: The case of the BnL." *WARCnet Papers*. Aarhus: WARCnet.

Gebeil, Sophie. 2021. *Website Story. Histoire, mémoires et archives du web*. Bry-sur-Marne: INA.

Gebeil, Sophie. 2014. "Le web, nouvel espace de mobilisation des mémoires marginales. Les mémoires de l'immigration maghrébine sur l'internet français (2000–2013)." *Cahiers Mémoires et Politique* 2. https://doi.org/10.25518/2295-0311.115

Geeraert, Fridel, and Nicola Bingham. 2020. "Exploring special web archives collections related to COVID-19: The case of the IIPC Collaborative collection, An interview with Nicola Bingham (British Library) conducted by Friedel Geeraert (KBR)." *WARCnet Papers*. Aarhus: WARCnet.

https://cc.au.dk/fileadmin/user_upload/WARCnet/Geeraert_et_al_COVID-19_IIPC__1_.pdf

Graham, Pamela. 2017. "Guest Editorial: Reflections on the Ethics of Web Archiving." *Journal of Archival Organization* 14, 3-4: 103–10. 10.1080/15332748.2018.1517589)

Healy, Sharon, Helena Byrne, Katharina Schmid, Nicola Bingham, Olga Holownia, Michael Kurzmeier, and Robert Jansma. 2022. "Skills, tools, and knowledge ecologies in web archive research." *WARCnet special report*. Aarhus: WARCnet. https://cc.au.dk/fileadmin/dac/Projekter/WARCnet/Healy_et_al_Skills_Tools_and_Knowledge_Ecologies.pdf

Holownia, Olga, and Sacha Kelsey. 2022. "Web Archving the War in Ukraine." *International Internet Preservation Consortium Blog*, July 20, 2022. https://netpreserveblog.wordpress.com/2022/07/20/web-archiving-the-war-in-ukraine/

Huc-Hepher, Saskia. 2021. "Queering the web archive: a xenofeminist approach to gender, function, language and culture in the London French special collection." *Humanities and Social Sciences Communications* 8, 1–15.

https://doi.org/10.1057/s41599-021-00967-8

Jacobsen, Ben, and David Beer. 2021. *Social Media and the Automatic Production of Memory: Classification, Ranking, and Sorting of the Past*. Bristol: Bristol University Press.

Jaillant, Lise, Annalina Caputo. 2022. "Unlocking digital archives: cross-disciplinary perspectives on AI and born-digital data." *AI & Society* 37, 823–35. https://doi.org/10.1007/s00146-021-01367-x

Jules, Bergis, Ed Summers, and Vernon Mitchell. 2018. "Documenting The Now White Paper. Ethical Considerations for Archiving Social Media Content Generated by Contemporary. Social Movements: Challenges, Opportunities, and Recommendations." https://www.docnow.io/docs/docnow-whitepaper-2018.pdf

Lor, Peter, and Johannes Britz. 2004. "A moral perspective on South-North web archiving." *Journal of Information Science* 30, 6: 540–49. https://doi.org/10.1177/0165551504047925

Lynch, Clifford. 2017. "Stewardship in the 'Age of Algorithms'." *First Monday* 22, 12. https://doi.org/10.5210/fm.v22i12.8097

Milligan, Ian. 2018. "The Ethics of Studying Geocities." Ethics and Archiving the Web

Conference. New York: New Museum.
https://ianmilli.wordpress.com/2018/03/27/ethics-and-the-archived-web-presentation-the-ethics-of-studying-geocities/

Nielsen, Jane. 2021. "Quantitative approaches to the Danish Web Archive." In *The Past Web: Exploring Web Archives*, edited by Daniel Gomes, Elena Demidova, Jane Winters, and Thomas Risse, 165–79. Cham: Springer Nature.

Quintanilla, Olivia, and Jeanelle Horcasitas. 2023. "A call to research action: transnational solidarity for digital humanists." In: *Debates in the Digital Humanities* 2023, edited by Matthew Gold, and Lauren Klein. Minneapolis: University of Minnesota Press.

Padilla, Thomas. 2019. *Responsible Operations: Data Science, Machine Learning, and AI in Libraries*. Dublin, OH: OCLC Research. https://doi.org/10.25333/xk7z-9g97.

Pendergrass, Keith, Walker Sampson, Tim Walsh, and Laura Alagna. 2019. "Toward Environmentally Sustainable Digital Preservation." *The American Archivist* 82, 1: 165–206.

Priem, Karin, and Ian Grosvenor. 2022. "Future Pasts: Web Archives and Public History as Challenges for Historians of Education in Times of COVID-19." In *Exhibiting the Past. Public Histories of Education,* edited by Frederik Herman, Sjaak Braster, and Maria del Mar del Pozo Andrés, 177–96. Berlin, Boston: De Gruyter Oldenbourg. https://doi.org/10.1515/9783110719871-009

Rollason-Cass, Sylvie, and Scott Reed, "Living Movements, Living Archives: Selecting and Archiving Web Content During Times of Social Unrest." *New Review of Information Networking* 2, 1–2: 241–47.

Ruest, Nick, Jimmy Lin, Ian Milligan, and Sam Fritz. 2020. "The Archives Unleashed Project: Technology, Process, and Community to Improve Scholarly Access to Web Archives." *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (JCDL '20), 157–166. New York: Association for Computing Machinery. https://doi.org/10.1145/3383583.3398513

Schafer, Valérie, Gérome Truc, Romain Badouard, Lucien Castex, and Francesca Musiani. 2019. "Paris and Nice terrorist attacks: Exploring Twitter and web archives." *Media, War & Conflict* 12, 2: 153–70. https://doi.org/10.1177/1750635219839382

Schafer, Valérie, and Ben Els. 2020. "Exploring special web archive collections related to COVID-19: The case of the BnL An interview with Ben Els (BnL) conducted by Valérie Schafer (C^2DH, University of Luxembourg)." *WARCnet Papers*. Aarhus: WARCnet https://cc.au.dk/fileadmin/user_upload/WARCnet/Schafer_et_al_COVID-19_BnL.pdf

Schafer, Valérie, and Jane Winters. 2021. "The values of web archives." *International Journal of Digital Humanities* 2, 129–44.

Schostag, Sabine. 2020. "The Danish Coronavirus web collection – Coronavirus on the curators' minds." *International Internet Preservation Consortium Blog,* July 29, 2020.
https://netpreserveblog.wordpress.com/2020/07/29/the-danish-coronavirus-web-collection/

Sönmez, Çağıl, Arzucan, Özgür, and Yörük Erdem. 2016. "Towards building a political protest database to explain changes in the welfare state." *10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities.* https://doi.org/10.18653/v1/W16-2113

Winters, Jane. 2017. "Coda: Web archives for humanities research – some reflections." In *The Web as History: Using Web Archives to Understand the Past and Present*, edited by Niels Brügger, and Raph Schroeder, 238–48. London: UCL Press.

Zuanni, Chiara. 2022. "Contemporary Collecting in a Pandemic: Challenges and Solutions for Documenting the COVID-19 Pandemic in Memory Organizations." *Heritage* 5, 4: 3616–27. 10.3390/heritage5040188