We're all experts now? Archiving public health discourse in the UK Web Archive

Alice Austin, Leontien Talboom

Abstract: Emerging from COVID-19 collecting initiatives that underscored the fragility of online health discourse, the Archive of Tomorrow was an ambitious collaborative project that set out to curate a representative and diverse collection of public health websites in the UK. The project encountered a number of challenges, such as technical barriers in capturing interactive and dynamic sites, ethical considerations concerning how disputed or outdated information might be responsibly made available to researchers, and philosophical questions about how 'health information' is to be defined. This chapter reports on the outcomes of the project and discusses future directions for improving the production and use of large-scale archived web collections.

Keywords: collection development, metadata, legal deposit, health information, misinformation.

As other chapters in this volume have reflected, the Covid-19 pandemic catalyzed a rising effort to archive information from the web, with libraries and archives rushing to document the traces of a 'new normal' that saw life for many move online. The speed at which information entered the public realm and was subsequently discussed, debated, and debunked served to crystalize some of the key issues that heritage organizations encounter when trying to capture a nebulous and rapidly-evolving information landscape: How do we capture history when it is still happening? How do we respectfully and responsibly reflect the dissent and divisions in a moment without a single, unifying narrative? And what preparations can we make now to meet the needs of the researchers of the future?

In an attempt to document the 'unprecedented' and historic events that unfolded at the start of the decade, numerous institutions and community groups turned to web archiving as a means of ensuring collecting could continue remotely. As Amanda Greenwood (2022) has detailed in a thorough literature review, the scope of these initiatives varied greatly—from global to local and community to individual—serving to reflect the myriad ways in which the pandemic's impacts were felt (Greenwood 2022). The potential (and limitations) of the web and (by extension) web archiving as a means of producing a historical record that serves as both correlative and counterpart to institutional and state narratives has long been recognized in the literature (Milligan, Ruest, and Lin 2016; Barrowcliffe 2021) and many curators experimented with participatory archiving practices to this end, inviting contributions of personal narratives, reflections, and experiences in a manner that Kerrie M. Davies has termed

Alice Austin, University of Edinburgh, Scotland, United Kingdom, alice.austin@ed.ac.uk, 0009-0007-5586-2571
Leontien Talboom, University of University of Cambridge, United Kingdom, lkt39@cam.ac.uk, 0000-0001-7408-5471

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

Alice Austin, Leontien Talboom, *We're all experts now? Archiving public health discourse in the UK Web Archive*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0413-2.25, 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. 295-307, 2024, published by Firenze University Press, ISBN 979-12-215-0413-2, DOI 10.36253/979-12-215-0413-2

'crowd-coaxing' (Davies 2023). As Tizian Zumthurm and Stefan Krebs (2022) have reflected, however, creating a web archive that comprehensively represents the divergences of experience in a moment like the pandemic is a complex endeavor, as underscored by recent efforts to explore how web archives are understood and used in online discourse (Odgen, Summers, and Walker 2021) and even deployed and weaponized in the service of misinformation (Acker and Chaiet 2020). Such work has prompted greater consideration on the part of archivists, curators, and other memory workers as to the role of web archive collections in an ever-evolving information nexus.

Evolving from these attempts to document the pandemic and its attendant 'infodemic', the Archive of Tomorrow (AoT) was an ambitious pilot project that sought to build a test-bed collection in the UK Web Archive (UKWA) through which such questions could be explored. Led by the National Library of Scotland in partnership with Cambridge University Library, Bodleian Libraries Oxford, and the University of Edinburgh, the project was funded by a Wellcome Research Resources Award in Humanities and Social Science and ran from February 2022 to April 2023. The technical team comprised three web archivists (each based within a university library), a project manager, a metadata analyst, and a rights officer (all based within the National Library of Scotland). A research data engineer was initially engaged for the project but resigned from the post at an early date. Throughout the project the team benefitted from the support and expertise of colleagues at the British Library, and was guided by an Advisory Board comprising academic researchers and industry experts from a diverse range of relevant disciplines.

The project aims were both tangible and exploratory. The primary goal was to curate a 'research-ready' collection of websites within the UKWA around the theme of Health Information and Misinformation. This curated collection would then serve as a test-bed around which options for metadata, computational analysis, ethics, and rights issues could be explored. The team further aimed to build a network of researchers across relevant disciplines in order to involve potential users in the process of building, evaluating, and using collections. It was anticipated that the project would serve to concretize recommendations for future work and provide a focus for advocacy for change to make web archives more representative, inclusive, and open for health research. After setting out the legal and technical contexts in which the project operated, this chapter will outline the processes and deliberations involved in the production of a large-scale web archive collection, describe the challenges encountered when trying to capture such a hotly debated area, and outline areas for future work.

2. Capture

2.1 Legislative background of the UK Web Archive

The UKWA is a partnership of the six Legal Deposit Libraries (LDLs) that performs the web function of the LDLs' legislative responsibility to collect and preserve a copy of all material published in the UK and Ireland. The UKWA has been systematically collecting non-print material since 2013, with the majority of material being captured through an annual domain crawl that attempts to make a copy of any content published to a website with a recognizable UK top-level domain (e.g., .uk, .scot), or hosted on a server physically located in the UK (identified via a GeoIP lookup). The yearly crawl is supplemented by curated collecting, achieved by manually adding targets to the Annotation and Curation Tool (W3ACT), a web-based interface that allows a user to create an entry for a specified URL, establish parameters such as depth or frequency of a crawl and record metadata for description and rights-management purposes. Access to content archived in the UKWA is by default restricted to users at computer terminals onsite in Legal Deposit Libraries, unless open access permission has been explicitly granted by the website owner. The project's dedicated Rights Officer was responsible for corresponding with site owners and issuing these permission requests. If significant concerns about making a site's content accessible were identified, the Rights Officer may choose not to pursue access permission.

2.2 Technical specification

The UKWA uses Heritrix, an open-source web crawler written in Java to perform crawls. Crawled content is written to a WARC file and stored alongside metadata necessary for interpreting the crawl. There are technical and legislative limitations to what the UKWA is able to capture. For example, dynamic pages (pages in which user interaction initiates server-side scripts) present difficulties as the crawler is unable to perform any such interactions—and so any content retrieved by querying a database cannot be captured. Material that requires login credentials to access is excluded on two fronts, as the crawler cannot input such credentials and the legislation that enables non-print legal deposit only covers material that is made publicly available. The legislation also does not extend to broadcast material that is primarily audio-visual, excluding content on video-centric platforms such as YouTube or TikTok.

2.3 Scope and focus

Establishing the boundaries and scope of any collecting activity is always a challenge, and the nature of the subject in question only compounded this

difficulty. The scoping process was a continuous one. Initial efforts used the Collection Development Framework created by the Web Archiving Team at the University of North Texas. This comprehensive document was particularly useful in encouraging an holistic view of the collection throughout the curation lifecycle, and was invaluable in guiding discussions about the kinds of material that it was anticipated could be encountered during the project and the potential issues we might have to navigate.

Health is an exceedingly broad term and area of study, and the project team were keen to ensure that collecting did not focus solely on the biomedical but also reflected how debates around physical, mental, and social wellbeing intersect with other issues, such as those of politics, economics, and technology. Early in the project web archivists met with academics and students at their respective institutions to gauge areas of interest for current health-related research, and surveyed published research in these areas in order to consider how existing studies might be extended into the digital realm. It was initially anticipated that the project would focus predominantly on the issues of misinformation and disinformation that had accompanied the Covid-19 pandemic, however as the project unfolded it became clear that that was just one aspect of a much larger picture of how the internet is used to find, share, and debate issues of health. One particular observation to emerge from these conversations concerned the value of social media data for research; however this also emphasized the challenges of using archived social media content for research within the confines of the legal deposit legislation.

2.4 Identification of material

A number of different methods were employed for the identification of material, both systematic and serendipitous. Existing directories (for example, a list of health-related charities exported from the Charity Commission Register) were used to locate content, and participatory collecting methods—engaging with health researchers to determine particular areas of interest—was also a valuable approach. Web Archivists also experimented with the use of various search engines to explore how the results differed, and noted that it was challenging to record the impacts of such tools on the resulting resources that were found as many of these impacts were obscured through algorithms designed to promote and suppress particular types of content.

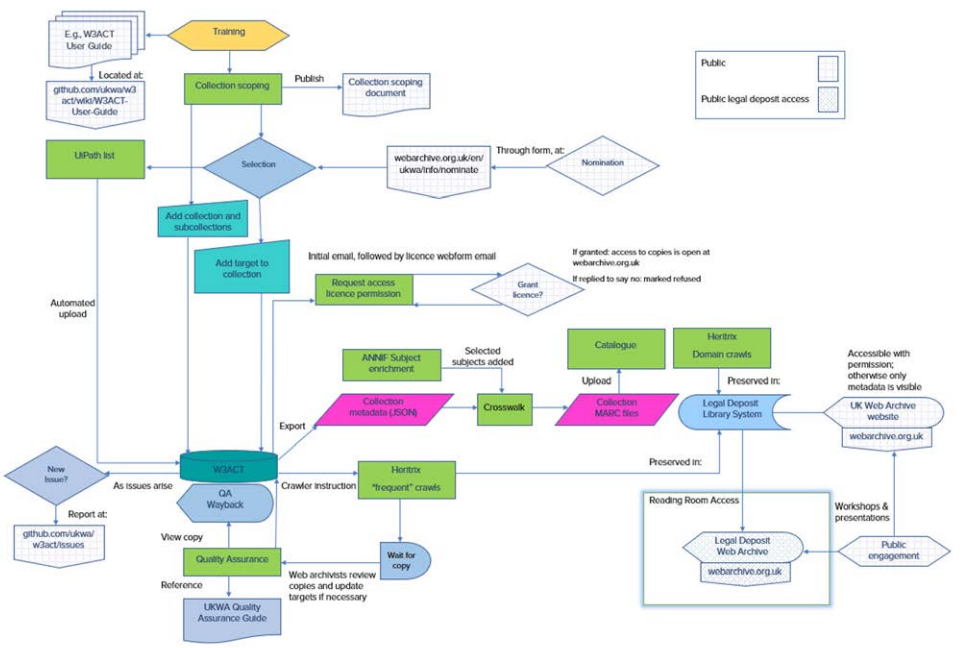
The collection of social media was an important objective for the project. Web archiving technologies allow for greater representation of ‘everyday’ or ‘street-level’ opinions through collection of social media resources, and collecting from platforms such as Twitter, Reddit, and Mumsnet provided a balance to the commercial algorithms that tend to promote commercial

enterprises. Again, there are technical and legislative limits to what can be captured from these platforms, and in order to comply with non-print legal deposit legislation, selection had to be meticulous: for example, search terms for specific topics on Twitter were combined with location tags (e.g. '(#wellbeing) near:edinburgh'), and collection from Reddit was limited to forums that explicitly located themselves in the UK (r/UKNurses, r/UKMCPatientCommittee/, r/UKantilockdown).

2.5 Workflow

The illustration below provides an overview of the team's workflow. The Web Archivists selected sites for inclusion, and records were created in W3ACT for each target URL either manually (URL-by-URL) or by means of a UiPath sheet for bulk creation. Each targeted URL was recorded in a shared-access spreadsheet, which also served as a mechanism for the Web Archivists to flag sites with potential access concerns to the Rights Officer. After an initial crawl had been performed the site was checked for quality, however as the collection grew the available capacity for quality checking diminished.

Figure 1: Project workflow



3. Collection overview

The resulting collection—named "Talking About Health" (TAH)—comprises around 3,500 individual targets. The TAH collection is subdivided along various lines through the use of W3ACT's tagging feature, whereby a target can be 'tagged' or marked for inclusion into any number of different collections or subcollections. As collecting progressed a number of areas that warranted a deeper level of collection emerged, and the tagging function has been used to group targets on focused areas such as dental health, substance abuse, menopause, and nutrition.

As a matter of course, sites were tagged into the main collection at the highest level appropriate, but in some cases more specific pages have also been targeted and tagged into lower-level sub collections. For example, the British Heart Foundation's website[1] has been included in the main collection, and in addition, a child page dedicated to 'heart-healthy recipes' has been tagged into the Nutrition sub-collection. This provides a more direct entry point to the specific content of that page, but also ensures the wider context of the site as a whole is preserved.

4. Access

4.1 Structure

Navigation of the collection is achieved through W3ACT's tagging feature which, as described above, provides some high-level description of targets. The project team was keenly aware of the narrative power of archival description and the potential for the terms used to convey a judgment of value or otherwise reflect curatorial bias. As collecting progressed it became apparent that the initial framework developed by the project team was not sufficient for representing the breadth of information and subjects covered, and that flexibility in the collecting framework was needed to allow current and future users to add areas or topics of interest beyond project completion to help ensure the collection remains relevant over time.

As the group experimented with different structures and descriptive frameworks it was interesting to observe how the diverse discipline backgrounds of the team members influenced these discussions: those who worked primarily in a library setting were inclined towards using tags to provide a description of the content and focus of a site, whereas those colleagues operating in a more archival context generally attempted to use terms to describe the creator as best as possible. As none of the team have a medical background and are therefore unqualified to make such

[1] https://web.archive.org/web/20240515150135/https://www.bhf.org.uk/

assessments, a decision was made to avoid the use of any labels that could be read as a judgment on the efficacy or suitability of a practice or its proponents. This appeared to be a straightforward approach in the context of medical information, but became a more complex challenge as the foci moved towards areas where medicine intersects with the legal, social, and discursive contexts in which it operates. The team combined these approaches to develop a multifaceted structure that provides a broad descriptive overview of the content, but also invites critical deliberation as to the circumstances and context of a site as a publication.

This process was significantly informed by the efforts of the Rights Officer to gather feedback from site owners and publishers on the hesitations, concerns, and barriers that held them back from providing access permission. A number of site owners requested information on how their site would be categorized within the collection, and voiced concern that their site might be misrepresented. The initial name of the project ('Health Information and Misinformation') was found to be causing particular consternation, and the collection name was subsequently changed to 'Talking About Health' in order to reflect the discursive nature of the content. Such insights into the concerns of site owners and creators were also influential in prompting the project team to consider the responsibilities that the UKWA has as a 'secondary publisher' of material, and how those responsibilities extend to both content creators and potential users. With regard to content creators, it is important that we fully consider the extent to which archiving practices are understood and anticipated by content creators, and what recourse those who find their content included in an archive collection have to challenge the preservation and republication of material about them. Such considerations are necessary for any kind of collecting from social media, but the team felt this was particularly crucial given the sensitive nature of much content about health, and the probability that content could have been made at a moment of crisis on the part of the poster.

4.2 Rights

The most significant factor affecting access to content in the UKWA is the legal deposit legislation that enables capture: a site owner must provide explicit approval in order for archived copies of their site to be viewed from outside legal deposit library premises. Standard procedure is for this permission to be requested through an automated email sent via the curation tool but a large proportion of these requests usually go unheeded, with the result that less than 1% of UKWA content was accessible offsite at the outset of the AoT project.

In the course of the project the dedicated Rights Officer issued offsite

access permission requests to 1,840 email addresses against 58% of the targets collected. Permission was granted for 7% of those requests and refused for 3%. The Rights Officer reported that reasons for refusing access permission were varied: some website owners saw no benefit of granting permission, and some creators were concerned that archiving itself may ultimately be damaging, either in terms of increased reputational risk from information remaining in publicly accessible online spaces beyond the original intention, risk to data subjects or risk of copyright infringement, or in terms of diverting website traffic to the archived resource and away from a live site.

At the close of the AoT project, 21% of the collected targets were accessible from outwith a legal deposit library, with the table below providing an overview of the levels of open access achieved across the subcollections:

Table 1. Breakdown of license status by category.

Heading	Remote access permission granted (%)
Blogs & social media	10.6
Charities & non-profit organizations	44.3
Commercial/industrial health sector	6.7
Focus	20.2
Government	58.6
Health professions	20.6
NHS	90.2
News & commentary	9.5
Politics & health	33.7
Research	23

While this is a vast improvement on the UKWA average, it still leaves a significant proportion of the collection inaccessible to the majority of researchers. In order to explore options for improving access a review of existing license agreements and approaches was conducted, and it was noted that the non-print legal deposit framework already recognized and respected where content had been published under an Open Government License. A paper was submitted proposing that a similar approach be adopted for content published under other open licenses such as Creative Commons licenses. This proposal was approved by the Legal Deposit Libraries Committee, and implementation is now underway.

4.3 Metadata

A key goal of the project was to investigate making use of the collection metadata as a means of ‘surrogate’ access to closed captures: that is, while full computational access to captured content is not possible within the framework of the legal deposit legislation, it was hoped that making

collection metadata available for use would encourage experimentation with this dataset and the identification of new tools for 'reading' and interpreting large-scale collections such as this. In collaboration with colleagues at the British Library, a tool was developed for obtaining robust, structured JSON exports of metadata from the UKWA curation tool via API. This contains technical information such as the URL(s) targeted, the date a target was created, the frequency and depth of a crawl, and rights status, along with any descriptive metadata, and significantly improves access to (and supports reuse of) UKWA metadata.

This metadata work was also valuable in helping to improve the representation of the archived web in library catalogs and to encourage a critical consideration of the material as archival source. A frequent hesitation voiced by creators was the concern that preservation in the archive might inadvertently lead a user to access outdated or inaccurate information, and the project team were conscious of the need to ensure archived web content was properly contextualized in order to protect creators and potential users. Archived copies appear with a blue 'UK Web Archive' banner at the top, marked with the capture date in order to clearly indicate that they are not live pages, however some website owners expressed concern that this branding may not be enough to discourage misunderstanding. Led by the Metadata Analyst and Rights Officer, the team explored options for stewarding access to archived captures and communicating the nature of archived web content to users. In response, the following short statement was included in the catalog records for subcollections identified as containing a higher incidence of broadly sensitive material:

> Please take care when accessing, using and sharing information from this collection, which may contain outdated or offensive language, sensitive information about living individuals, or otherwise sensitive material.

The UKWA-derived metadata was also used to produce library catalog records for each target. The Metadata Analyst developed a crosswalk to prepare metadata for ingest into the ALMA cataloging system used by the National Library of Scotland. This repurposed descriptions generated by the web archivists and technical metadata to populate a MARC record for each target, with the intention that such surrogate records would not only facilitate a more user-friendly means of accessing the collection, but also serve as a public-facing source of information about the existence of the UK Web Archive and the archive web more generally. Additionally, it was hoped that representing the archived websites alongside other more 'traditional' sources would serve to impress that they should be understood with the same level of critical deliberation as to the circumstances and context of a site as a publication, and of the collection as an entity.

4.4 Transparency

In an effort to further shed light on the archival processes that contribute to a collection, the exported metadata is accompanied by documentation that provides further information on the technical and non-technical circumstances of the collection's creation. This is based on the 'datasheets for datasets' framework adapted for web archive collections by Emily Maemura, and members of the project team met with Maemura in November 2022 and began drafting a datasheet based on Maemura's questions about a dataset's provenance, parameters, omissions. The resulting document provides context for the Talking about Health collection data made available to researchers, communicating technical details and outlining potential uses of the data. Similar efforts were made to find routes to 'open up' the collection development process to potential researchers, and the team participated in a number of researcher-focused workshops and seminars discussing questions around what contextual and collection-development information researchers need access to and the best means for delivering these. A Discourse channel was established as a space in which documentation about the project's background, aims, and approaches could be shared, and the project team also used this space as a means to share relevant reading materials and other resources.

5. Lessons learned

The AoT project offered a valuable opportunity to study the various processes, workflows, deliberations, and interactions that contribute to the development of a large-scale archived web collection, and a number of useful lessons can be observed. Firstly, the AoT project reinforced the need for continued and sustained resourcing for web archiving work. All web archivists were employed on a part-time basis, and this may have contributed to the shortfall in targets collected—the total sits around 3,500, rather than the 10,000 aimed for, reflecting the level of work necessary to curate a collection around such a sensitive and potentially fraught topic. The presence of the dedicated Rights Officer on the project was particularly illustrative of how transformative adequate resourcing can be: the feedback gathered from website and content creators was invaluable in helping to guide the collection development process, and as rights management work is usually performed by a web archivist alongside their selection, curation, collection, and quality assurance duties, the majority of web archiving staff simply do not have the capacity or resources necessary for such discussions. The value of these conversations can be seen in the final statistics: 21% of this collection can be accessed from outwith a legal deposit library, compared with 1% of the UKWA collections as a whole.

A related (if not unexpected) finding concerns the value of collaboration between individuals and institutions when approaching topics of this scale. The final collection is broad in its coverage with over 70 different subtopics represented. Such broad coverage could not have been achieved by a single institution, and would not have been possible without the input of a range of stakeholders from a variety of disciplines and backgrounds. Scoping the collection was an iterative and discursive process, and having web archivists located within different academic environments meant the team was able to discuss the collection development process with an established group of researchers from health and health-adjacent disciplines. Not only was this extremely valuable in pushing collection efforts beyond the strict categories of 'medical' or 'health' information and in encouraging the team to consider the relative value of different types of information sources, but was also instrumental in helping the team to better understand the needs, concerns, and ambitions of potential users.

Recognizing the subjective and discursive nature of the collection development process, the team considered how curatorial decision-making might be better communicated to potential users and content creators. This could be achieved through access to further technical metadata (for example, information on which collections a target appears in, when it was added to a collection, or how the crawl parameters for a target have changed over time) or reflective documents that describe and communicate how a term or subject has been interpreted. Similarly, while the ethical issues that arise when capturing personal narratives and discourse cannot be avoided, improved documentation of how such concerns have influenced and shaped collecting may go some way to strengthening not only the research value of this material, but the relationships with creators and users.

6. Conclusion

The AoT project offered a rare opportunity for observing the processes and deliberations involved in collaboratively building large-scale collections of archived web content. The collaborative nature of the project allowed for better identification of the points where decisions were based on assumptions and expectations that required probing. This in turn impressed the need for digital historical representations like web archives to provide clear contextual and provenancial information alongside records, and to integrate the mechanisms of archival representation into our understanding of context. Such information will allow users to fully interrogate the relevance, integrity, and reliability of records for themselves.

The project also demonstrated the impact of dedicated resourcing for web archiving, particularly with regard to rights work. By engaging with web creators and site owners, the Rights Officer has been able to articulate

the concerns and hesitations that prevent rights-holders from granting access, and the next step will be to seek means of addressing and alleviating these concerns. First among these will be efforts to increase the visibility and understanding of web archives amongst the general public, and continued advocacy of web archives as a key part of UK research infrastructure.

Finally, while the experiences of the AoT project found that there is researcher interest in and appetite for archived web resources, it also further illustrated the hurdles that must be overcome in order for collections like this to be more widely utilized in research. As has been noted, the biggest barrier to use of the collections are the conditions of the legislation under which they are created, and it is anticipated that expanding the recognition of open licenses to include creative commons licensing will significantly increase the proportion of the archived web that is accessible for research. Improving the ease of the export of data from W3ACT is a constructive step towards an alternative means for access, and providing a route to datasets that can be used for computational analysis while still respecting the boundaries of legal deposit legislation represents significant progress towards increasing the visibility—and with hope, the research use—of the archived web.

With thanks to the Archive of Tomorrow project team: Eddie Boyle (Research Data Engineer, National Library of Scotland); Cui Cui (Web Archivist, Bodleian Library, University of Oxford); Mark Simon Haydn (Metadata Analyst, National Library of Scotland); Mary Garner (Project Manager, National Library of Scotland); Jasmine Hide (Rights Officer, National Library of Scotland); Agnieszka Kurzeja (Metadata Coordinator, University of Cambridge); Eilidh MacGlone (Web Archivist, National Library of Scotland). The full project report is available via the British Library's Research Repository https://doi.org/10.23636/6q6k-8369

References

Acker, Amelia and Mitch Chaiet. 2020. “The weaponization of web archives: Data craft and COVID-19 publics.” *Harvard Kennedy School (HKS) Misinformation Review* 1, no. 3. https://doi.org/10.37016/mr-2020-41

Barrowcliffe, Rose. 2021. “Closing the Narrative Gap: Social Media as a Tool to Reconcile Institutional Archival Narratives with Indigenous Counter-Narratives.” *Archives and Manuscripts* 49, no. 3: 151–66. http://www.doi.org/10.1080/01576895.2021.1883074

Davies, Kerrie M. 2023. “Crowd Coaxing and Citizen Storytelling in Archives of Crisis.” Life Writing 20, no. 2: 351–365. https://doi.org/10.1080/14484528.2022.2106611

Greenwood, Amanda. 2022. “Archiving COVID-19: A Historical Literature Review.” *The American Archivist* 85, no. 1: 288–311. https://doi.org/10.17723/2327-9702-85.1.288

Milligan, Ian, Nick Ruest, and Jimmy Lin. 2016. “Content Selection and Curation for Web Archiving: The Gatekeepers vs. the Masses.” In JCDL '16: *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*, 107–110. https://doi.org/10.1145/2910896.2910913

Ogden, Jessica, Ed Summers, and Shawn Walker. 2021.“Patterns of Use: Conceptualising the role of web archives in online discourse.” Paper presented at *Fourth Research Infrastructure for the Study of Archived Web Materials (RESAW) Conference: Mainstream vs. marginal content in Web history and Web archives*, University of Luxembourg, Luxembourg, June 17–18. https://hdl.handle.net/1983/59169b00-10ac-435b-8179-f6b88cff9c1c

Zumthurm, Tizian and Stefan Krebs. 2022. “Collecting Middle-Class Memories? The Pandemic, Technology and Crowdsourced Archives.” *Technology and Culture* 63 no. 2: 483–493. https://doi.org/10.1353/tech.2022.0059