Lost Archives? Preservation of Ephemeral COVID-19 Data for Research

Tolulope Balogun*

ARTICLE INFO

Article history:
Received 31 March 2025
Revised 11 May 2025
Accepted 04 June 2025
Online First 05 June 2025

Keywords:
Pandemic,
COVID-19,
Web Archiving,
Digital History,
Digital Preservation,
Africa

ABSTRACT

This study explores the significance of web archives in preserving COVID-19-related data and information, emphasizing their potential for future research and their role in documenting major events like the pandemic. Adopting a qualitative research approach, the study utilizes web content analysis to examine the availability of selected COVID-19-related websites on the Internet Archive's Wayback Machine between January 20, 2020, and April 17, 2020. The analysis focuses on the presence of these webpages and the frequency of their archival crawls. Findings reveal that while the Internet Archive actively harvested COVID-19-related webpages, the frequency of crawls varied, and some pages were not archived as frequently as needed. Rather than examining update rates, the study highlights the accessibility and archival frequency of these webpages. The research is limited in scope, concentrating on a selected group of COVID-19-related websites and emphasizing the importance of web archiving without delving into technical analyses or the issue of online misinformation. Despite these limitations, the study highlights the value of web archives as essential tools for preserving digital records of significant events and demonstrates their potential for future research. By showcasing the function and value of web archiving, this study contributes to ongoing discussions about its role in research, particularly in an interdisciplinary context. It also provides a foundation for discussions on web archiving in Africa, offering insights into its relevance and possibilities for further exploration.

1. Introduction

The internet has revolutionized the way people access, create, and share information. The COVID-19 pandemic in 2020 disrupted commercial activities and restricted movement, requiring the web to serve as a vital global information and communication source. The COVID-19 pandemic prompted a global surge in information dissemination through various digital platforms, including news media

* Department of Information Science, College of Human Sciences, University of South Africa, Pretoria, South Africa (balogtb@unisa.ac.za) (First & Corresponding Author)
International Journal of Knowledge Content Development & Technology, 16(1): 67-82, 2026.
http://dx.doi.org/10.5865/IJKCT.2026.16.1.067

websites, social media, and dedicated health organization portals (Mcclain et al., 2021; Ferguson, Merga, & Winn, 2021; Wu, 2024). Although the web played a vital role in the dissemination of information and communication during the COVID-19 pandemic, it is important to note that it was also used as a tool to misinform, disinform and generally spread fake news and cause unnecessary panic prompting the World Health Organization (WHO) to declare an 'infodemic' (WHO, 2020a). The impact of COVID-19 pandemic has stimulated researchers to strive to understand how it affected different aspects of society (Jaumotte, 2023; Sohrabi et al., 2021; Sindhwani, Kumar, & Saddikuti, 2022).

In this context, preserving the vast amount of digital information generated during the pandemic has become increasingly important. Web archiving has emerged as a solution for preserving online content (Hendry & Stock, 2014). It involves the collection, storage, and retrieval of web pages to create an archive of online content (Vlassenroot et al., 2019). The archived pages can then be accessed later, preserving the original context and content. Web archiving is an important tool for digital preservation, providing an opportunity to preserve ephemeral online data for future generations.

Therefore, the aim of this article is to discuss the preservation of ephemeral COVID-19 data, highlighting the significant sources of information during the pandemic, and assessing the efforts made to archive this critical information. The study also seeks to highlight the importance of web archiving for future research and digital history preservation, particularly in the context of public health crises. With this objective in mind, certain limitations and delimitations are outlined to set the scope of this paper. First, the paper highlights the importance of research on web archives, focusing specifically on COVID-19-related websites. This focus represents both a limitation and a deliberate delimitation, as it narrows the scope to a select number of websites rather than a comprehensive or representative sample. The selected websites are adequate to explain the concept of web archiving and underscore its significance for future research. This paper provides a descriptive overview of web archiving and its prospects, intended to benefit a wide range of interdisciplinary audiences. However, it does not cover technical aspects of web archiving, such as selection, appraisal, and curation. Additionally, the paper does not delve into the issue of online misinformation and fake news during the COVID-19 pandemic. Instead, the paper, which is largely descriptive, aims to illustrate the importance of web archiving and to stimulate further research in this field, particularly in the context of Africa.

2. Literature review

2.1 The Ephemeral Nature of Online Data and Its Preservation Challenges

The ephemeral nature of online data poses significant challenges for web archiving (Davis, 2014; Duncan, Sumitra, & Blumenthal, 2016). Ephemeral online data refers to digital content that is volatile, rapidly changing, and has a short lifespan. This includes social media posts, news articles, and other digital content that is often not preserved through traditional archival methods. The rapid

evolution of technology and the internet presents a challenge for preserving ephemeral online data (Brügger et al., 2017). As technology evolves, digital content is often presented in new and innovative ways, making it challenging to capture and preserve the content accurately.

Despite the popular impression that contents on the internet last forever, websites cease to exist daily (Slania, 2013). According to Antracoli et al. (2014: 156), "the ubiquity and public accessibility of the World Wide Web tend to foster an impression of permanence." Several studies have been able to highlight and establish the ephemeral nature of the web or decay of URL contents (e.g., Habibzadeh, 2013; Slania, 2013; Schneider & Foot, 2004; Kumar & Prithviraj, 2015; Antracoli et al., 2014; Duncan & Blumenthal, 2016; Duncan, 2015; Brügger, 2005). In fact, it is estimated that around 80% of online/web content changes or completely disappears within a year (Brügger, 2005; Gomes et al., 2011), and this instability poses a serious challenge for researchers (Kumar & Prithviraj, 2015). Although the ephemeral nature of the web has been established, the web is considered as a mixture of ephemeral and permanent (Schneider & Foot, 2004). Schneider and Foot (2004) discussed the permanence of the web in the context of its form in which information is presented to be transmitted, which is a shared characteristic with other forms of media such as print, sound recordings and films. However, Schneider and Foot (2004: 115) acknowledged that the permanence of the web is fleeting, and unlike other permanent media, a website which is updated frequently mostly overrides each previous edition or versions. Indeed, it is important to note that most COVID-19 related websites were frequently updated, making it relatively difficult to compare some of the websites with their previous versions in less than a week in some cases.

2.2 Research value of Web archives

The growing recognition of the web as a critical resource for research across various fields has led to increased interest in the preservation of online content through web archiving. Web archives offer born-digital primary research resources that are particularly valuable for digital scholars (Vlassenroot et al., 2019; Nielsen, 2016). Despite this, there remains limited awareness within the humanities and social sciences regarding the full research potential of web archives, emphasizing the need for greater scholarly engagement with these digital repositories (Costea, 2018). The importance of web archiving extends beyond mere preservation as it plays a crucial role in enabling future research on the web itself. However, current web archiving efforts have not yet achieved the reliability and flexibility needed to fully support long-term research needs (Dougherty & Meyer, 2014). For historians and researchers in other disciplines, web archives present both opportunities and challenges. While they offer significant research value, historians often struggle to replicate traditional research processes when working with these digital collections, necessitating the development of new methodologies (Belovari, 2017).

Web archiving efforts related to COVID-19 have emerged globally to preserve digital content about the pandemic. For instance, the National Library of Medicine has collected over 2,500 items, including websites, blogs, and social media posts, to create a comprehensive historical record (Speaker & Moffatt, 2020). Also, strategies like "screensampling," where archived conspiracy content is shared on social media platforms, contribute to the "misinfodemic" (Acker & Chaiet, 2020). These efforts

highlight the significance of web archiving during global health events while also revealing the potential for misuse in spreading harmful information. Web archives are particularly valuable for preserving the ephemeral information published online, which would otherwise be lost. This preservation is crucial for humanities research, where such content can serve as important primary sources (Gomes & Costa, 2014). Additionally, the ability to search web archives offers the potential for novel research practices and analytical approaches that go beyond the traditional single URL access model, enabling more comprehensive studies of digital phenomena (Ben-David & Huurdeman, 2014). Despite the clear research benefits, methodological and ethical challenges persist. For example, the use of web archives to study social media users' communicative practices provides a rich data corpus, but researchers must navigate the complexities associated with such studies (Lomborg, 2012). Moreover, while national libraries often house valuable web archive collections, defining meaningful collections and addressing the accompanying ethical and methodological issues remain significant challenges for researchers (Stirling, Chevallier & Illien, 2012). In the context of historical research, the web is increasingly recognized as a significant resource, and web archiving is seen as essential for conducting history in the digital age. However, to fully realize the potential of web archives, scholars must overcome the challenges associated with their use and continue to advocate for more reliable and feasible archiving solutions (Sheldon, 2019).

However, it is also important to note the use of web archives in disinformation research. Weigle (2023) discussed how journalists and researchers utilize web archives to investigate changes in webpages, study archived social media content, including deleted posts, and analyze known disinformation, underscoring the role of web archives in understanding and combating misinformation.

3. Methodology

The study adopted a qualitative research approach, utilizing web content analysis to explore the preservation of ephemeral COVID-19 data for future research. A selective approach was employed in the choice of websites, focusing on those from key organizations that played a critical role in disseminating COVID-19 information during the pandemic. Specifically, the study analyzed websites from the World Health Organization (WHO), Worldometers, the Centers for Disease Control and Prevention (CDC), the South African Department of Health, and the Nigeria Centre for Disease Control (NCDC). These websites were selected due to their authoritative status and their significant contributions to the global and regional response to the COVID-19 pandemic. For instance, the World Health Organization (WHO) was included as a global leader in health emergencies, providing comprehensive and reliable data for international audiences. Similarly, the Centers for Disease Control and Prevention (CDC) offered detailed and region-specific insights, making it an important source for analysis. To represent Africa, the Nigeria Centre for Disease Control (NCDC) and the South African Department of Health were chosen, ensuring a balance between global and regional perspectives. These organizations' frequent updates during the pandemic made their websites ideal for assessing the Wayback Machine's ability to capture rapidly changing data. in addition, the accessibility of these sites through the Wayback Machine ensured the feasibility of conducting a detailed

analysis. While many other websites contributed to COVID-19 information dissemination, this focused approach allows the study to emphasize web archiving's potential and limitations, laying a foundation for future research that may include broader or less prominent sources.

Furthermore, the selective inclusion of these websites was justified by the need to assess how web archives can effectively preserve crucial data from leading health authorities and data aggregators. By concentrating on these specific sources, the study aimed to demonstrate the potential of web archiving tools, such as the Internet Archive's Wayback Machine, in capturing and storing valuable digital content for future research. Data collection was conducted using the Internet Archive's Wayback Machine, a powerful web archiving tool that captures and stores web pages for long-term access. Its extensive reach and accessibility make it a representative and reliable tool for studying web archiving in the context of the COVID-19 pandemic. Its established credibility and widespread use in academic and research settings further justify its exclusive use in this study. In addition, the Wayback Machine effectively captures global and regional data, as evidenced by its archiving of key COVID-19 information sources such as the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC), as well as data from African health organizations.

The study focused on web pages archived during the early phase of the COVID-19 pandemic, from January 20, 2020, to April 17, 2020, a period marked by the rapid spread of the virus and the implementation of global lockdown measures. The web content analysis examined the availability and frequency of archival of the selected websites, highlighting the importance of preserving these digital records for ongoing and future research. The study highlights the need for systematic preservation of COVID-19-related digital content, particularly from key health organizations, to ensure valuable data remains accessible for future analysis and research.

4. Limitations of the Wayback Machine for COVID-19 research

The use of the Internet Archive's Wayback Machine as a tool for web archiving in the context of COVID-19 research has the potential to introduce biases and limitations that could affect the results of the study. One potential bias is the selection bias, which may occur if certain websites or web pages are not archived, or if the archiving process does not capture all the relevant information. This could lead to a biased sample of data, which could affect the findings of the study. For example, if the Wayback Machine only archived pages from certain types of media outlets, it could lead to a bias towards certain perspectives or narratives about the pandemic. Another potential limitation of using the Wayback Machine as a tool for web archiving is the lack of real-time data, which may not accurately reflect the current discourse and public opinion about the pandemic. For example, the Wayback Machine captures the pages at a specific point in time, and it may not include the most recent updates or discussions about the pandemic, which could lead to an incomplete understanding of the narrative evolution. Also, using the Wayback Machine as a tool for web archiving may also introduce technical limitations that could affect the data. For example, some web pages may not be archived correctly, or certain types of web pages may not be able to be archived at all. This could lead to missing data or errors in the data, which could affect the findings of the study.

Thus, it is important to be aware of its potential biases and limitations of Internet Archive's Wayback Machine when using it in the context of COVID-19 research. A critical examination of the tool and its potential impacts on the study's findings should be done to ensure the reliability of the data and the conclusions derived from it.

5. Research Findings

This section presents the research findings from the study on the preservation of ephemeral COVID-19 data for research purposes. It details the significant sources of information during the COVID-19 pandemic, including prominent websites like those of the World Health Organization (WHO), Worldometer, the South African Department of Health, and the Nigeria Centre for Disease Control (NCDC). These platforms were instrumental in disseminating timely and reliable COVID-19 data and guidance.

5.1 World Health Organization (WHO)

The World Health Organization (WHO) is a specialized United Nations (UN) agency with the primary role of directing international health and leading its partners in global health response. The World Health Organization (WHO) has a webpage dedicated to the worldwide coverage of the COVID-19 hosted on its official website. The webpage accessible through https://www.who.int/emergencies/diseases/novel-coronavirus-2019 covered issues such as public advice, country and technical guidance, travel advice, situation reports, media resources, research and development, myth busters, a dashboard, among other information. The COVID-19 outbreak dashboard of WHO gave live updates with a global map showing the spread of the COVID-19 pandemic across the world. The WHO was considered as the most authoritative source of information for the COVID-19 during the pandemic. The situation reports were presented in PDF format and edited daily is a compilation of daily reports of the COVID-19 pandemic from 21st January 2020.

5.2 Centers for Disease Control and Prevention (CDC)

The CDC is a national public health institute in the United States of America (USA or US), and it is a US federal agency under the Department of Health and Human Services (Wikipedia, 2020). The main goal of the institute is to protect safety and public health by controlling and preventing diseases, disability, and disease in the US and internationally (Wikipedia, 2020). The CDC COVID-19 webpage accessible through https://www.cdc.gov/coronavirus/2019-ncov/index.html contains information about the virus with a focus on the US. Although the webpage covers a summary of global cases and situation reports under its Data and Surveillance section, it focuses more on the spread of the virus in the US. The cases of COVID-19 in the US are detailed and broken down by the states in the US and different demographics (age, race, ethnicity), which are not usually the kind of in-depth analysis covered by other websites.

5.3 Worldometer

Worldometer is a website run by an international team of researchers, developers, and volunteers with the goal of making timely world statistics available to people around the world (Worldometer, 2020b). The website, without affiliation to any government, political group or corporation was a major provider of global COVID-19 live statistics. Worldometer's data was trusted and used by several organizations around the world including BBC, The New York Times, the UK Government, John Hopkins University, Financial Times, etc. (Worldometer, 2020b). The Worldometer COVID-19 webpage accessible through https://www.worldometers.info/coronavirus/ covered live updates of the COVID-19 outbreak. The webpage reported the cases, deaths, recoveries of the COVID-19 pandemic around the world by country, territory, or conveyance. The web page also included a summary of the COVID-19 cases, deaths and recoveries displayed with graphs, and updates regarding the COVID-19 outbreak within a period of seven (7) days. The Worldometer webpage was a very popular source of COVID-19 statistics and information. Sometime in mid-March 2020, there was a DDoS attack on the website by hackers which compromised the data presented on the webpage.

5.4 South African Department of Health

The South African Department of Health has a COVID-19 online resource and news portal accessible through https://sacoronavirus.co.za/. The portal was dedicated to the dissemination of COVID-19 related statistics and information to South African citizens at the peak of the pandemic. The web portal covered vital information such as explanation of the virus, its symptoms, preventive tips, FAQs, statements from the minister of health, up to date COVID-19 statistics in South Africa, news and update, COVID-19 resources and guide, and other important information about the pandemic which are useful to South Africans.

5.5 Nigeria Centre for Disease Control (NCDC)

The NCDC is a Nigerian government agency established to deal with challenges of public health emergencies and for the enhancement of Nigeria's preparedness and response to epidemics through detection, prevention, and control of communicable and non-communicable diseases.

The NCDC website had a webpage accessible through https://covid19.ncdc.gov.ng/ where COVID-19 related statistics and information are uploaded. The contents of this page include COVID-19 highlights, case summary in Nigeria (total tests, confirmed cases, discharged, deaths), latest updates, regulations and guides, awareness videos, global situation reports, etc. The case summary in Nigeria was broken down into states to know what the situation is in all the affected states are so far. Some of the data on the website was retrieved from WHO COVID-19 webpage.

5.6 Web archiving efforts for COVID-19 related information online

There are several Web archiving initiatives all over the world. Some of the most popular of

these initiatives are the Internet Archive (Brown, 2006; Szydlowski, 2010; Boyle, 2016), Library of Congress Web Archives (Guenther & Myrick, 2006; Chen et al., 2008; Nielsen, 2016), Danish Web archiving project (Sutton, 2004), PANDORA Archive (Boyle, 2016; Srivastava et al., 2016) and the UK Web Archive (UKWA, 2018; Bailey & Thompson, 2006).

Within the last ten years, the Ebola and the Zika virus outbreaks between 2014 and 2016 are examples of widespread viruses that were declared as epidemics. There were efforts to harvest Ebola and Zika virus information online for future research (Mofatt, 2016; Chen, 2018). While the UK Web Archive has an Ebola Outbreak collection (Chen, 2018), the National Library of Medicine (NLM) has also archived web content related to the Ebola outbreak and the Zika virus (Moffatt, 2016). The Library of Congress also has a collection of archived web site for the Ebola outreach known as Ebola Deeply which is accessible through https://www.loc.gov/item/lcwaN0010800/. However, the Library of Congress' Ebola Deeply Web archive collection has a restricted access policy, and it is only accessible by on-site users.

Ever since the outbreak of the COVID-19 and its declaration as a pandemic by WHO, efforts were made by different organizations to harvest web contents and COVID-19 related online information. For example, as part of Global Health Events Web archive, the National Library of Medicines (NLM) with a collaborative effort started archiving the web including social media sites documenting the COVID-19 outbreak (Moffatt, 2020), and this is accessible through https://archive-it.org/collections/4887?fc=websiteGroup%3AEbola+Outbreak+2014. The NLM's Web archiving initiative which began since WHO declared the outbreak as a Public Health Emergency of International Concern (PHEIC) was expected to continue throughout the duration of the pandemic (Moffatt, 2020). The Content Development Group of the International Internet Preservation Consortium (IIPC) in collaboration with Archive-It were also involved in a Web archive project with the aim to preserve web content related to the novel COVID-19 outbreak (IIPC, 2020). The Web archive which is accessible through https://archive-it.org/collections/13529 focuses on high priority subtopics such as coronavirus origins; information about the spread of infection; regional or local containment efforts; medical and scientific aspects; social aspects; economic aspects; and political aspects.

Other websites such as WHO and Worldometer COVID-19 webpages were popular for the dissemination of COVID-19 related data and information that were harvested by the Internet Archive and accessible through the Wayback Machine.

6. Assessment of Internet Archive's Wayback Machine

The Internet Archive is a non-profit organization founded by Brewster Kahle in 1996 with the purpose of building a digital library that offers permanent access to historical collections existing in digital form (Brown, 2006). The Internet Archive which currently contains a large amount of digital information such as documents, scholarly journals, TV programs, audio, and e-books (Boyle, 2016) is currently the largest Web archive in the world (Brown, 2006). The Internet Archive's Wayback Machine accessible through https://archive.org/web/ is an important tool of the Internet Archive as the access point for its archived contents. According to Brown (2006: 9-10), the Wayback

Machine is "an innovative time-based index and interface which not only provides access to individual website snapshots but also allows them to be browsed through within historical context." Therefore, the assessment of some COVID-19 related websites mentioned earlier is outlined below.

6.1 Websites captured by Internet Archive's Wayback Machine

Table 1 shows the presence of the COVID-19 websites or webpages analyzed in this study on the Wayback Machine. The data collected through web content analysis of the individual web pages in the Wayback Machine shows that out of the five websites analyzed, they are all being archived by the Internet Archive and made available on the Wayback Machine. Out of these five websites, two (South African Department of Health and Nigeria's NCDC) appear to be partially archived. This shows that three of five (WHO, Worldometers, and CDC) representing 60% of the websites assessed in the study are fully archived by the Internet Archive and made accessible through the Wayback Machine. This data also shows that the archived webpages are those which were currently considered as the most reliable sources of COVID-19 related statistics and information. The details of the archived websites are further broken down below.

Table 1. Websites' presence on the Wayback Machine

Name/organization	Website	Coverage	Available on Wayback Machine (Yes/No)
World Health Organization (WHO)	https://www.who.int/emergencies/diseases/novel-coronavirus-2019	Worldwide	Yes
Worldometers	https://www.worldometers.info/coronavirus/	Worldwide	Yes
Centers for Disease Control and Prevention (CDC)	https://www.cdc.gov/coronavirus/2019-ncov/index.html	USA	Yes
South African Department of Health	https://sacoronavirus.co.za/	South Africa	Yes
Nigeria Centre for Disease Control (NCDC)	https://covid19.ncdc.gov.ng/	Nigeria	Yes

Source: Author's own work

6.2 Analysis of the Websites' harvest details on the Wayback Machine

This section is a breakdown of the details of the archived websites as presented in Table 1. From the Table, all the five websites included in the study have been harvested and archived by the Internet Archive and accessible on the Wayback Machine. Table 2 shows that the crawling of the harvested webpages by the Internet Archive started between 20th January 2020 and 27th March 2020. The earliest website to be harvested is the Centers for Disease Control and Prevention (CDC) COVID-19 webpage followed by WHO and Worldometer which were first crawled on 20th, 24th and 29th January 2020 respectively. This shows that information contained on these websites as far back as January 2020 can be retrieved on the Wayback Machine. The other websites were

harvested in March 2020 and harvesting on the Africa related COVID-19 webpages all started in March 2020. While the NCDC webpage was first crawled on 15th March 2020, the South African Department of Health COVID-19 portal was first crawled on 15th March 2020. The index COVID-19 cases in Africa were reported between late February and mid-March 2020, so it is obvious that these COVID-19 webpages or portals started publishing reports about the COVID-19 pandemic around this period which explains the late harvesting of the webpages by the Internet Archive in comparison to other webpages. Also, the harvest period of the COVID-19 webpages presented in Table 2 revealed that most of the webpages were harvested between the date of their first crawl and the 17th of April 2020 when the data was analyzed.

Furthermore, Table 2 shows the frequency of crawl of the webpages assessed. It is important to note that the frequency of crawled data presented refers to how frequently the webpages were crawled by the Internet Archive and not the number of times the websites were updated. Some of the webpages also crawled multiple times a day. Worldometer's COVID-19 webpage is the webpage that is harvested most frequently with 5019 crawls between 29th March 2020 and 17th April 2020. Within this period the webpage was crawled almost every day except 9th February 2020. The CDC page was crawled 2,058 times between 20th March 2020 and 17th April 2020 with multiple crawls almost every day except 22nd and 23rd January 2020. Similarly, WHO webpage crawled 1,501 times between 24th January and 17th April 2020 was crawled multiple times daily except 4th, 7th, and 10th February 2020. The South African Department of Health COVID-19 portal was also crawled multiple times daily between its harvest period for a total of 202 times except 16th March 2020.

The data analyzed revealed that NCDC COVID-19 webpages were the least crawled. Data on the Wayback Machine reveals that the webpages are not harvested regularly. Between 27th March and 17th April, the page crawled only seven times within six inconsecutive harvest days. It is therefore evident through the data that some of the webpage's crawl more frequently than others. While some crawled multiple times daily, some crawled almost daily apart from 1-2 days while others like the NCDC webpages crawled less frequently. This reveals that some of the webpages are more complete than others in the Web archive and would be most likely more useful for future research.

Table 2. Webpages harvest details on Wayback Machine

Name/organization	Date of first crawl	Harvest period	Frequency of crawl
World Health Organization (WHO)	24.01.2020	24.01.2020 - 17.04.2020	1,501
Worldometers	29.01.2020	29.01.2020 - 17.04.2020	5,019
Centers for Disease Control and Prevention (CDC)	20.01.2020	20.01.2020 - 17.04.2020	2,058
South African Department of Health	15.03.2020	15.03.2020 - 17.04.2020	202
Nigeria Centre for Disease Control (NCDC)	12.03.2020	12.03.2020 - 17.04.2020	59

Source: Author's own work

7. Discussion

The findings of the study revealed that web archiving is a feasible and effective technique for preserving ephemeral online data related to COVID-19 and other digital history. The analysis of the contents, such as COVID-19 data on the Internet Archive's Wayback Machine, showed that the extent of the capture of COVID-19 related websites at the peak of the pandemic was generally high. The ease of access and dissemination of the content was also satisfactory, with the archived digital content being easily accessible through web archives. However, it was discovered that the capture of online content on social media sites like Twitter is generally low. Although the Wayback Machine seems to have captured the URLs, the actual contents of these websites are not accessible.

The findings on archiving frequencies highlight critical implications for future research usability of the archived web. Websites with higher crawl frequencies, such as those of WHO, CDC, and Worldometer, are likely to provide more comprehensive and consistent data for longitudinal studies. On the other hand, websites with lower crawl frequencies, such as those from the NCDC and the South African Department of Health, may result in incomplete datasets, potentially limiting their usefulness for detailed research or comparative analyses. These disparities emphasize the need for targeted efforts to enhance the frequency and consistency of web crawls for underrepresented regions to ensure equitable access to archived data. In addition, the challenges associated with archiving social media content warrant deeper exploration, given the pivotal role social media played in disseminating COVID-19 information and shaping public discourse. Unlike static web pages, social media platforms feature dynamic, user-generated content that is often ephemeral and more difficult to capture comprehensively. For example, Twitter threads or Facebook posts containing critical updates, personal testimonies, or real-time discussions may disappear or become inaccessible due to privacy settings, account deletions, or platform policies. To address these challenges, web archiving initiatives must consider developing specialized tools and frameworks tailored to the unique nature of social media content. Such tools should prioritize capturing metadata, preserving multimedia elements, and maintaining context to ensure the usability of archived social media data for future research.

The study also identified some challenges associated with the web archiving of ephemeral online data, including the technical complexity of the web archiving process, the need for continuous monitoring and maintenance of the archived material, and the legal and ethical issues associated with the preservation of online data. In addition, there seems to be a lack of awareness of the potential of web archiving as a tool for preserving content online, especially in Africa. Archivists and other stakeholders involved in digital preservation lack the required knowledge and training on the use of web archiving tools and techniques to ensure the effective preservation of digital heritage.

Analysis of several digital preservation policies revealed that web archiving is not widely recognized or incorporated into the policies of several public institutions with an online presence. These findings call for an increased effort to address technical, legal, and regional challenges, ensuring that valuable data is preserved not only for historical records but also for academic and public health research. Overall, the findings of this study highlight the importance of web archiving as a valuable technique for preserving ephemeral online data related to COVID-19 and other potential digital history.

8. Conclusion

In conclusion, this study explored the use of web archiving techniques for preserving ephemeral online data related to digital history like COVID-19 and political events. The study showed that web archiving is a feasible and effective technique for preserving ephemeral data, information, and records. The study highlighted prominent sources like the World Health Organization (WHO), Worldometer, and national health departments provided critical data that were extensively archived. However, it also exposed the relatively lower frequency of archiving for websites focused on Africa, emphasizing a regional disparity in digital preservation efforts. Furthermore, technical complexities and legal and ethical considerations present significant hurdles in the archiving process, necessitating comprehensive strategies to address these issues. Additionally, the findings underscore the importance of web archiving in preserving not just public health information but also the broader socio-economic and political narratives shaped by the pandemic. This ensures that future researchers have a holistic view of the impact and response to such global crises. The study also calls attention to the potential for web archiving to support public accountability and transparency by preserving official communications and public reactions in real-time.

Based on the study's findings, several key recommendations are proposed to strengthen web archiving for preserving digital history. Increasing awareness and providing targeted training on web archiving tools is essential to equip archivists and stakeholders with the necessary skills. Continuous monitoring and maintenance of archived content should be ensured through clear policies to support long-term preservation and accessibility. Developing robust legal and ethical frameworks, with contributions from archivists, policymakers, legal experts, and affected communities, is vital to balance preservation efforts with individual and collective rights, especially for marginalized groups. Also, greater collaboration among stakeholders and partnerships with organizations like the International Internet Preservation Consortium (IIPC) and the UK Web Archive can help address regional disparities, particularly in Africa. Governments and academic institutions should also establish regional centers to build local capacity and advocate for sustainable funding. Furthermore, a comprehensive preservation framework should be developed to guide consistent archiving practices, and specialized tools must be created to improve the archiving of social media platforms like Twitter. Finally, targeted initiatives should address regional disparities by increasing the archiving of underrepresented websites. Implementing these measures will enhance the preservation of valuable digital content for future research and historical reference.

Declaration of Interest Statement

The author declares no conflicts of interest, financial assistance or funding with regards to the publication of this article.

References

Acker, A., & Chaiet, M. (2020). The weaponization of web archives: Data craft and COVID-19 publics. *Harvard Kennedy School-Shorenstein Center on Media, Politics, and Public Policy.* https://repositories.lib.utexas.edu/server/api/core/bitstreams/fe7f8a0d-d91f-4fc2-857c-fe005ca67473/content

Antracoli, A., Duckworth, S., Silva, J., & Yarmey, K. (2014). Capture all the URLs: first steps in Web archiving. *Pennsylvania Libraries: Research and Practice, 2*(2), 155-170.

Bailey, S., & Thompson, D. (2006). Building the UK's First Public Web Archive. *D-Lib Magazine, 12*(1). http://www.dlib.org/dlib/january06/thompson/01thompson.html

Belovari, S. (2017). Historians and Web Archives. *Archivaria, 83*, 59-79.

Ben-David, A., & Huurdeman, H. (2014). Web Archive Search as Research: Methodological and Theoretical Implications. *Alexandria, 25*(1-2), 93-111. https://doi.org/10.7227/ALX.0022

Boyle, A. (2016). *Archiving the Internet, what does it mean in practice?* (Master thesis: Leiden University). https://openaccess.leidenuniv.nl/bitstream/handle/1887/40119/Anya%20Boyle%20Thesis.pdf?sequence=1

Brown, A. (2006). *Archiving Websites: A Practical Guide for Information Management Professionals.* London: Facet Publishing.

Brügger, N. (2005). *Archiving websites: General considerations and strategies*. Centre for Internet Research. http://cfi.au.dk/fileadmin/www.cfi.au.dk/publikationer/archiving_underside/archiving.pdf

Brügger, N. (2016). Introduction: The Web's first 25 years. *New Media & Society, 18*(7), 1059-1065. https://doi.org/10.1177/1461444816643787

Chen K. H., Chen, Y. L., & Ting, P. F. (2008). Developing National Taiwan University Web Archiving System. *Proceedings from IWWA '08: 8th International Workshop for Web Archiving*, Denmark (pp. 1-8).

Chen, E. (2018, February). The UK Web Archive Ebola Outbreak collection. *Archives and Manuscript at the Bodleian Library: a Bodleian Library blog.* https://blogs.bodleian.ox.ac.uk/archivesandmanuscripts/2018/02/09/the-uk-web-archive-ebola-outbreak-collection/

Costea, M. D. (2018). *Report on the scholarly use of web archives*. NetLab.

Davis, C. (2014). Archiving the Web: A Case Study from the University of Victoria. *Code4Lib Journal,* (26).

Dougherty, M., & Meyer, E. T. (2014). Community, tools, and practices in web archiving: The state-of-the-art in relation to social science and humanities research needs. *Journal of the Association for Information Science and Technology, 65*(11), 2195-2209.

Ducharme, J. (2020). World Health Organization Declares COVID-19 a 'Pandemic.' Here's What That Means. *Time. 2020*. https://time.com/5791661/who-coronavirus-pandemic-declaration/

Duncan, S. (2015). Preserving born-digital catalogues raisonnés: Web archiving at the New York Art Resources Consortium (NYARC). *Art Libraries Journal, 40*(2), 50-55.

Duncan, S., & Blumenthal, K. R. (2016). A Collaborative Model for Web Archiving Ephemeral Art Resources at the New York Art Resources Consortium (NYARC). *Art Libraries Journal, 41*(2), 116-126.

Ferguson, C., Merga, M., & Winn, S. (2021). Communications in the time of a pandemic: the readability of documents for public consumption. *Australian and New Zealand Journal of Public Health, 45*(2), 116-121.

Gomes, D., & Costa, M. (2014). The importance of web archives for humanities. *International Journal of Humanities and Arts Computing, 8*(1), 106-123.

Guenther, R., & Myrick, L. (2007). Archiving web sites for preservation and access: MODS, METS and MINERVA. *Journal of Archival Organization, 4*(1-2), 141-166.

Habibzadeh, P. (2013). Decay of references to Web sites in articles published in general medical journals: mainstream vs small journals. *Applied clinical informatics, 4*(04), 455-464.

Hendry, R., & Stock, G. (2014). Forget Me Net, Not: Inside the Struggle to Preserve the World's Data. *Newsweek Global, 168*(2), 1-6.

International Internet Preservation Consortium. (2020). *Novel Coronavirus outbreak: help us collect websites.*
https://netpreserveblog.wordpress.com/2020/02/13/cdg-collection-novel-coronavirus/

Jaumotte, F., Oikonomou, M., Pizzinelli, C., & Tavares, M. M. (2023, March). How Pandemic Accelerated Digital Transformation in Advanced Economies. *IMF Blog.*
https://www.imf.org/en/Blogs/Articles/2023/03/21/how-pandemic-accelerated-digital-transformation-in-advanced-economies

Kumar, B. T., & Prithviraj, K. R. (2015). Bringing life to dead: Role of Wayback Machine in retrieving vanished URLs. *Journal of Information Science, 4*(1), 71-81.
DOI: 10.1177/0165551514552752.

Lomborg, S. (2012). Researching Communicative Practice: Web Archiving in Qualitative Social Media Research. *Journal of Technology in Human Services, 30*(3-4), 219-231.
https://doi.org/10.1080/15228835.2012.744719

Mcclain, C., Vogels, E. A., Perrin, A., Sechopoulos, S., & Rainie, L. (2021, September). The Internet and the Pandemic. *Pew Research Center.*
https://www.pewresearch.org/internet/2021/09/01/the-internet-and-the-pandemic/

Moffatt, C. (2006, March). Future historical collections: archiving the 2014 Ebola outbreak. *Circulating Now: From the Historical Collections of the National Library of Medicine.*
https://circulatingnow.nlm.nih.gov/2016/03/10/future-historical-collections-archiving-the-2014-ebola-outbreak-2/

Moffatt, C. (2020, March). Archiving Web content on the Coronavirus disease (COVID-19). *Circulating Now: From the Historical Collections of the National Library of Medicine.*
https://circulatingnow.nlm.nih.gov/2020/03/26/archiving-web-content-on-the-coronavirus-disease-covid-19/

Nielsen, J. (2016). Using web archives in research-an introduction.
http://www.netlab.dk/wp-content/uploads/2016/10/Nielsen_Using_Web_Archives_in_Research.pdf

Schneider, S. M., & Foot, K. A. (2004). The Web as an Object of Study. *New Media & Society, 6*(1), 114-122. https://doi.org/10.1177/1461444804039912

Sheldon, Z. (2019). The archived web: doing history in the digital age: by Niels Brügger, Cambridge, MA, MIT Press, 2018, 185 pp., $30, Hardcover. ISBN 978-0262039024. *Information, Communication & Society, 23*(2), 309-311. https://doi.org/10.1080/1369118X.2019.1682636

Sindhwani, R., Kumar, G. P., & Saddikuti, V. (2022). Effect of COVID-19 pandemic on social factors. In COVID-19 and the Sustainable Development Goals (pp. 259-284). Elsevier.

Slania, H. (2013). Online Art Ephemera: Web Archiving at the National Museum of Women in the Arts. *Art Documentation: Journal of the Art Libraries Society of North America, 32*(1), 112-126.

Sohrabi, C., Mathew, G., Franchi, T., Kerwan, A., Griffin, M., Soleil C Del Mundo, J., Ali SA., Agha, M., & Agha, R. (2021). Impact of the coronavirus (COVID-19) pandemic on scientific research and implications for clinical academic training-A review. *International Journal of Surgery, 86,* 57-63.

Speaker, S. L., & Moffatt, C. (2020). The National Library of Medicine Global Health Events web archive, coronavirus disease (COVID-19) pandemic collecting. *Journal of the Medical Library Association: JMLA, 108*(4), 656.

Srivastava, M., Singh, D. S., & Abbas, D. S. (2016). Web archiving: past present and future of evolving multimedia legacy. *International Advanced Research Journal in Science, Engineering and Technology, 3*(3).

Stirling, P., Chevallier, P., & Illien, G. (2012). Web archives for researchers: Representations, expectations and potential uses. *D-Lib, 18*(3/4). https://www.dlib.org/dlib/march12/stirling/03stirling.html.

Sutton, M. (2004). Preserving our Internet history-Websites on archiving via the Internet. *SA Journal of Information Management, 6*(4).

Szydlowski, N. (2010). Archiving the web: It's going to have to be a group effort. *The Serials Librarian, 59*(1), 35-39.

Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., & Mechant, P. (2019). Web archives as a data resource for digital scholars. *International Journal of Digital Humanities, 1*, 85-111.

Weigle, M. C. (2023). The Use of Web Archives in Disinformation Research. arXiv preprint arXiv:2306.10004.

Wikipedia (2020). *Centers for Disease Control and Prevention.* https://en.wikipedia.org/wiki/Centers_for_Disease_Control_and_Prevention#cite_note-3

World Health Organization (WHO). (2020a, September). Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation: Joint statement by WHO, UN, UNICEF, UNDP, UNESCO, UNAIDS, ITU, UN Global Pulse, and IFRC. https://www.who.int/news/item/23-09-2020-managing-the-covid-19-infodemic-promoting-healthy-behaviours-and-mitigating-the-harm-from-misinformation-and-disinformation

World Health Organization. (2020b). Coronavirus.

https://www.who.int/health-topics/coronavirus#tab=tab_1

Worldometer. (2020a). *COVID-19 Coronavirus pandemic.* https://www.worldometers.info/coronavirus/

Worldometer. (2020b). *About*. https://www.worldometers.info/about/

Wu, F., Yang, T., Zhang, C., Yu, Y., & Xu, D. (2024). Internet-Based Media as Information Sources in Risk Communication: Comparing Three Media Sources During COVID-19 Pandemic. *Journalism Practice*, 1-24.

[About the author]

Tolulope Balogun is a Postdoctoral researcher at the University of South Africa with research interest in digitization, digital preservation, Artificial Intelligence, records and archives.