Web Archives Preserve Our Digital Collective Memory

Daniela Major and Daniel Gomes

Abstract This chapter discusses the importance of web archiving, briefly presents its history from the beginning with the Internet Archive in 1996 and exposes the challenges with archiving certain types of online data.

Contemporary generations have the responsibility of preserving current information for future ones, in the same way that previous generations have contributed their knowledge legacy to us. But is it possible to provide a meaningful description of our current times for future generations that ignores web heritage? Most web authors do not have either the resources nor the awareness to preserve their digital works. The mission of web archives is to provide web authors with the right to be remembered. In 1996, Brewster Kahle initiated the Internet Archive, which aimed to preserve the information on the pages published through a new technology named the World Wide Web. Fortunately, many other initiatives with the same objective followed around the world. Web archives collect and preserve information published online, so that it can be consulted after it is no longer available on its original websites. They contribute to solving numerous everyday problems, for example web users coming across broken links; journalists looking for past information to reference in articles; software engineers searching for technical manuals to fix legacy systems; webmasters recovering past versions of their sites' pages; historians searching for web documents describing past events; or researchers looking for related work published open-access. In general, all the use cases existing for the current Web are applicable to the past Web when we adopt a historical perspective, because all current information used today will one day be part of the past. Use cases such as longitudinal studies of sociological events like political trends or the

D. Major (✉)
School of Advanced Study, University of London, London, UK
e-mail: daniela.major@sas.ac.uk

D. Gomes
Fundação para a Ciência e a Tecnologia, Lisbon, Portugal
e-mail: daniel.gomes@fccn.pt

D. Gomes et al. (eds.), *The Past Web*,
https://doi.org/10.1007/978-3-030-63291-5_2

evolution of media communication come to mind. However, other use cases may not be so obvious. For instance, web archives can contribute to the prevention of terrorist attacks. Terrorists' communications have been shared using the public Web, through special code keywords and nicknames, so that anyone interested in their actions could easily find information. There are people interested in terrorist organisations who are not yet active—they are called *hobbyists*. However, terrorist organisations incrementally incite them to perform more incisive actions. The probability of a hobbyist becoming violent increases with time. A hobbyist with a given nickname active on the Web for several years is potentially more dangerous than a newcomer. If the hobbyist's publications present indications of different geographical locations across time, this could also be an alert indicator. Web platforms evolve but terrorists leave online traces among them over time which can be tracked to measure their level of engagement and dangerousness. This historical analysis can only be performed by security agencies through web archives. Focused web archives for national security or counterterrorism that collect information from the dark Web are important tools to ensure security. Maybe they already exist; or maybe this awareness has not reached security agencies and governments yet.

1 Word Wide Web Archiving

Web archives are crucial for preserving information published on the Web and making it reusable in the future. Since 1996, several web archiving initiatives have been created around the world. In 2020, there were at least 94 web archiving initiatives (Wikipedia 2020). Web archives worldwide host at least 981 billion web pages, which are stored on 29 petabytes of disk space. This volume of data will continue to grow as they acquire information incrementally to build their historical collections documenting the past Web. In 2006, Julien Masanès edited the book *Web Archiving*, which was a major milestone in establishing the preservation of the Web as a scientific and professional field (Masanès 2006). Since 2006, a lot has changed in the world and in the area of web archiving. In the early 2000s, web archiving was an effort pursued mainly by librarians and a few computer scientists. The challenges focused mainly on how to collect information from the Web and store it. During the 2010s, all kinds of professionals became involved in web archiving activities. The web archiving community was widespread and engaged social scientists, data scientists, IT professionals, librarians and historians. In this new decade, collecting and storing web data remains an essential and demanding task, but the main challenges today concern how to provide efficient access so that the stored web data can be effectively preserved. There is no preservation without proper access methods. The professionals that work at cultural heritage organisations do not hold themselves the knowledge contained in the artefacts they preserve (e.g. books, paintings, songs or movies). Their mission has been to grant access to that knowledge so that it can be explored by suitable experts, students or citizens. The breakthrough of the digital era was to enable large-scale and low-cost access to knowledge through the widespread usage of the Web as it became a ubiquitous communication medium.

For the first time, “universal access to all knowledge” became an achievable goal for humankind. Web archives are gatekeepers that safeguard the collective memory of societies in the digital era. However, without providing open and efficient online tools for engaging with web archives, the precious historical artefacts that document the digital era are as inaccessible as books stored in distant monasteries during the Middle Ages. Knowledge exists but nobody can access it. It was with these principles in mind that organisations such as the International Internet Preservation Consortium (IIPC) and the Digital Preservation Coalition (DPC) were created. The IIPC was formally set up in 2003, sponsored by the Bibliothèque nationale de France, and now includes libraries, museums and cultural institutions in an effort to “acquire, preserve and make accessible knowledge and information from the Internet for future generations everywhere” (IIPC 2021, *about* . . .). The DPC is a non-profit organisation which allows “members to deliver resilient long-term access to digital content and services, helping them to derive enduring value from digital assets” (Digital Preservation Coalition 2021a, b) and counts as members the British Library, Cambridge University Library, CERN, as well as universities or private companies such as banks.

2 A Brief History of Web Archiving

The first webpage was published on the Internet in 1990, although access to the Internet was only made available to the general public in the following year (World Wide Web Foundation 2021). Five years later, in 1996, the Internet Archive (IA) was founded by Brewster Kahle and Bruce Gilliat. Months earlier, Kahle and Gilliat had also founded Alexa Internet, a for-profit search engine that catalogued the Web and provided data about sites visited to a browser toolbar. Alexa then went into partnership with Microsoft and Netscape, so that it could have a presence in most desktop computers (Masanès 2006). Alexa contributed to the Internet Archive by taking snapshots of webpages and donating them to the IA, which then created collections (Morh et al. 2004). While initially the material archived was not publicly available, as the Internet grew in size and usage, it became clear that the next step was to make the content easily accessible. Consequently, the Wayback Machine service, which the IA had been using since 1996, was made public in 2001. To use the Wayback Machine, users simply insert the URL of a page into a search bar, and then, if the page has been archived, they are able to peruse all the old versions that have been captured along time. However, the crawler utilised to harvest the Web and provide the archived pages, the aforementioned Alexa, did not belong to the IA. It had been bought by Amazon in 1999. The IA needed to develop its own software and techniques to gather web data. Thus, a joint team of computer scientists at the IA and the Nordic national libraries developed the Heritrix software, “an open source, extensible web crawler” (IIPC 2017) whose purpose was to archive “websites and to support multiple different use cases including focused and broad crawling” (Lee et al. 2011). Web crawling is the process by which a software component automatically gathers information from the Web (Manning 2008).

Heritrix made it possible for the IA to enlarge the scope of the material archived, and the software could also be moulded to meet the needs of users. Alexa was mainly dedicated to indexing text, which is and continues to be the most common content on the Internet; it was only with Heritrix that the IA began systematically to archive the images contained in the websites they crawled and archived.

Web archiving may be defined as the process of “gathering up data that has been recorded on the World Wide Web, storing it, ensuring the data is preserved in an archive, and making the collected data available for future research” (Niu 2012). When Yahoo! announced the closure of Geocities in 2009, the IA and groups of independent web archivists such as the Archive Team (Archive Team 2021) worked together to save the site’s content. This effort ensured that as early as 2010 a huge file of 652 GB was made available online, encompassing data from the now-defunct Geocities platform (Gilbertson 2010). Much of this content can today be accessed using the Wayback Machine, and it documents an important part of Internet history (Internet Archive 2009).

The concern to archive webpages is relatively new, but awareness is spreading. In 2015, a survey of academics from the social sciences and humanities in New Zealand universities found that 44% of correspondents had “at some point used some of the following international sources of web archives” (the list included the Internet Archive, the Library of Congress Web Archive and the UK Web Archive). Furthermore, a large majority of the correspondents stated they believed that “it is important for New Zealand websites and blogs to be archived”, and half of those surveyed answered that the New Zealand Web Archive would be “important for their current research within the next 5 years” (Riley and Crookston 2015).

As of 2016, the indexed Web, that is, those websites indexed by major search engines such as Google and Bing, contained 4.75 billion websites. It is estimated that since 2012, the Web has doubled in size every year. It is further estimated that over 3 billion people use the Web worldwide (Fischer 2020). While it is not possible to archive all of the Web, web archives do manage to gather impressive amounts of information. According to the IA’s website, the Web Archive Collection amounts to 330 billion webpages, 20 million books and texts, 4.5 million audio recordings, 4 million videos, 3 million images and 200,000 software programmes (Internet Archive 2021).

3 Archiving Online Social and Multimedia

The sheer variety of the material collected by web archives demonstrates that the Web is a medium that includes multiple content formats—text, images, videos, audio—but these have not been collected equally. A breakthrough came when it became possible to archive moving images, namely videos. Here the Institut national de l’audiovisuel (INA) has assumed a very important role. Established in 1975, INA had a mission to preserve the archives of existing radio and television channels. In 1992, together with the Bibliothèque nationale de France and the Centre national du cinéma et de l'image animée, INA became a legal repository of French audiovisual patrimony. Finally, in 2006, INA also became responsible for archiving a part of the Web. Its work focused on French websites, especially those that concerned web television and radio, radio and television websites, and institutional websites on the subject of communication and audiovisual media. Since 2008, INA has archived social media accounts whose prime medium of communication is video (INA 2021). INA is a very important initiative both for the scope of its content and for how accessible it is to ordinary users. However, it is not the only institution that is concerned with the archiving of moving images from the Web. Coca-Cola, a private corporation, has a web archive which, apart from containing "all Coca-Cola sites and other selected sites associated with Coca Cola", includes "social media and video" (Digital Preservation Coalition 2021a, b).

The development of web archives necessarily has to keep pace with the development of the Web. Many people use the Web, among other things, for social media. In 2018, around 2.62 billion people used social media platforms. In 2020, a global average of 49% of Internet users had some kind of social media presence (Clement 2020). Ever since the 2016 American presidential election, debate has been growing about the influence of social media on current events. It has become clear that social media produces large volumes of relevant information that concerns the general public. But who is archiving this information? The National Archives of the UK has an initiative to archive tweets from official government Twitter accounts, such as those from various governmental departments (National Archives 2021). INA also has incorporated tweets into its web archive, although they are not yet widely available for consultation as it is only possible to access the database from specific locations in France. Although recently abandoned (Bruns 2018), for several years the Library of Congress undertook a very ambitious project to archive all public Twitter data, with help from Twitter itself, which gave the Library of Congress access to its own archive (Stone 2010). With regard to Facebook pages, the IA and other web archives tend to keep a record of those made public. For instance, Donald Trump's Facebook page has been captured nearly every day since 2016 by the IA. Some additional social media is archived by Internet Archive such as the Instagram page of the photographer responsible for Humans of New York[1] or the Tumblr blog of the fiction writer Neil Gaiman.[2]

4 Cloud Web-Archiving and Crowdsourcing

An important, and relatively recent, development in web archiving is that it can be undertaken through cooperation between Internet users and institutions. Archive-It, for instance, is a collaboration between the IA and institutions all over the world,

[1] https://www.instagram.com/humansofny/?hl=en

[2] https://neil-gaiman.tumblr.com/

from state and university libraries, national archives and museums to governmental bodies. Archive-It helps these institutions to archive content considered important in exchange for a fee and then creates thematic collections which can be made accessible to everyone. For instance, the Human Rights Documentation Initiative, a collaboration between Archive-It and the University of Texas at Austin, features “fragile websites containing human rights documentation and related content from human rights organizations and advocates across the globe” which can be browsed by topic, geographic area, language or keyword search (Archive-it 2021). Furthermore, it is possible for an independent user to contribute to the archiving of webpages. When a user searches for a URL in the IA and the page cannot be found, the service automatically suggests that the user can add the missing page to the archive, which can be done by simply clicking on a link. In the case of Arquivo.pt., the Portuguese web-archive, users can suggest a website or a list of websites that they consider worth archiving by completing an online form. In 2016, Rhizome, an arts organisation that supports the preservation of new media art, launched Webrecorder, a free tool that allows users to create collections of recorded webpages, which they can then keep for themselves or share with others (Espenschied 2016). Webrecorder saves more than just a screenshot of the page; it saves all the associated links and images that are presented on a given page.

5 Preserving the Web: New Media, Same Problem

The issue of how to preserve new media formats is not novel. It is estimated that about 50% of the movies made before 1950 are irremediably lost; 90% of movies made in the USA before 1929 have completely disappeared (Film Foundation 2021); and 75% of American silent movies have been lost (Ohlheiser 2013). This means that a significant portion of film history has disappeared both because there was insufficient space to archive movies in the studios and because the material used to make the film—nitrate—deteriorated quickly without proper preservation. The cost of such preservation was often deemed too expensive. There were other causes for film destruction. In Europe, the First and Second World Wars contributed to the destruction of film. For instance, Bezhin Meadow, an unfinished movie by Sergei Eisenstein, was destroyed during the bombing of Moscow in the Second World War (Kenez 2001).

However, during the 1930s institutional attempts began to be made to try to preserve film. “Numerous departments for audiovisual media were set up within existing, well-established institutions” such as the Museum of Modern Art Film Library, the British Film Institute, the Cinemathèque Française and the Reichsfilmarchiv. In 1938, all of these came together to found the International Federation of Film Archives (FIAF). Its members were “committed to the rescue, collection, preservation, screening, and promotion of films, which are valued both as works of art and culture and as historical documents” through “the creation of moving image archives in countries which lack them” and the establishment of “a code of ethics for film preservation and practical standards for all areas of film archive work”. As of May 2018, FIAF “comprises more than 166 institutions in 75 countries” (FIAF 2021).

Samewise to the archiving of the Web, archiving film was not considered of historical importance for decades. Awareness was slowly raised, and now it is universally believed that film ought to be preserved, not only to be viewed after its release but because of its cultural and social significance. The same can be said about the archiving of the Web, with the added knowledge that the Web is not only a medium of culture but has become an integral part of all aspects of our daily lives.

6 A Future of Open Challenges

Web archiving must keep up with the development of the Web. Video, for example, is currently a frequently used medium of content creation, but the majority of web archiving projects are still “unable to properly download and archive a YouTube video and preserve it for posterity as part of their core crawling activity” (Leetaru 2017). Much of the content that is produced on social media is locked away not only because many accounts are private but because in some cases logins are needed to access the websites. There is no easy solution to this problem, as privacy must be a concern for web archivists. There is also much content that is produced by online services such as Google Maps that are not currently being archived, as well as pages that use JavaScript, which some web archives are still unable to crawl.

In the early days of the Web, it became clear that acquiring and storing online information before it quickly vanished was a challenge. But it was a rather simple challenge in comparison with ensuring the accessibility of the stored web data over time. It is a naive common belief that everything is online, leading to the misleading conclusion that, if it is not online, it does not exist; or even more concerning, that it never happened. The tremendously fast pace at which the Internet has penetrated societies, without the development of proper digital preservation services, may actually have created an online amnesia about recent events. On the other hand, for the first time in the history of humankind, the technology exists to make information published in the past as accessible as the information about the present.

References

Archive Team (2021) Geocities. Archive Team. Available at: https://www.archiveteam.org/index.php?title=GeoCities. Accessed on 29 January 2021

Archive-it (2021) Human rights documentation initiative. Archive-it. Available at: https://archive-it.org/collections/1475. Accessed on 29 January 2021

Bruns A (2018) The library of congress twitter archive: a failure of historic proportions. Medium. Available at: https://medium.com/dmrc-at-large/the-library-of-congress-twitter-archive-a-failure-of-historic-proportions-6dc1c3bc9e2c. Accessed on 2 January 2021

Clement (2020) Number of social network users worldwide from 2010 to 2023. Statista. Available at: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/. Accessed on 1 April 2020

Digital Preservation Coalition (2021a) About. Digital Preservation Coalition. Available at: https://www.dpconline.org/about. Accessed on 29 January 2021

Digital Preservation Coalition (2021b) Web-archiving. Digital Preservation Coalition. Available at: https://www.dpconline.org/handbook/content-specific-preservation/web-archiving. Accessed on 29 January 2021

Espenschied D (2016) Rhizome releases first public version of webrecorder. Rhizome. Available at: https://rhizome.org/editorial/2016/aug/09/rhizome-releases-first-public-version-of-webrecorder/. Accessed on 9 August 2020

FIAF-International Federation of Film Archives (2021) FIAF's Mission. FIAF. Availble at: https://www.fiafnet.org/pages/Community/Mission-FIAF.html. Accessed on 29 January 2021

Film Foundation (2021) About us-film preservation. The Film Foundation. Available at: https://web.archive.org/web/20130312021638/http://www.film-foundation.org/common/11004/aboutAboutUs.cfm?clientID=11004&sid=2&ssid=5. Accessed on 29 January 2021

Fischer T (2020) How big is the web? Lifewire. Available at: https://www.lifewire.com/how-big-is-the-web-4065573. Accessed on 15 January 2020

Gilbertson S (2010) Geocities lives on a massive torrent download. Wired. Available at: https://www.wired.com/2010/11/geocities-lives-on-as-massive-torrent-download/. Accessed on 1 November 2020

IIPC-International Internet Preservation Consortium (2017) Tools and Software. Git Hub. Available at: https://github.com/iipc/iipc.github.io/wiki/Tools-and-Software. Accessed on 29 January 2021

IIPC-International Internet Preservation Consortium (2021) About IIPC. Net-Preserve. Available at: http://netpreserve.org/about-us/. Accessed on 29 January 2021

INA (2021) Dépôt légal radio, télé et web. Institut national de l'audiovisuel. Available at: https://institut.ina.fr/institut/statut-missions/depot-legal-radio-tele-et-web. Accessed on 29 January 2021

Internet Archive (2009) Geocities special collection 2009. Internet Archive. Available at: https://archive.org/web/geocities.php. Accessed on 29 January 2021

Internet Archive (2021) About the Internet Archive. Internet Archive. Available at: https://archive.org/about/. Accessed on 29 January 2021

Kenez P (2001) A history of Bezhin meadow. In: LaValley AJ, Scherr BP (eds) Eisenstein at 100: a reconsideration. Rutgers University Press, New Jersey

Lee HB, Nazareno F, Jung SH, Cho WS (2011) A vertical search engine for school information based on Heritrix and Lucene. In: Lee G, Howard D, Ślęzak D (eds) Convergence and hybrid information technology. Springer, Berlin

Leetaru K (2017) Are web archives failing the modern web: video, social media, dynamic pages and the mobile web. Forbes. Available at: https://www.forbes.com/sites/kalevleetaru/2017/02/24/are-web-archives-failing-the-modern-web-video-social-media-dynamic-pages-and-the-mobile-web/#53a22d3845b1. Accessed on 24 February 2020

Manning CD, Raghavan P, Schutze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge

Masanès J (2006) Web archiving. Springer, New York

Mohr G, Kimpton M, Stack M, Ranitovic I (2004) Introduction to heritrix, an archival quality web crawler. In: Proceedings of the 4th International Web Archiving Workshop IWAW'04

National Archives (2021) Twitter archives. National archives-UK Government. Available at: https://webarchive.nationalarchives.gov.uk/twitter/. Accessed on 29 January 2021

Niu J (2012) An overview of web archiving. D-Lib Mag 18(3–4). Available at: http://www.dlib.org/dlib/march12/niu/03niu1.html

Ohlheiser A (2013) Most of America's silent films are lost forever. The Atlantic. Available at: https://www.theatlantic.com/culture/archive/2013/12/most-americas-silent-films-are-lost-forever/355775/. Accessed on 4 December 2020

Riley H, Crookston M (2015) Use of the NZ web archive: introduction and context. National Library of New Zealand. Available at: https://natlib.govt.nz/librarians/reports-and-research/use-of-the-nz-web-archive/introduction

Stone B (2010) Tweet preservation. Blog Twitter. Available at: https://blog.twitter.com/official/en_us/a/2010/tweet-preservation.html. Accessed on 14 April 2020

Wikimedia Foundation, Inc (2020) List of web archiving initiatives. https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives, last update on 21 April 2020. Accessed on 28 April 2020

World Wide Web Foundation (2021) History of the Web. Available at: https://webfoundation.org/about/vision/history-of-the-web/. Accessed on 29 January 2021