

Web archiving

Web archiving is the process of collecting, preserving, and providing access to material from the [World Wide Web](#). The aim is to ensure that information is [preserved](#) in an [archival](#) format for research and the public.^[1] The process of platformizing archives, digitizing historical records via interfaces patterned on social media platforms, can reshape collective memory by privileging content that aligns with social-media logic such as popularity, connectivity, and programmability.^[2]

Web archivists typically employ automated [web crawlers](#) to capturing the massive amount of information on the Web. A widely known web archive service is the [Wayback Machine](#), run by the [Internet Archive](#).

The growing portion of human culture created and recorded on the web makes it inevitable that more and more libraries and archives will have to face the challenges of web archiving.^[3] [National libraries](#), [national archives](#), and various consortia of organizations are also involved in archiving Web content to prevent its loss.

Commercial web archiving software and services are also available to organizations that need to archive their own web content for corporate heritage, regulatory, or legal purposes.

History and development

While curation and organization of the web has been prevalent since the mid- to late-1990s, one of the first large-scale web archiving projects was the [Internet Archive](#), a non-profit organization created by [Brewster Kahle](#) in 1996.^[4] The Internet Archive released its own search engine for viewing archived web content, the [Wayback Machine](#), in 2001.^[4] As of 2018, the Internet Archive was home to 40 petabytes of data.^[5] The Internet Archive also developed many of its own tools for collecting and storing its data, including [PetaBox](#) for storing large amounts of data efficiently and safely, and [Heritrix](#), a web crawler developed in conjunction with the Nordic national libraries.^[4] Other projects launched around the same time included a web archiving project by the [National Library of Canada](#), Australia's [Pandora](#), Tasmanian web archives and Sweden's [Kulturarw3](#).^{[6][7]}

From 2001 to 2010, the International Web Archiving Workshop (IWAW) provided a platform to share experiences and exchange ideas.^{[8][9]} The [International Internet Preservation Consortium](#) (IIPC), established in 2003, has facilitated international collaboration in developing standards and open source tools for the creation of web archives.^[10]

The now-defunct [Internet Memory Foundation](#) was founded in 2004 and founded by the [European Commission](#) in order to archive the web in Europe.^[4] This project developed and released many open source tools, such as "rich media capturing, temporal coherence analysis, spam assessment, and terminology evolution detection."^[4] The data from the foundation is now housed by the Internet Archive, but not currently publicly accessible.^[11]

Despite the fact that there is no centralized responsibility for its preservation, web content is rapidly becoming the official record. For example, in 2017, the [United States Department of Justice](#) affirmed that the government treats the President's [tweets](#) as official statements.^[12]

Methods of collection

Web archivists generally archive various types of web content including HTML web pages, style sheets, JavaScript, images, and video. They also archive metadata about the collected resources such as access time, MIME type, and content length. This metadata is useful in establishing authenticity and provenance of the archived collection.

Transactional archiving

Transactional archiving is an event-driven approach, which collects the actual transactions which take place between a web server and a web browser. It is primarily used as a means of preserving evidence of the content which was actually viewed on a particular website, on a given date. This may be particularly important for organizations which need to comply with legal or regulatory requirements for disclosing and retaining information.^[13]

A transactional archiving system typically operates by intercepting every HTTP request to, and response from, the web server, filtering each response to eliminate duplicate content, and permanently storing the responses as bitstreams.

Difficulties and limitations

Crawlers

Web archives which rely on web crawling as their primary means of collecting the Web are influenced by the difficulties of web crawling:

- The robots exclusion protocol may request crawlers not access portions of a website. Some web archivists may ignore the request and crawl those portions anyway.
- Large portions of a website may be hidden in the Deep Web. For example, the results page behind a web form can lie in the Deep Web if crawlers cannot follow a link to the results page.
- Crawler traps (e.g., calendars) may cause a crawler to download an infinite number of pages, so crawlers are usually configured to limit the number of dynamic pages they crawl.
- Most of the archiving tools do not capture the page as it is. It is observed that ad banners and images are often missed while archiving.

However, it is important to note that a native format web archive, i.e., a fully browsable web archive, with working links, media, etc., is only really possible using crawler technology.

The Web is so large that crawling a significant portion of it takes a large number of technical resources. Also, the Web is changing so fast that portions of a website may suffer modifications before a crawler has even finished crawling it.

General limitations

Some web servers are configured to return different pages to web archiver requests than they would in response to regular browser requests. This is typically done to fool search engines into directing more user traffic to a website and is often done to avoid accountability or to provide enhanced content only to those browsers that can display it.

Not only must web archivists deal with the technical challenges of web archiving, they must also contend with intellectual property laws. Peter Lyman^[14] states that "although the Web is popularly regarded as a public domain resource, it is copyrighted; thus, archivists have no legal right to copy the Web". However national

libraries in some countries^[15] have a legal right to copy portions of the web under an extension of a legal deposit.

Some private non-profit web archives that are made publicly accessible like WebCite, the Internet Archive or the Internet Memory Foundation allow content owners to hide or remove archived content that they do not want the public to have access to. Other web archives are only accessible from certain locations or have regulated usage. WebCite cites a recent lawsuit against Google's caching, which Google won.^[16]

Laws

In 2017 the Financial Industry Regulatory Authority, Inc. (FINRA), a United States financial regulatory organization, released a notice stating all the businesses doing digital communications are required to keep a record. This includes website data, social media posts, and messages.^[17] Some copyright laws may inhibit Web archiving. For instance, academic archiving by Sci-Hub falls outside the bounds of contemporary copyright law. The site provides enduring access to academic works including those that do not have an open access license and thereby contributes to the archival of scientific research which may otherwise be lost.^{[18][19]}

See also

- Anna's Archive
- Archive site
- Archive Team
- archive.today (formerly archive.is)
- Collective memory
- Common Crawl
- Digital hoarding
- Digital preservation
- Digital library
- Ghost Archive
- Google Cache
- List of Web archiving initiatives
- Memento Project
- Minerva Initiative
- Mirror website
- National Digital Information Infrastructure and Preservation Program (NDIIPP)
- National Digital Library Program (NDLP)
- PADICAT
- PageFreezer
- Pandora Archive
- UK Web Archive
- Virtual artifact
- Wayback Machine
- Web crawling
- WebCite
- Webrecorder



General bibliography

- Brown, A. (2006). *Archiving Websites: A Practical Guide for Information Management Professionals*. London: Facet Publishing. ISBN 978-1-85604-553-7.
- Brügger, N. (2005). *Archiving Websites. General Considerations and Strategies* (<https://web.archive.org/web/20090129171453/https://www.cfi.au.dk/en/publications/cfi>). Aarhus: The Centre for Internet Research. ISBN 978-87-990507-0-3. Archived from the original (<https://www.cfi.au.dk/en/publications/cfi>) on January 29, 2009.
- Day, M. (2003). "Preserving the Fabric of Our Lives: A Survey of Web Preservation Initiatives" (<https://purehost.bath.ac.uk/ws/files/569662/ecdl-2003-final.pdf>) (PDF). *Research and Advanced Technology for Digital Libraries*. Lecture Notes in Computer Science. Vol. 2769. pp. 461–472. doi:10.1007/978-3-540-45175-4_42 (https://doi.org/10.1007%2F978-3-540-45175-4_42). ISBN 978-3-540-40726-3. Archived (<https://web.archive.org/web/20231029012250/https://purehost.bath.ac.uk/ws/files/569662/ecdl-2003-final.pdf>) (PDF) from the original on October 29, 2023. Retrieved November 16, 2023.
- Eysenbach, G. & Trudel, M. (2005). "Going, going, still there: using the WebCite service to permanently archive cited web pages" (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1550686>). *Journal of Medical Internet Research*. 7 (5) e60. doi:10.2196/jmir.7.5.e60 (<https://doi.org/10.2196%2Fjmir.7.5.e60>).

- PMC 1550686 (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1550686>). PMID 16403724 (<https://pubmed.ncbi.nlm.nih.gov/16403724>).
- Fitch, Kent (2003). "Web site archiving—an approach to recording every materially different response produced by a website" (<https://web.archive.org/web/20030720111610/https://ausweb.scu.edu.au/aw03/papers/fitch/>). *Ausweb 03*. Archived from the original (<https://ausweb.scu.edu.au/aw03/papers/fitch/>) on July 20, 2003. Retrieved September 27, 2006.
 - Jacoby, Robert (August 19, 2010). "Archiving a Web Page" (<https://web.archive.org/web/20110103095915/https://www.seoq.com/archiving-a-web-page/>). *seoq.com*. Archived from the original (<https://www.seoq.com/archiving-a-web-page/>) on January 3, 2011. Retrieved October 23, 2010.
 - Lyman, Peter (2002). "Archiving the World Wide Web" (<https://www.clir.org/pubs/reports/pub106/web.html>). *Building a National Strategy for Preservation: Issues in Digital Media Archiving* (<https://www.clir.org/wp-content/uploads/sites/6/pub106.pdf>) (PDF). Council on Library and Information Resources. pp. 38–51. ISBN 978-1-887334-91-4.
 - Masanès, J., ed. (2006). *Web Archiving*. Berlin: Springer-Verlag. ISBN 978-3-540-23338-1.
 - Pennock, Maureen (2013). *Web-Archiving*. DPC Technology Watch Reports. Great Britain: Digital Preservation Coalition. doi:10.7207/twr13-01 (<https://doi.org/10.7207%2Ftwr13-01>). ISSN 2048-7916 (<https://search.worldcat.org/issn/2048-7916>).
 - Toyoda, M.; Kitsuregawa, M. (2012). "The History of Web Archiving" (<https://doi.org/10.1109%2FJPROC.2012.2189920>). *Proceedings of the IEEE*. **100** (special centennial issue): 1441–1443. doi:10.1109/JPROC.2012.2189920 (<https://doi.org/10.1109%2FJPROC.2012.2189920>).

References

1. "Web Archiving" (<https://netpreserve.org/web-archiving/>). *Netpreserve - International Internet Preservation Consortium*. August 14, 2024. Archived (<https://web.archive.org/web/20240712174404/https://netpreserve.org/web-archiving/>) from the original on July 12, 2024.
2. Ringel, Sharon; Ribak, Rivka (January 1, 2024). "Platformizing the Past: The Social Media Logic of Archival Digitization" (<https://doi.org/10.1177%2F20563051241228596>). *Social Media + Society*. **10** (1) 20563051241228596. doi:10.1177/20563051241228596 (<https://doi.org/10.1177%2F20563051241228596>). ISSN 2056-3051 (<https://search.worldcat.org/issn/2056-3051>).
3. Truman, Gail (2016). "Web Archiving Environmental Scan" (<https://nrs.harvard.edu/urn-3:HUL.InstRepos:25658314>). *Harvard Library*.
4. Toyoda, M.; Kitsuregawa, M. (May 2012). "The History of Web Archiving" (<https://doi.org/10.1109%2FJPROC.2012.2189920>). *Proceedings of the IEEE*. **100** (Special Centennial Issue): 1441–1443. doi:10.1109/JPROC.2012.2189920 (<https://doi.org/10.1109%2FJPROC.2012.2189920>). ISSN 0018-9219 (<https://search.worldcat.org/issn/0018-9219>).
5. Crockett, Zachary (September 28, 2018). "Inside Wayback Machine, the internet's time capsule" (<https://thehustle.co/inside-wayback-machine-internet-archive>). *The Hustle*. sec. Wayyyy back. Archived (<https://web.archive.org/web/20181002145800/https://thehustle.co/inside-wayback-machine-internet-archive>) from the original on October 2, 2018. Retrieved July 21, 2020.
6. Costa, Miguel; Gomes, Daniel; Silva, Mário J. (September 2017). "The evolution of web archiving". *International Journal on Digital Libraries*. **18** (3): 191–205. doi:10.1007/s00799-016-0171-9 (<https://doi.org/10.1007/s00799-016-0171-9>). S2CID 24303455 (<https://api.semanticscholar.org/CorpusID:24303455>).
7. Consalvo, Mia; Ess, Charles, eds. (April 2011). "Web Archiving – Between Past, Present, and Future" (<https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781444314861.ch2>). *The Handbook of Internet Studies* (<https://onlinelibrary.wiley.com/doi/book/10.1002/9781444314861>) (1 ed.). Wiley. pp. 24–42. doi:10.1002/9781444314861 (<https://doi.org/10.1002%2F9781444314861>). ISBN 978-1-4051-8588-2. Archived (<https://web.archive.org/web/20220910053354/https://onlinelibrary.wiley.com/doi/book/10.1002/9781444314861>) from the original on September 10, 2022. Retrieved September 11, 2022.
8. "IWAW 2010: The 10th Intl Web Archiving Workshop" (<https://web.archive.org/web/2020112212839/http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=9651>). *WikiCFP*. Archived from the original (<https://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=9651>) on November 12, 2020. Retrieved August 19, 2019.
9. "IWAW - International Web Archiving Workshops" (<https://web.archive.org/web/20121120035514/https://bibnum.bnf.fr/ecdl/index.html>). *bibnum.bnf.fr*. Archived from the original (<https://bibnum.bnf.fr/ecdl/index.html>) on November 20, 2012. Retrieved August 19, 2019.

10. "About the IIPC" (<https://netpreserve.org/about-us/>). *IIPC*. Retrieved April 17, 2022.
11. "Internet Memory Foundation: Free Web: Free Download, Borrow and Streaming" (<https://archive.org/details/internetmemoryfoundation>). *archive.org*. Internet Archive. Retrieved July 21, 2020.
12. Regis, Camille (June 4, 2019). "Web Archiving: Think the Web is Permanent? Think Again" (<https://www.historyassociates.com/resources/blog/web-archiving-challenges/>). History Associates. Archived (<https://web.archive.org/web/20190715150212/https://www.historyassociates.com/resources/blog/web-archiving-challenges/>) from the original on July 15, 2019. Retrieved July 14, 2019.
13. Brown, Adrian (January 10, 2016). *Archiving websites: a practical guide for information management professionals*. Facet. ISBN 978-1-78330-053-2. OCLC 1064574312 (<https://search.worldcat.org/oclc/1064574312>).
14. Lyman (2002)
15. "Legal Deposit I IIPC" (<https://netpreserve.org/legal-deposit>). *netpreserve.org*. Archived (<https://web.archive.org/web/20170316103200/https://netpreserve.org/legal-deposit>) from the original on March 16, 2017. Retrieved January 31, 2017.
16. "WebCite FAQ" (<https://www.webcitation.org/faq>). *Webcitation.org*. Retrieved September 20, 2018.
17. "Social Media and Digital Communications" (https://www.finra.org/sites/default/files/notice_doc_file_ref/Regulatory-Notice-17-18.pdf) (PDF). *finra.org*. FINRA.
18. Claburn, Thomas (September 10, 2020). "Open access journals are vanishing from the web, Internet Archive stands ready to fill in the gaps" (https://www.theregister.com/2020/09/10/open_access_journal/). *The Register*. Archived (https://web.archive.org/web/20211029191612/https://www.theregister.com/2020/09/10/open_access_journal/) from the original on October 29, 2021. Retrieved October 22, 2020.
19. Laakso, Mikael; Matthias, Lisa; Jahn, Najko (2021). "Open is not forever: A study of vanished open access journals". *Journal of the Association for Information Science and Technology*. **72** (9): 1099–1112. arXiv:2008.11933 (<https://arxiv.org/abs/2008.11933>). doi:10.1002/ASI.24460 (<https://doi.org/10.1002%2FASI.24460>). S2CID 221340749 (<https://api.semanticscholar.org/CorpusID:221340749>).

External links

- International Internet Preservation Consortium (IIPC) (<https://www.netpreserve.org/>)—International consortium whose mission is to acquire, preserve, and make accessible knowledge and information from the Internet for future generations
- National Library of Australia, Preserving Access to Digital Information (PADI) (<https://www.nla.gov.au/padi/topics/92.html>)
- Library of Congress—Web Archiving (<https://www.loc.gov/webarchiving/>)
- Data Hoarding non-profit organization (<https://datahoarding.org/>)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Web_archiving&oldid=1316852182"