Archivage des données du web : sauvegarde et citabilité pour la recherche

Jean-Christophe Peyssard (CNRS – MMSH UAR 3125)
Ecole d’été DISTAM Juillet 2022
1. Courte introduction aux archives du Web
2. Des outils pour la collecte
The World Wide Web by numbers

• “The first-ever website (info.cern.ch) was published on August 6, 1991 by British physicist Tim Berners-Lee while at CERN, in Switzerland”

• In the June 2022 survey we received responses from 1,146,976,964 sites across 273,010,403 unique domains and 12,224,786 web-facing computers. This reflects a loss of 8.75 million sites and 583,000 domains, but a gain of 155,000 computers. [from previous 30th May, 2022], https://news.netcraft.com/archives/2022/05/30/may-2022-web-server-survey.html

• “The Indexed Web contains at least 5.36 billion pages” (Monday, 04 July, 2022, https://www.worldwidewebsize.com/)
The structure of URLs on the World Wide Web (www):

protocol://subdomain.domain.top-level domain/path/page/

Ex:

https://distam.hypotheses.org/category/ecole-dete
Life and death of Web pages

“The average life span of a Web page is only 44 days, and 44 percent of the Web sites found in 1998 could not be found in 1999. [...] As ubiquitous as the Web seems to be, it is also ephemeral, and much of today's Web will have disappeared by tomorrow.” (Lyman, 2002 p. 38)

“40% of the material on the Internet disappears within a year, while another 40% has been changed, which is why today we can only expect to find 20% of the material that was on the Internet one year ago.” (Brügger, 2005 p. 15)

“We now know that Web pages only last about 100 days on average before they change or disappear.” (Kahle, 2015)

In 2013, the average life span of a URL is 9.3 years (Musiani at al., 2019, https://books.openedition.org/oep/8743)

In 2019, according to the Wayback Machine team the average life span of a Web page is 92 days (2018)

“Since January 2013, 66.5% of the links pointing to the 2,062,173 websites we sampled have rotted. We found another 6.45% with temporary errors. We don’t know if they’re still there or not.” (Stox, 2022, https://ahrefs.com/blog/link-rot-study/)
When the Web is broken

A problem for research and scholarly communication

- Error 404, Broken links, Link rot, Reference rot, Infosuicide, digital ruins, content drift, zombie media...
- Shut down & take down, mergers and acquisitions:
  On March 18, 2019, it was revealed that MySpace lost all of their user content from 2016 and earlier in "a server migration gone wrong". It was widely reported that over 50 million songs and 12 years worth of content was permanently lost, and there was no backup (https://en.wikipedia.org/wiki/Myspace)
- History:
  - Yougoslavia (.yu) breakup (now Serbia and Montenegro, .rs and .me)
  - Czechoslovakia (.cs) dissolution (now Czech Republic and Slovakia, .cz and .sk)
- Reference rot, a combination of:
  - Content decay: The content of the linked resource may change over time and, as a result, the degree to which that content remains representative of the content that was intended to be linked to may decrease over time.
  - Link rot: The linked resource may disappear altogether. (Thoughts on Referencing, Linking, Reference Rot http://mementoweb.org/missing-link/)
- The integrity of research is at risk! (James G. Neal, http://library.ifla.org/id/eprint/907)
Why Web Archiving?

- To maintain our digital cultural heritage
- To stabilize and preserve web materials as a research object
- To be able to document and illustrate claims based on analyses of web materials (whether the web itself is the research object or a source of knowledge about other research objects).

A definition of Web archiving

- Web archiving is the process of collecting portions of the World Wide Web, preserving the collections in an archival format, and then serving the archives for access and use. (IIPC Web site, http://netpreserve.org/web-archiving/)

- “Web archiving is 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, J. (2012). An Overview of Web Archiving. *D-Lib Magazine*, 18 (3/4). https://doi.org/10.1045/march2012-niu1)
A short chronology of Web archives

1537 Legal deposit in France (1619 Spain, 1710 United Kingdom)  
1989 The World Wide Web was invented by Tim Berners-Lee  
1996 Internet Archive is founded by Brewster Kahle  
1996 Kulturarw3 in Sweden for archiving the .se top level domain name  
1998 Google is launched  
2001 The Wayback Machine (Internet Archive)  
2003 UNESCO Charter on the Preservation of Digital Heritage  
2003 the International Internet Preservation Consortium is formally chartered at the National Library of France with 12 participating institution  
2005 Youtube is launched  
2006 [The] Facebook and Twitter are launched  
2006 National Library of France is in charge of collecting and preserving the “French Internet” (new French Copyright Law) alongside with the National Audiovisual Institute (Ina)  
2013 EU Web Archive
Different strategies & methods for Web archiving

- Macro archiving
- Micro archiving
- Thematic or selective archiving
- Bulk or snapshot harvesting, or broad crawls
- “Exhaustivity” of a National domain name archiving (.se in Sweden)
- Event and ‘real-time’ institutional archiving (after 2015 terrorist attacks and Notre Dame fire in 2019 France)
- Shared archiving among institutions (in France btwn BnF and Ina)

...
Different ways of accessing Web archives

  The contents of the archive can be only accessed from special legal deposit workstations that are available in selected libraries within Finland (including The National Library of Finland).

- Portugal (since 2008) accessible in Open Access at [Arquivo.pt](http://arquivo.pt)


There is a lot of different types of Web archives

As much as for other kind of archives, one must know the history of a Web archive and how it was constructed to better understand it and use it in research work. What you see is a reconstruction, not a copy of the site.

- “What is harvested is both a point in time (the time of harvesting) and a period of time (the period up to the time of harvesting).” (Brügger, 2008 p. 158)

- “On the one hand the archive does not look like the internet as it actually was in the past (we have lost something), but on the other hand the archive might look like the internet as it never was in the past (we get something different).” (Brügger, 2001 p. 6)

Web archiving projects often needs to gather diverse and multiples expertises and skills : archivists and librarians, researchers, legal officers, IT and computer specialists and... users and stakeholders
Legal and ethical issues

As for any other kind of archives one must act lawfully and ethically when archiving and using Web archives:

● The materials in the web archives are protected by copyright law as they were on the live web
● There is “tensions around the archival principles of preserving the public record vs the individual’s expectation of the right to be forgotten” (http://netpreserve.org/ga2018/programme/abstracts/#paper21)
● The processing of personal data is submitted to laws and even more to the research project ethics
● New laws to take into account ex. General Data Protection Regulation (GDPR, https://gdpr-info.eu/)
The web is a unique and dynamic resource that is of high value to current and future researchers.

Learn about the value of our work.

IIPC Members
Members are organizations from over 45 countries, including national, university and regional libraries and archives.

Working groups
IIPC members join working groups that engage in short and long-term projects to advance the practice of web archiving.

Events
Our community comes together annually to share experiences and present solutions during the Web Archiving Conference and the General Assembly.

http://netpreserve.org/
New research fields and projects

https://respadon.hypotheses.org/

https://resaw.eu/

https://web90.hypotheses.org/

https://archivesunleashed.org/
Challenges and setbacks in Web archiving

- Robots.txt
- Captcha (ie Completely Automated Public Turing-test to tell Computers and Humans Apart)
- User interaction needed
- Password protected content
- Technologies and dynamic content : Flash, java scripts,…
- Distant content
- Temporal inconsistencies
- Bot traps
- …
Method and tools for Web archiving

- Define a strategy
- Use a log and write throughout the life of the project
- You may need to use additional methods and tools

- Screen capture and screen recording
- Link crawling
- On demand archiving
- ...
A standard format for Web archives

- WARC file format = Web ARChive archive format
- ARC was accepted as an international standard in 2009 (ISO 28500:2009)
- WARC is now recognised by most national library systems as the standard to follow for web archival

https://en.wikipedia.org/wiki/Web_ARChive
https://wiki.archivematica.org/Significant_characteristics_of_websites
https://wiki.archivematica.org/File:WARCdiagram.png
Internet Archive’s Wayback Machine

- Launched in 2001
- 357 billion archived Web pages so far
- Archived content going back to 1996

https://archive.org/
Structure of URLs in the Internet Archive’s Wayback Machine:

**Wayback Machine URL/collection/time shown as yyyymmddhhmmss/URL**

**Ex:**

https://web.archive.org/web/20220707064851/
https://distam.hypotheses.org/2656
Internet Archive / Wayback Machine extension for browsers

Welcome to the Official Internet Archive Wayback Machine Browser Extension! Go back in time to see how a website has changed through the history of the Web. Save websites, view missing 404 Not Found pages, or read archived books & papers.

Read all 236 reviews


https://chrome.google.com/webstore/detail/wayback-machine/fpnmgdkabkmnadcjpehmlllkndpkmiak?hl=fr
Conifer
Collect and revisit web pages.

Conifer is a web archiving service that creates an interactive copy of any web page that you browse, including content revealed by your interactions such as playing video and audio, scrolling, clicking buttons, and so forth.

Create a Free Account

Existing Users Login

Free accounts with 5GB of storage. Get more and support this project by becoming a supporter.

https://conifer.rhizome.org/
You're Invited: Rhizome's 2022 Benefit Honoring Rachel Rossin & Julie Martin

Studies in Perception IV: Julie Martin

Off: Endgame

https://rhizome.org/
Webrecorder provides a **suite of open source tools and packages** to **capture** interactive websites and **replay** them at a later time as **accurately as possible**. Learn more about our key tools and efforts:

- **ArchiveWeb.page**
- **ReplayWeb.page**
- **pywb**
- **Browsertrix Crawler**

[https://webrecorder.net/](https://webrecorder.net/)
Interactive browser-based web archiving, a Webrecorder project.

ArchiveWeb.page is the latest tool from Webrecorder to turn your browser into a full-featured interactive web archiving system!

ArchiveWeb.page is available as an extension for any Chrome or Chromium based browsers. (A standalone app version is also in development.)

To create web archives, the extension (or app) will be needed. Once created, the archives can be viewed in any modern browser using ReplayWeb.page -- no extension required!

View the ArchiveWeb.page User Guide

Sorry, the ArchiveWeb.page browser Extension requires Chrome or a Chromium based browser. Please try ArchiveWeb.page in Chrome or check back soon for additional options.

https://archiveweb.page/
An original methodological framework for approaching the archived web, both as a source and as an object of study in its own right.

As life continues to move online, the web becomes increasingly important as a source for understanding the past. But historians have yet to formulate a methodology for approaching the archived web as an object in its own right.
Thank you / Merci / شكراً

Ecole d’été DISTAM 2022

jean-christophe.peyssard@univ-amu.fr