Archives of the World Wide Web

How to access the internet of the past

Dr. Nuria Plattner
The World Wide Web - a huge, but short-lived database

- Most content in the internet is short-lived, as webpages are updated, replaced or removed.
- The average webpage lifetime is less than a year, depending on the definitions used.
- Web content is therefore continuously disappearing.
- The use of the internet as a huge database is hampered by the short lifetime of its content.
Web archiving purpose and challenges

Why are internet archives important?

- Internet archives can save content from disappearing completely or being untraceable
- Web archives main advantages:
  - **find missing content**
  - **cite web content**
  - **certified web records** (e.g. for legal proceedings)
- **Challenge**: the number of webpages is still increasing and the content is continuously updated

→ **Web archives therefore need to focus on specific content and limited time intervals**

Number of Websites from 1991 to 2021 according to *Statista*

* As of August 1, 1991.
Source: Internet Live Stats
How to archive the internet?
Creating a web archive

- Living Web
  - Selection: Create list of URLs to archive
  - Crawl and fetch: Access URLs with web crawlers and fetch content
  - Quality assurance: Automatic or manual

- Access: Create index and manage access

- Web Archive
How to archive the internet?

Technical challenges

• Web crawling methods (avoid ads, detect spam, …)

• Personalized content: different page versions shown for different users

• Archiving weblinks: include linked content in archiving?

• Websites are updated in irregular time intervals – how often should they be archived?
Limitations of public internet archives

What content can they provide?

• Limited crawling of linked pages in order to keep the content manageable
• Some webpage types can not be crawled e.g. non-public databases
• Many pages can not be made accessible publicly due to copyright or license restrictions
• Web archives often have specific agreements and permissions to store and provide content

Web archiving projects over time

Development of archival resources

- The number of web archives is increasing over time
- Many countries maintain their own web archives

Countries hosting web archiving services in 2014

Useful Resources

The Internet Archive

@ https://archive.org

• Non-Profit organization headquartered in San Francisco

• Digital library of internet sites

• In addition, it provides E-books, audio recordings, videos and more to registered users (free of charge)

• Wayback Machine: archive of web pages; webpage versions are collected over time
Useful Resources

Archive.today

- Archive site storing snapshots of webpages
- Users can create new snapshots or access the archive
- In addition to searching for URLs, a full text search is provided

@ https://archive.today

archive.today webpage capture

My url is alive and I want to archive its content
http://www.domain.com/url

Archive.today is a time capsule for web pages:
It takes a ‘snapshot’ of a webpage that will always be online even if the original page disappears.
It saves a text and a graphical copy of the page for better accuracy and provides a short and reliable link to an unalterable record of any webpage.
Including those from Web 2.0 sites:

This can be useful if you want to take a ‘snapshot’ of a page which could change soon (price list, job offer, real estate listing, drunk blog post, ...)
Saved pages will have no active elements and to scripts, so they keep you safe as they cannot have any popups or malware.

I want to search the archive for saved snapshots

search queries by example
- *.microsoft.com for snapshots from the-host microsoft.com
- *.teknik.com for snapshots from teknik.com and all its subdomains (e.g. www.microsoft.com)
- http://thesite.com/* for snapshots from exact url (search is case-sensitive)
Useful Resources

National Web Archives

- Many countries have their own national web archives storing URLs in the countries web domain
- Access to national archives is often very restricted, e.g. only metadata search possible via remote access
- For webpages with no specific intellectual property or access restrictions, it makes sense to search in other web archives first

Web Archive Switzerland
https://www.e-helvetica.nb.admin.ch

UK Web Archive
https://www.webarchive.org.uk

DEUTSCHE NATIONALBIBLIOTHEK
http://dnb.oia-dwa.de/index.aspx
Useful Resources

Time Travel (Memento project)

@ http://timetravel.mementoweb.org

- Part of the Memento project led by Los Alamos National Laboratory and Old Dominion University
- Searches for specific URLs in various web archives
- To be included in the search, web archives need to support the Memento "Time travel for the Web" protocol
Archives of the World Wide Web

Summary and Conclusions

- Web content is short-lived; web archives can help finding missing content and provide certified web records

- Web archives provide access to different versions of web pages over time

- Archiving is either done automatically or upon users request

- Challenges: capture entire web sites with full functionality, provide content to users without intellectual property or license restrictions

- Useful resources: various organizations offer free access to their web archives
Thanks for your attention

Questions?

Coffee & Bit(e)s
THE COFFEE LECTURES FOR SCIENTISTS

UNIVERSITY LIBRARY BERN
Nuria Plattner, Aline Frank, Michael Horn, and Silvan Christen
www.unibe.ch/ub/sciencelibrary