

# Coffee & Bit(e)s

THE COFFEE LECTURES FOR SCIENTISTS

u<sup>b</sup>

b  
UNIVERSITÄT  
BERN

## Archives of the World Wide Web

*How to access the internet of the past*

Dr. Nuria Plattner

UNIVERSITY LIBRARY BERN

Nuria Plattner, Aline Frank, Michael Horn, and Silvan Christen

[www.unibe.ch/ub/sciencelibrary](http://www.unibe.ch/ub/sciencelibrary)



# 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

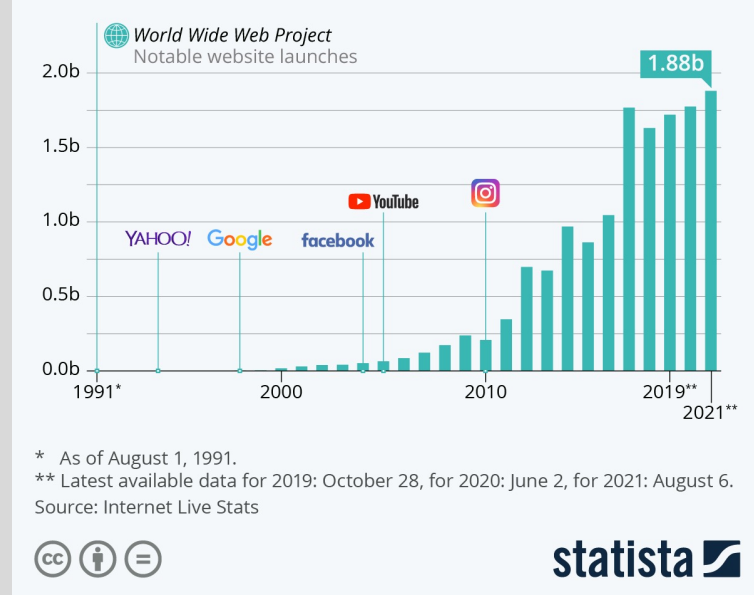


# Web archiving purpose and challenges

## Why are internet archives important?

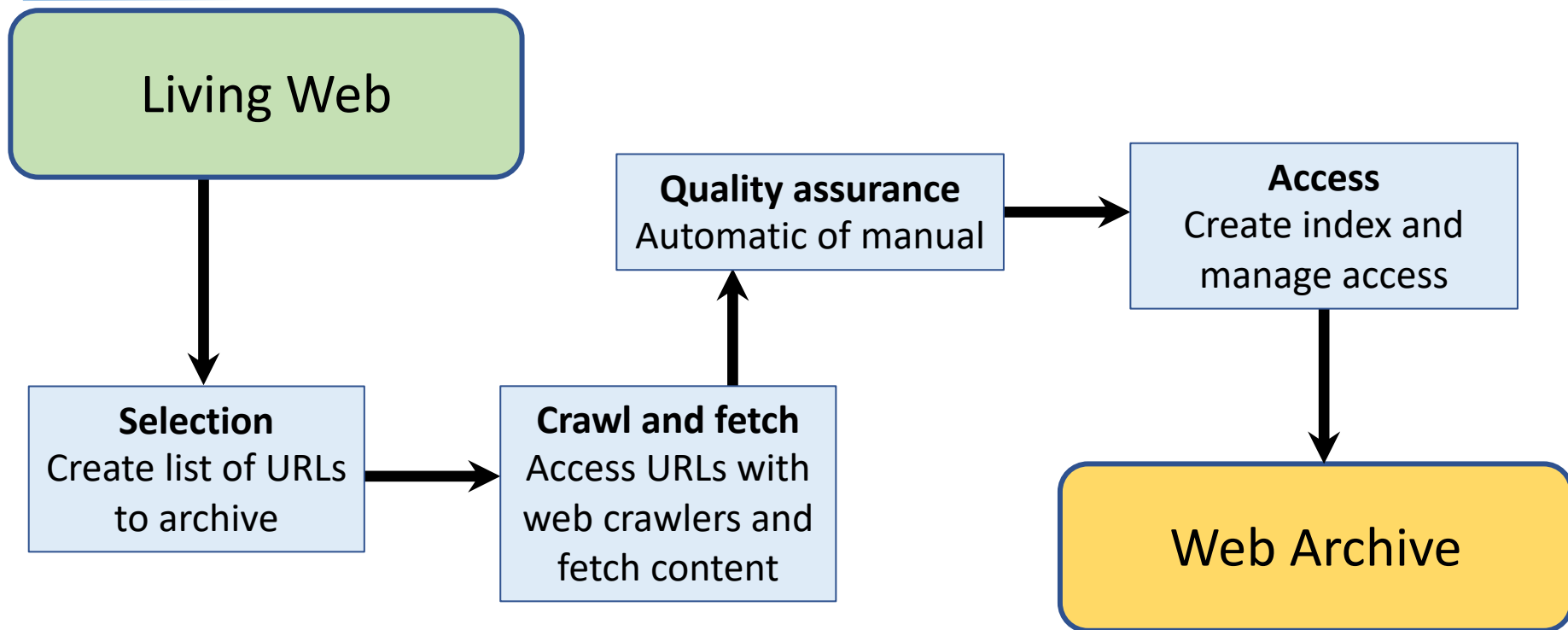
- Internet archives can save content from disappearing completely or being untraceable
- Web archives main advantages:
  - *find missing content*
  - *cite web content*
  - *certified web records (e.g. for legal proceedings)*
- **Challenge:** the number of webpages is still increasing and the content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



# How to archive the internet?

## Creating a web archive



# How to archive the internet?

## Technical challenges

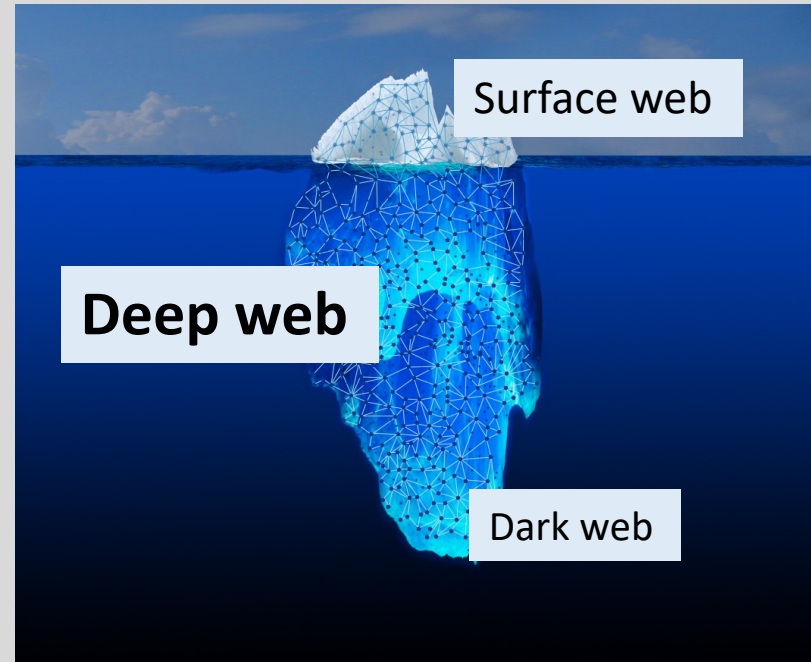
- Web crawling methods (avoid ads, detect spam, ...)
- Personalized content: different page versions shown for different users
- Archiving weblinks: include linked content in archiving?
- Websites are updated in irregular time intervals – how often should they be archived?



# 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

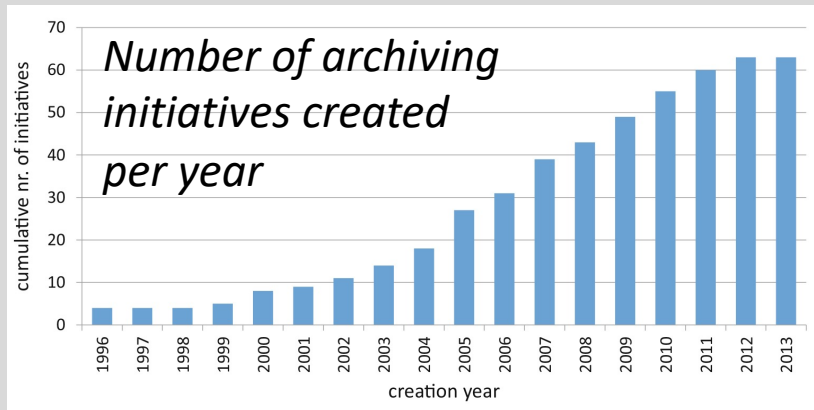


<https://phys.org/news/2015-05-deep-web-scientists.html>

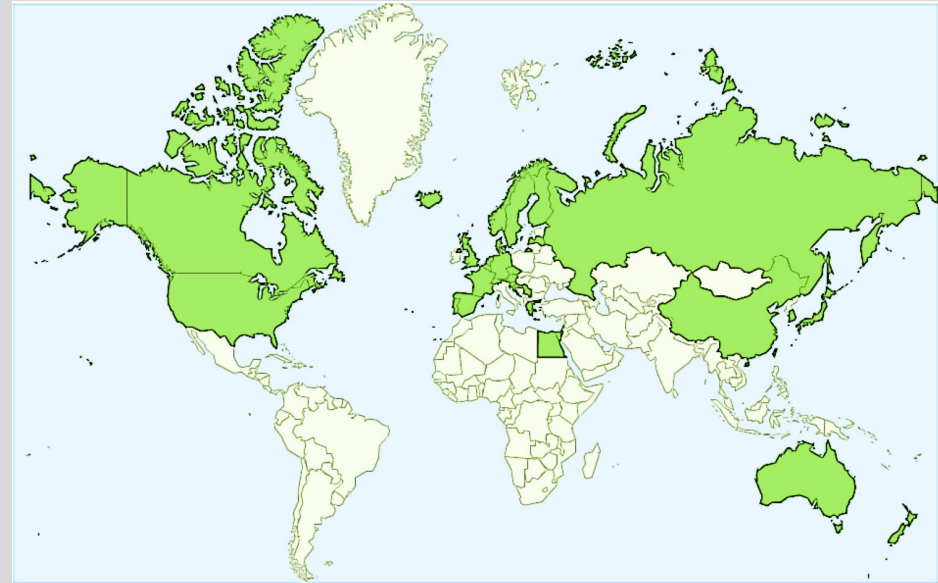
# Web archiving projects over time

## 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*



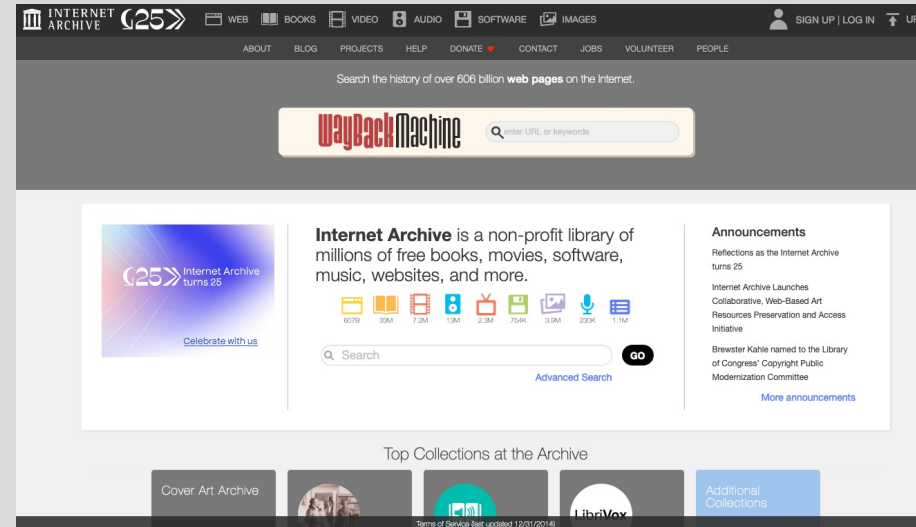
M. Costa *et al.*, *Int. J. Digit. Libr.* **18**, 191–205, (2017)

# 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



The screenshot shows the Internet Archive website homepage. At the top, there is a navigation bar with the Internet Archive logo and a '25' anniversary badge. Below the navigation bar, there is a search bar for the Wayback Machine. The main content area features a large banner for the 25th anniversary of the Internet Archive, a search bar, and a section for 'Top Collections at the Archive' which includes links to the Cover Art Archive, Internet Archive, and LibriVox. The footer contains the text 'Terms of Service Last updated: 12/31/2014'.



# Useful Resources

## Archive.today

@ <https://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

The screenshot shows the Archive.today website interface. At the top, there are navigation links for 'email', 'blog', 'ask me', 'FAQ', and 'Donate'. A prominent red banner reads 'My url is alive and I want to archive its content'. Below this banner is a text input field containing 'http://www.domain.com/url' and a 'save' button. The main content area explains that Archive.today is a time capsule for web pages, taking a 'snapshot' that remains online even if the original page disappears. It lists examples of URLs that can be archived, such as 'https://archive.ph/2020\_04\_21/rt.live/' and 'https://archive.ph/2014\_06\_26/google.com/maps/...'. A search bar is located at the bottom, with the text 'I want to search the archive for saved snapshots' above it. Below the search bar, there are example search queries: 'microsoft.com' for snapshots from the host, '\*microsoft.com' for snapshots from the domain and subdomains, 'http://twitter.com/burgenking' for exact URIs, and 'http://twitter.com/burg\*' for URIs starting with a specific prefix.

# 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

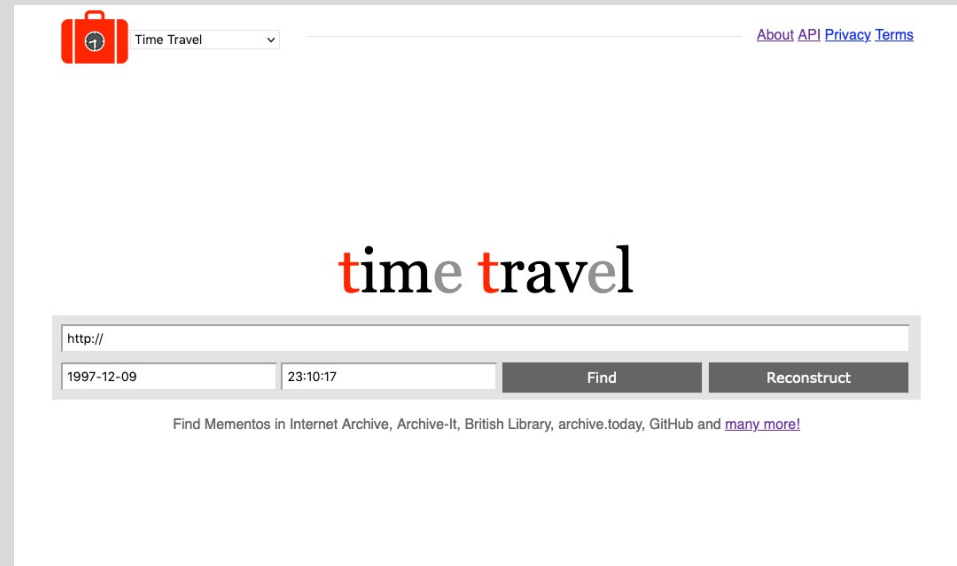


# Useful Resources

## 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



## 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?

*u<sup>b</sup>*

u<sup>b</sup>  
UNIVERSITÄT  
BERN



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](http://www.unibe.ch/ub/sciencelibrary)