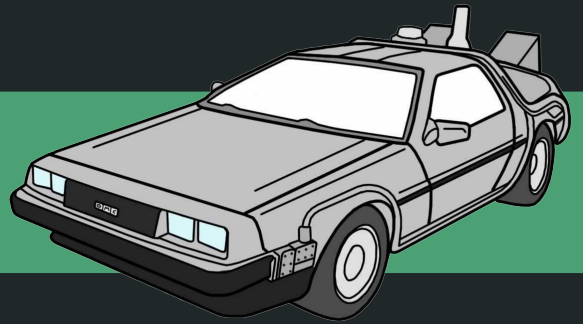
The Internet Archive and *The Wayback Machine*



The Internet Archive (IA) is a non-profit that was founded in 1996 to build an Internet library. Its primary purpose is to support a free and open internet by offering permanent access to historical collections that exist in digital format.

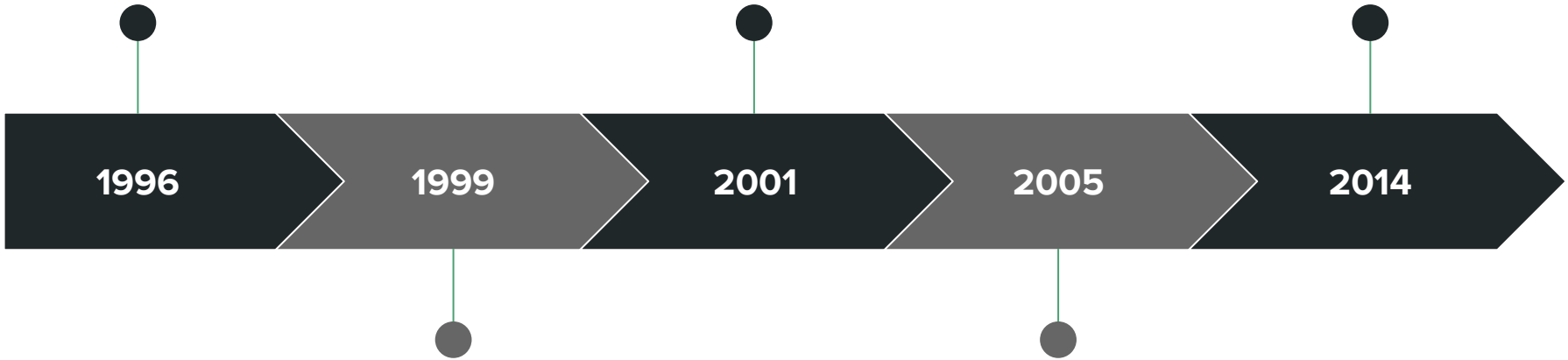
It currently holds more than 20 petabytes (that's *20 million gigabytes*) of data in its collections.

archive.org

The IA includes:

- ★ text
- ★ audio
- ★ video
- ★ images
- ★ software
- ★ web pages

It also provides adaptive reading services and information access for the blind and others with disabilities.



IA is founded with support from Alexa Internet

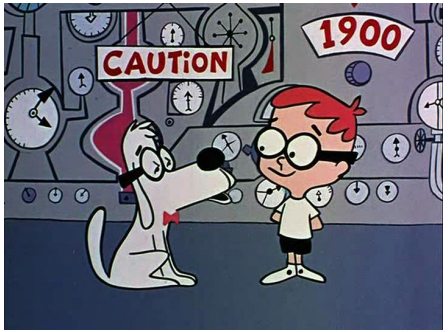
The Wayback Machine is launched, providing access to the general public

Collections reach 15 petabytes of data stored and 400 billion web pages archived

IA expands beyond Web archiving, beginning with the Prelinger film archives

Archive-It.org is launched as subscription service for institutions to create digital libraries and website backups

INTERNET ARCHIVE
WayBackMachine



The Wayback Machine was designed by Brewster Kahle and launched in 2001.

- ➔ Automated web crawlers periodically capture websites as they are at that moment in time.
- ➔ Designed to be searchable, usable, and able to be referenced.
- ➔ “Save Page Now” feature allows users to add pages to the Archive.
- ➔ 600,000+ users a day.
- ➔ 469 billion web pages archived.

What is it used for?

Reviving dead links

65% of users are attempting to view a dead link encountered elsewhere.

Viewing a web site from a particular date/time

Nostalgia, design development, sites/pages that are no longer maintained.

Tracking the evolution of the Internet and of language

How have design, interfaces, usability changed over time? How has language developed?

Researching contemporary history

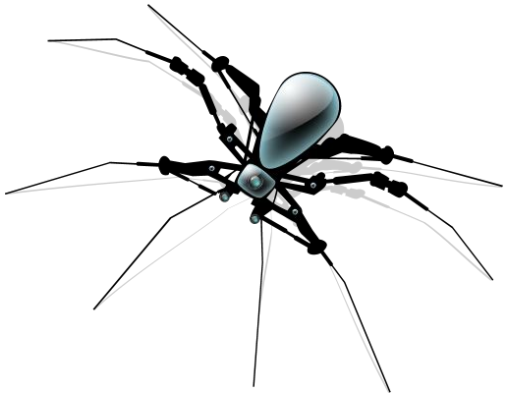
Example: looking at a news site on a historically significant date.

AlNoamany, Yasmin, Ahmed AlSum, Michele Weigle, and Michael Nelson. 2014. "Who and what links to the Internet Archive." *International Journal On Digital Libraries* 14, no. 3/4: 101-115.

How does it work? (the simple version)

Open-source/in-house software

Kahle: “We use as much open source software as we can.”
Running on various operating systems, including Solaris and Linux.



Web crawlers

Mining and capture of publicly available websites and links.
Written using Perl and C.

Automation

Excludes pages with robots.txt file on a server that is set to disallow *User-Agent: ia_archiver.*

http://webservices.xml.com/pub/a/ws/2002/01/18/brewster.html
Shein, Esther. 2016. "Preserving the Internet." *Communications Of The ACM* 59, no. 1: 26-28.

How does it work? (the tech version)

Structure

Data is stored in WARC or ARC files which are written at time of capture by the Heritrix and other crawlers.

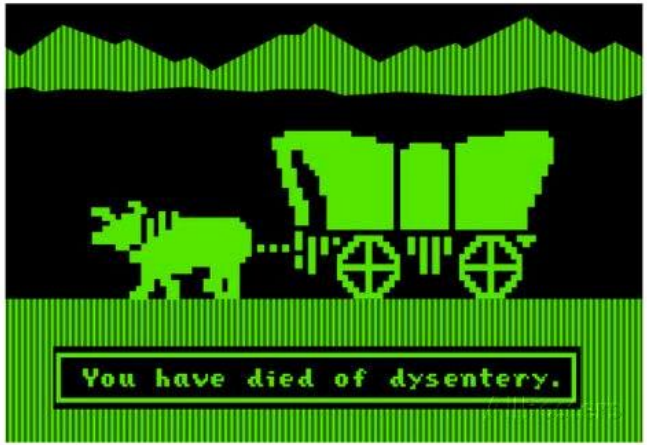
Indexing

“Three dimensional indexing”: 2-level index points into WARC data, one a compressed sorted list and the other CDX records.

Querying

Binary searching the first level list stored in core, then HTTP range-request loading the appropriate second-level blocks from the CDX index.

http://highscalability.com/blog/2014/5/19/a-short-on-how-the-wayback-machine-stores-more-pages-than-st.html



How it doesn’t work

1. Forms, JavaScript, and other interaction with host site.
2. No archiving of email, chats, or text messaging.
3. Orphan pages, bad links, broken images, and incomplete sites.

The Internet Archive and Rights

Copyright

- Copyright status stated at time of content upload.
- IA cannot guarantee copyright status of collection items, users cautioned to use them at their own risk.
- Most items labeled with a Creative Commons license.

Privacy

- Items can be removed from IA at the request of the owner.
- "The Internet Archive is not interested in preserving or offering access to Web sites or other Internet documents of persons who do not want their materials in the collection."

“Internet Archive's Terms of Use, Privacy Policy, and Copyright Policy”: https://archive.org/about/terms.php

archive.org