NATION WIDE WEBS
The Internet Archive

Non-Profit Library
Founded in 1996 by Brewster Kahle

Universal Access to All Knowledge
42,000,000,000,000,000,000 Bytes Archived

(42 Petabytes)
25,000  Software Titles
2,000,000  Moving Images
2,300,000  Book Archive
2,400,000  Audio Recordings
3,000,000  Hours of Television
4,000,000  eBooks
690,000,000,000+  URLs (~21PB)
40,000,000,000,000,000 Bytes Archived
(40 Petabytes)

Explore more than 286 billion web pages saved over time

Tools
Wayback Machine Availability API
Build your own tools.

WordPress Broken Link Checker
Banish broken links from your blog.

404 Handler for Webmasters
Help users get where they were going.

Subscription Service
Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit Archive-It to build and browse the collections.

Save Page Now
Capture a web page as it appears now for use as a trusted citation in the future. Only available for sites that allow crawlers.
Archive-It Demo Account
3.6 TB archived since Sep 22, 2010

Archive-It
Web Archiving SaaS since 2006
625 Organizational Users
NWW Scope of Work

- National web archive access portal that replays only specific ccTLD web resources from the entirety of IA’s historical web archive
- Multiple additional search and metadata features for enhanced discovery
- Possibility to build branded or embedded interfaces to match and institution’s design
- Inclusion of future ccTLD-specific content collected via domain, topical, contributed, and SPN harvests
NWW Technical Approach

- **Portal:** RESTful frontend web app built on Bootstrap and jQuery; no server apps used, just HTML+JavaScript
- **Indexing:** Dedicated APIs for search and summary stats at the website and URL levels serving the results and details
- **Access:** Replay uses IA’s new Python Wayback Machine
- **Search:** URL/site, keyword, and possible full-text search is powered by ElasticSearch on Internet Archive’s central ES cluster.
Search Details

Webpage search: Keyword search for website homepages and URL-text search with a search index built on the anchor text of inbound links and the terms in the page’s URL. Generally high 100Ms in prod.

Media Search: URL-based keyword search for all image, audio, video, and PDF media using terms in the URL and inbound anchor link text into a searchable text index. Generally low 100Ms in prod.

Media Linkback: Addition link to the webpage upon which the media file is embedded (when relevant), a la GifCities https://gifcities.org/.
NWW: Search Box

Enter search terms

Enter text or URL search | Defaults to homepage search

Multiple media types can be searched simultaneously

Text
Audio
PDFs
Images
Video
Web-only search will display a screenshot of the homepage, when available.

Image results will display a thumbnail of the image.

Other formats will display an icon indicating the media type.

Multi-format searches will return interleaved results in the order of the most relevant for each format.
NWW: Result Details

www.blog-ums-bier.de
blog ums bier
Earliest capture | Latest capture | 103 capture(s) from 2008 to 2017 | Site stats

FIRST CAPTURE IN WAYBACK MACHINE

LATEST CAPTURE IN WAYBACK MACHINE

LINKS TO WAYBACK MACHINE CALENDAR PAGE

WEB SITES LINK TO CALENDAR PAGE & MEDIA LINK DIRECTLY TO ARCHIVAL FILE

LINKS TO HOST/DOMAIN SITE STATISTICS
Additional Discovery: Potential for genre (blog, news, etc), topic (sports, culture), special ("lost" websites), or other unique points of entry for users to browse the collection.

Seed Submission or Nomination: Potential to integrate "Save Page Now" services into the portal.

Other Data: Potential to ingest archival web data collected outside of IA or custom crawls and include it in the portal.

Added Services: Include "special seeds" (lang, non-ccTLD ranked, geo-IP, etc) data. Integrated with specific services (Archive-It, etc), UI localization, site embedding, custom metadata, visualizations, etc.
NWW: WHY

Additional Discovery: Potential for genre (blog, news, etc), topic (sports, culture), special (“lost” websites), or other unique points of entry for users to browse the collection.
Seed Submission or Nomination: Potential to integrate “Save Page Now” services into the portal.
Other Data: Potential to ingest archival web data collected outside of IA or custom crawls and include it in the portal.
Added Services: Include “special seeds” (lang, non-ccTLD ranked, geo-IP, etc) data. Integrated with specific services (Archive-It, etc), UI localization, site embedding, custom metadata, visualizations, etc.
Why do we need a NWW and what does it say about the state of global data sharing and transnational access?
- Past efforts at data sharing has had various success and not necessarily achieved reciprocality
- Just as research outputs are often paywalled; web archive collections are often "roomwalled"
- Just as publicly-funded research outputs are often paywalled; publicly-funded web archive collections are often "roomwalled"
- Legal requirements often restrict access
HOW DO WE SOLVE:

- Access Issues
- Data Aggregation Issues
- Infrastructure Challenges
- Search Challenges
THANK YOU