Internet Archive

- We are a Digital Library
- Founded in 1996 by Brewster Kahle
- In 2007 officially designated a library by the state of California
- Located in San Francisco
Software Library: MS-DOS Games

Software for MS-DOS machines that represent entertainment and games. The collection includes action, strategy, adventure and other unique genres of game and entertainment.

MORE

SORT BY: VIEWS · TITLE · DATE ARCHIVED · CREATOR

2,605 RESULTS

Search this Collection

PART OF

The Software Library: MS-DOS
Software Library
www.archive.org
Largest publicly available web archive in existence

480+ Billion URLs
  80+ million websites,
content in 40+ Languages
Over 10 PetaBytes & growing at rate of 20TB per week

We collect a broad snapshot of the web every 60 days
What the web said yesterday

The Cobweb: Can the Internet be archived?

By Jill Lepore

New Yorker, January 26, 2015

http://www.newyorker.com/magazine/2015/01/26/cobweb
A web archive is a collection of archived URLs grouped by theme, event, subject area, or web address.

A web archive contains as much as possible from the original resources and documents the change over time. It is a priority to recreate the same experience a user would have had if they had visited the live site on the day it was archived.
Web Archiving Services

- Launched in 2004
- Work with partner organizations to create collections of web content.
- Organizations needed the ability to manage their own collections
- Lined up with our goal to become a self sustaining non profit
Archive-It is a subscription service launched in February 2006

- Web based application that allows users to create, manage, access and store collections of digital content
- The service is a fully hosted solution, and includes access and storage.
- Provides tools for selection and scoping including cataloging with metadata
- Partner organization or the Archive-It team can manage the account
- Ability to capture content using 10 different frequencies
- Archived web content includes: html, text, videos, audio, social media, PDF, images, password protected content, static databases and newspapers
- Can browse archived content 24 hours after a capture is complete; and full text search is available within 7 days
- Private access options are available
How do we collect the Content?

Open Source Technology

**Heritrix**: Web crawler – crawls and captures web pages.

**Umbra**: Assists the crawler to access social media and other sites in the same way a browser would.

**Wayback Machine**: Access tool for rendering and the viewing pages - surf the web as it was.

**NutchWAX**: Search engine – provides full-text search

**SOLR**: Metadata search
356 Partners in 48 U.S. States and 16 Countries

- **Universities & Colleges** (37%)
- **State Archives & Libraries** (20%)
- **Museums & Art Libraries** (8%)
- **Public Libraries & Local Governments** (8%)
- **National Organizations** (10%)
- **NGOs** (14%)
- **Other** (3%)
What our partners are collecting
Archives-It Use Cases

- Create a thematic/topical web archive on a specific subject or event.
  - Different perspectives and social commentary (tweets, blogs, comments).
  - Spontaneous events
  - Often related to traditional collecting activity around the same focus

- Mandate to capture/preserve evolving web history.
  - Support an electronic records system to meet records retention requirements.
  - Capture publications/documents that are no longer being deposited in print form.
  - Construct a historical record of an institution or individual web/social media presence over time.

- Closure crawls
University of Texas at Austin: LAGDA

Use Case: Archive government documents from 18 different countries in Latin America

Content includes:

Full-text versions of official documents

Original video and audio recordings of key regional leaders

Thousands of annual and "state of the nation" reports

Specific collections for Latin American elections and political parties
University of Texas at Austin: LANIC
Honduras Presidential site 2008 (before the Coup)
www.presidencia.gob.hn/
University of Texas at Austin: LANIC
Honduras Presidential site 2010 (after the Coup)
www.presidencia.gob.hn/

Sociedad civil acuerpa esfuerzos del Gobierno por lograr acuerdo con el FMI

COMUNICADO
La Secretaría de Relaciones Exteriores, a la opinión pública nacional e internacional hace saber que el gobierno de Honduras ejecuta un programa de cooperación antidrogas con el gobierno de los Estados Unidos de América.
Use Cases:

1. Collection on domain/university web presence in order to meet required mandates

2. Collections on focused thematic/topical subjects as designated by researchers and academics at the university
Commission Taps Brewery as a Landmark, Expands SoHo Historic District

The Landmarks Preservation Commission on Tuesday, May 11, 2010 unanimously approved the expansion of the SoHo-Cast Iron Historic District in Manhattan by 135 buildings, and for the first time approved the designation of a brewery as a New York City landmark. The designations bring to more than 27,000 the total number of properties throughout the City that have earned landmark protection since the Commission was founded in 1965.

- Read more
- Read the William Ulmer Brewery Designation Report
- Read the SoHo-Cast Iron Historic District Extension Designation Report

Mid-Century Modern Midtown Office Tower Now a Landmark

June 08, 2010
Public Meeting

June 22, 2010
Public Hearing & Meeting
Research:

June 15, 2010
Public Hearing
Material Viewing:
June 11th

June 01, 2010
Public Hearing
Material Viewing:
Thursday
May 27th

May 18, 2010
Public Hearing
Material Viewing:
May 14th

May 11, 2010
Public Meeting

May 04, 2010
Public Hearing
Material Viewing:
April 30th
Use case: archive state agency websites and publications

Includes pages in a variety of formats: text, images, audio, video and social networking sites
Access to Public Collections

Partners:
- Can view through private web application with login/password

General Public:
- Landing Pages: view from organization’s website with a branded page that links back to Archive-It hosted data
- Integration of data and metadata with existing systems and catalogs
Storage & Preservation

Multiple ways to Store and Preserve

Storage:
- 2 copies of the archived data (primary and back-up) are stored at San Francisco Data Center
- Collections transferred to the General Archive as a third copy
- A copy of archived data can be shipped on a hard drive
- Ability to download files from Internet Archive servers

Digital Preservation:
- 2008: LOCKSS
- 2013: Duracloud
- 2015?: Archivematica
Data Repository
Learn More

http://www.archive-it.org/learn-more

Download our white paper on the web archiving life cycle and meet some of the team members

Follow us on Twitter: @archiveitorg

Like us on Facebook:
https://www.facebook.com/ArchiveIt

Check out our blog: archive-it.org/blog
Thank you!

Maria LaCalle

maria@archive.org