



# Archiving and Preserving the Web

**Kristine Hanna**

**Internet Archive**

**November 2006**



# About Internet Archive

- Non profit founded in 1996 by Brewster Kahle, as an ‘Internet library’
- Provide universal and permanent access to digital information for researchers, historians, scholars and the general public
- Built on open source principles and dedicated to Open Source software
- Largest publicly available (free!) web archive at [www.archive.org](http://www.archive.org)



# What do we collect?

## General Web Archive

- Take a broad snapshot of the web every 2 months
- 2 billion pages a snapshot (approx. 20 pages per site)
- Current archive is 55 billion pages
- Websites from every domain and content in 21 languages
- Stored on 2300 machines (servers)
- Servers in U.S. Egypt, France, Amsterdam
- Since 2002, Archive has collected texts, audio, moving images, and software (30,000 to 90,000 of each media)



# How do we collect it?

Open Source Technology  
primarily developed by Internet Archive and IIPC

- **Heritrix**: web crawler
- **Wayback Machine**: access tool for rendering and viewing files. Displays archived web pages--surf the web as it was.
- **NutchWAX**: Bundling of Nutch, open source search engine. Standard full-text search
- **Arc File**: archival file format used for preservation



# Web Archiving for Partners

## **What we found out: institutions needed more control and access over content.**

- Create focused collections.
- Crawl specific websites
- Decide when to crawl
- Monitor crawls, with post crawl reports
- Full text search of collections
- Hosted content



# Web Archiving Services

- Curated Crawls
- Domain Crawls
- Archive-It



# Curated/Domain Crawls

**Designed for large institutions (National libraries and archives)**

## **1. Curated Crawls**

- Large topic collections (100+ million documents)
- **On-going** crawls run by IA crawl engineer

## **2. Domain Crawls**

- Large domain specific collections (250+ million documents)
- **One time** crawls run by IA crawl engineer

### **Sample Collections:**

- National elections – since 2000 (**Library of Congress**)
- Iraq War (**Library of Congress**)
- 2006 congressional election (**U.S. National Archives**)
- .fr domain (**France**), .au domain (**Australia**)



# Archive-It

**Designed for smaller institutions (state archives, state libraries and university libraries)**

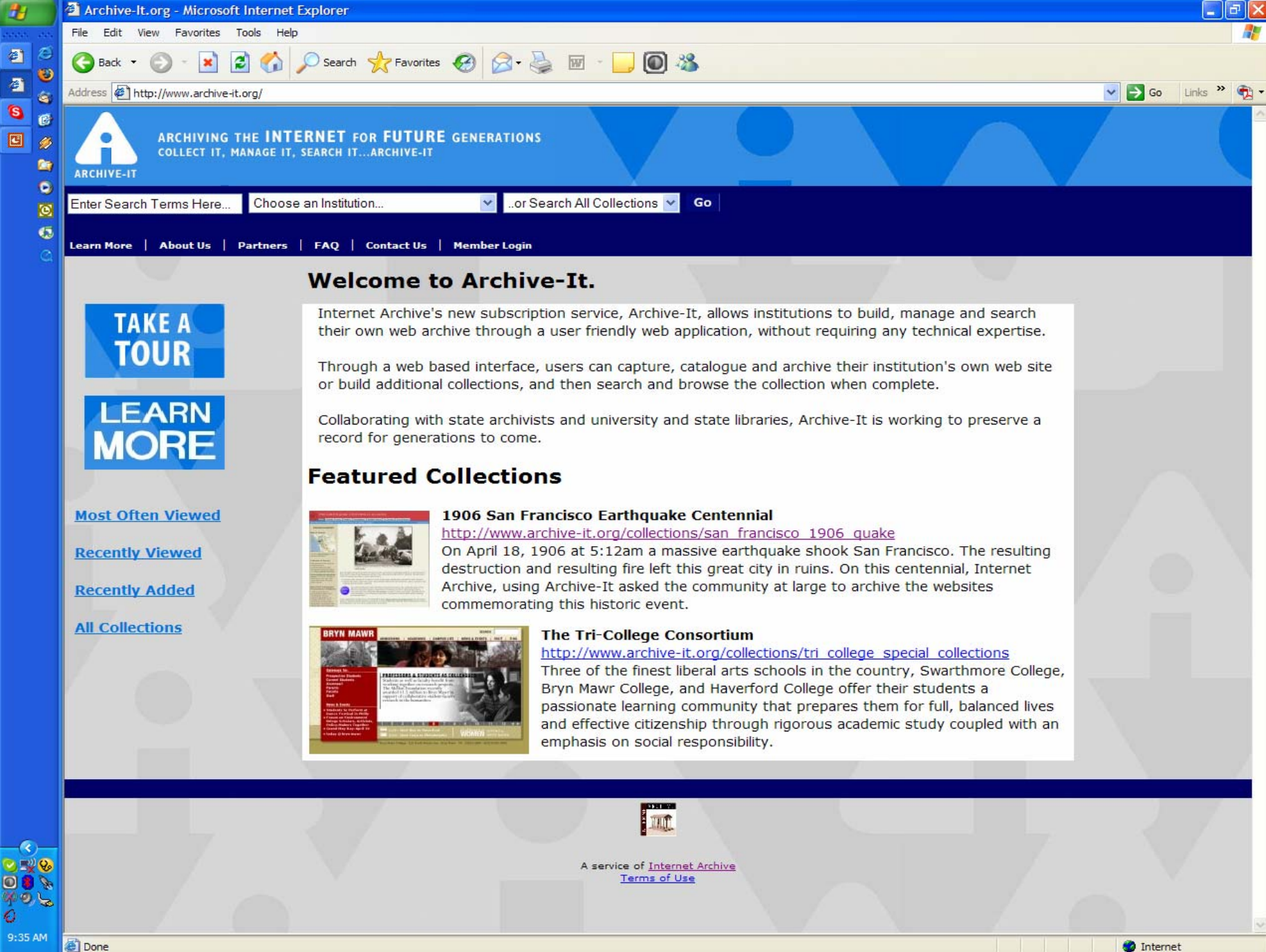
- Web based application that allows users to create, manage and search their web archives
- Annual subscription service, 3 collections and up to 10 million pages
- Functions include: harvesting, cataloging, managing and analysis of collections + full text search





# Archive-It

- Pilot was September – November 2005
- **1.0** release in February
- **1.5** release in April
- **2.0** release in July





# Archive-It features

- Admin + 10 users
- Customized crawl frequency controls
- Test Crawls
- Dublin Core Metadata fields
- XML Feed of seed and collection metadata
- Advanced full text search: by relevancy, by date and by metadata
- Scope extension (sub domains etc)
- Crawl constraints (limit # of documents per host or crawl)
- Post crawl Reports/Analysis
- Templates for collection home pages
- Online Help Section/ User Manual



# How Partners use Archive-It

- **University of Texas Libraries**
  - Internet directory/portal specializing in delivering filtered, organized, content-rich information about Latin America.

English

Español

Português

University of Texas Libraries



## Latin American Government Documents Archive

### Presidential Message in the Spotlight

#### MENSAJE PRESIDENCIAL 2005



- Discurso Co
- Compromis
- Leyes Apro

[Discurso en](#)

[Full text](#) of Chilean President Ricardo Lagos' annual Presidential Message, May 21, 2005

### About the Archive

The *Latin American Government Documents Archive* (LAGDA) seeks to preserve and facilitate access to a wide range of ministerial and presidential documents from 18 Latin American and Caribbean countries. The Archive contains copies of the Web sites of approximately 300 government ministries and presidencies. Capture of sites began on multiple dates in 2005 and 2006, and will continue with regularly scheduled captures.

Content in the Archive includes not only the full-text versions of official documents, but also original video and audio recordings of key regional leaders. Archive contents include thousands of annual and "state of the nation" reports; plans and programs; and speeches by presidents and government ministers. Content can be accessed via full-text search ([search help](#)), or by browsing by country or by specialized sample collection, such as "Presidential Messages" or "Ministerial Documents."

LAGDA is a joint project of the University of Texas Libraries, The Nettie Lee Benson Latin American Collection, and the Latin American Network Information Center at The University of Texas at Austin. Web archiving services are provided by the Internet Archive's [Archive-It](#) service.

Home

Browse Full Collection

Presidential Messages

Ministerial Documents

Related Collections

Search LAGDA

Go



# How Partners use Archive-It

- **Library of Virginia**

- serves as the Commonwealth's archival agency, the reference library at the seat of government and the research institution for Virginia history, politics and culture.
- Governor Mark Warner collection
- Jamestown 2007 - 400th anniversary of the first permanent English settlement
- Virginia State Government, Legislative Branch,



**What We Have**[Who We Are](#)[What We Do](#)**Archiving the Web: Virginia's Political Landscape, Fall 2005**

In the fall of 2005, the Library of Virginia was one of several cultural heritage institutions that participated in a pilot project with the [Internet Archive](#) to develop and refine a tool to collect, preserve, and provide access to Web sites that meet institutional collection policies and are considered to be of enduring value. This tool is called [Archive-It](#).

Spurred on by our mission to preserve and make accessible in perpetuity the Web heritage of each Virginia governor, the Library as part of its pilot project archived the Web-based materials for the administration of Governor Mark R. Warner. The Library also expanded its collection parameters for the project to include the campaign Web sites related to Virginia's statewide elections taking place in the fall of 2005. This included the sites of the candidates for Governor, Lieutenant Governor, and Attorney General, as well as related political party sites and several political blogs. Additional collections highlighting two statewide projects, the Jamestown 2007 observance and the renovation of Capitol Square in Richmond, are also available.

You can search our collections as a whole or browse specific archived Web site collections by URL. Please see [Search Help](#) for more detailed assistance.

Search the Library's Web archive of Virginia's Political Landscape, Fall 2005, here:

Or browse these collections:

**Things to Know**[Search Help](#)

Technical Facts  
about Web  
archiving (coming  
soon)

**Our Experts  
Suggest**[Archive-It  
homepage](#)[The Internet  
Archive](#)[MINERVA at the  
Library of Congress](#)[Web Harvest 2004  
\(NARA\)](#)



# How Partners use Archive-It

- **North Carolina State Archives**

- portal page includes all North Carolina government web sites. Government Agencies, Occupational Licensing Boards, and Commissions.
- Provides users a place to research their state agencies.



lobe

y X

, >>

W...

W...

ast ...

oday



# North Carolina State Government Web Site Archives

North Carolina State Archives State Library of North Carolina

Search the Web Archives



[Search Help](#)

## Welcome!

### Links

- Home
- About Us
- Browse by Agency
- Browse by Collection
- Site Map
- Help
- Contact

### Other Sites

- Archive-It
- Internet Archive

The *North Carolina State Government Web Site Archives* allows you to view North Carolina state agency web sites from past dates. The *Web Site Archives* contains web sites from the Fall of 2005, and from April 2006 forward allowing free and open access to this information long after the sites have changed on the live web.

The *Web Site Archives* can be searched via the search box on the top of every page on this site. For tips and helpful hints on searching, read our [Search Help](#) document. Users may also browse for web sites by [State Agency](#) or by [Collection](#).

The *Web Site Archives* began as a pilot project with the [Internet Archive](#) during the Fall of 2005. The purpose of this project was to refine a tool called [Archive-It](#) which collects, preserves, and provides access to web sites of enduring value.

The success of the Pilot Project led to the creation of the current version of *The North Carolina State Government Web Site Archives* which began archiving web sites in late April of 2006. The *Web Site Archives* contains copies of state agency web sites captured during the pilot project in the Fall of 2005 and after the official launch in April 2006 forward.



Above: Screenshot of the Governor's web site



# How Partners use Archive-It

- **University of Indiana**

- Responsible for the appraisal, acquisition, preservation and use of University records of permanent value
- archives the records of individuals and organizations associated with Indiana University



# Indiana University Digital Library Program

search the dlp website

SEARCH

[Home](#) [About Us](#) [Collections & Resources](#) [Research & Development](#) [Services](#) [Education & Outreach](#) [Employment](#)

## Collections & Resources

[home](#) > collections & resourcesSee also: [Digital Collections of IUPUI University Library](#)

### [The Algernon Charles Swinburne Project](#)

The Swinburne Project contains electronic editions of the works of prolific Victorian poet and critic Algernon Charles Swinburne (1837-1909). The collection includes both poetry and criticism.

### [Bloomington Faculty Council](#)

The Indiana University Bloomington Faculty Council collection will contain meeting notes made available to end-users through a fully functional website.

### [Board of Trustees](#)

The Indiana University Board of Trustees collection contains meeting notes from 1981 through 2001 will be made available to end-users through a fully functional website.





### [Central American and Mexican Video Archives \(CAMVA\) \(in progress\)](#)

The Central American and Mexican Video Archives (CAMVA) collection consists of historical records from El Salvador, Nicaragua and Mexico from 1970 through 1999. More than 200 hours of video, audio, and photographic digital materials from these countries histories are used for instructional purposes.

### [Charles W. Cushman Photograph Collection](#)

The Charles W. Cushman Photograph Collection contains almost 15,000 Kodachrome slides by amateur photographer Charles W. Cushman. Mr. Cushman lived from 1896-1972 and spent a good portion of his life photographing his travels in the United States and abroad.

### [The Chymistry of Isaac Newton](#)

-  = Text Collection
-  = Image Collection
-  = Audio Collection
-  = Video Collection



# Archive-It Partners & Trial Partners

- University of Texas
- University Southern California
- Utah State Library and Archives
- University of Hawaii
- Tennessee State Library and Archives
- South Dakota State Library
- State Archives of North Carolina
- Nebraska State Historical Society
- State Archives of South Carolina
- Institut d'Etudes Politiques de Grenoble
- Univ. of North Carolina, Chapel Hill
- Pennsylvania Historical Commission
- Texas State Library and Archives Commission
- University of Toronto
- University of Minnesota
- Library of Virginia
- New York State Archives
- Michigan Historical Center
- State Archives of Alabama
- Indiana University
- California Digital Library
- Library of Congress
- Tri College Consortium



# Archive-It moving forward

- **Current stats for Archive-It 2.0**
  - 23 partners (and growing!)
  - 65 collections
  - 216 million documents
- **Next steps:**
  - Enhanced application releases scheduled for: February 8, April 13 and June 14
  - Continued Partner feedback



# What's Next for Web Archive

- **Collaboration and Partnerships**
  - Technology partner in providing web archiving services
  - Develop partnerships with like minded organizations (CDL, OCLC etc)
  - Develop Open Source software
  - Develop common tools, storage formats and standards through the IIPC, and with our partners
- **Multiple copies around the world**
  - Within IA's own repository, and with partners such as LC, Bnf, Library of Alexandria



# Questions?

**Kristine Hanna**

**Director, Web Archiving Services**

**[kristine@archive.org](mailto:kristine@archive.org)**