Archiving “Katrina”

Lessons Learned

Kris Carpenter Negulescu, Director
Gordon Mohr, Chief Technologist

The Internet Archive, Web Group

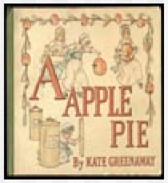
Agenda

- Introductions
- Project Overview
- Lessons Learned
  - What Worked
  - Challenges & Limitations
- Technology Review
- Future Recommendations

What is the Internet Archive?

A digital library of about 3 petabytes of information, including

- Web Pages
- Educational Courseware
- Films & Videos
- Music & Spoken Word
- Books & Texts
- Software

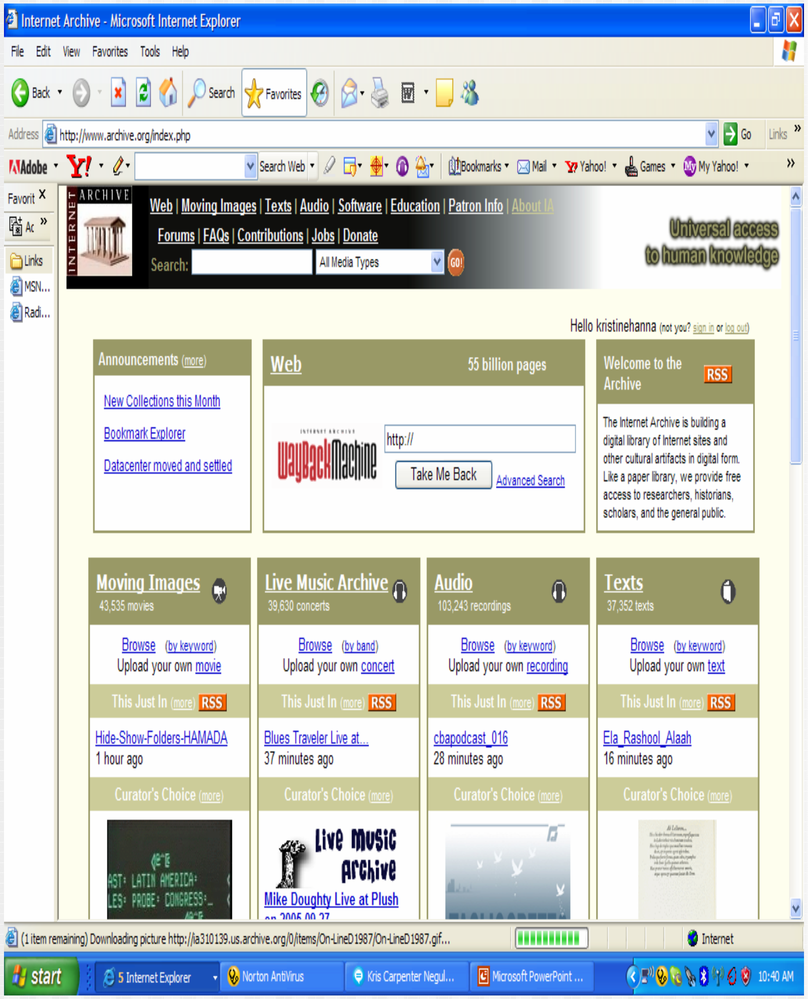


The archive's combined collections receive 6 million downloads a day

How is IA Content Collected?

- **Bi-monthly snapshots of www harvested by Alexa Internet**
- **Web harvests by the Internet Archive**
- **Contributions from the community**
- **Scanning & digitization of public domain texts, stills, moving images, etc.**

Entire collection accessible for free to the public at www.archive.org


Internet Archive - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.archive.org/index.php Go Links
Adobe Search Web Bookmarks Mail Yahoo! Games My Yahoo!
Favorit Ac Links MSN... Radi...
INTERNET ARCHIVE
Web | Moving Images | Texts | Audio | Software | Education | Patron Info | About IA
Forums | FAQs | Contributions | Jobs | Donate
Search: All Media Types GO!
Universal access to human knowledge
start 5 Internet Explorer Norton AntiVirus Kris Carpenter Negul... Microsoft PowerPoint ... 10:40 AM

Hello kristinehanna (not you? sign in or log out)

Announcements (more)

New Collections this Month

Bookmark Explorer

Datacenter moved and settled

Web 55 billion pages

WayBackMachine http:// Take Me Back Advanced Search

Welcome to the Archive RSS

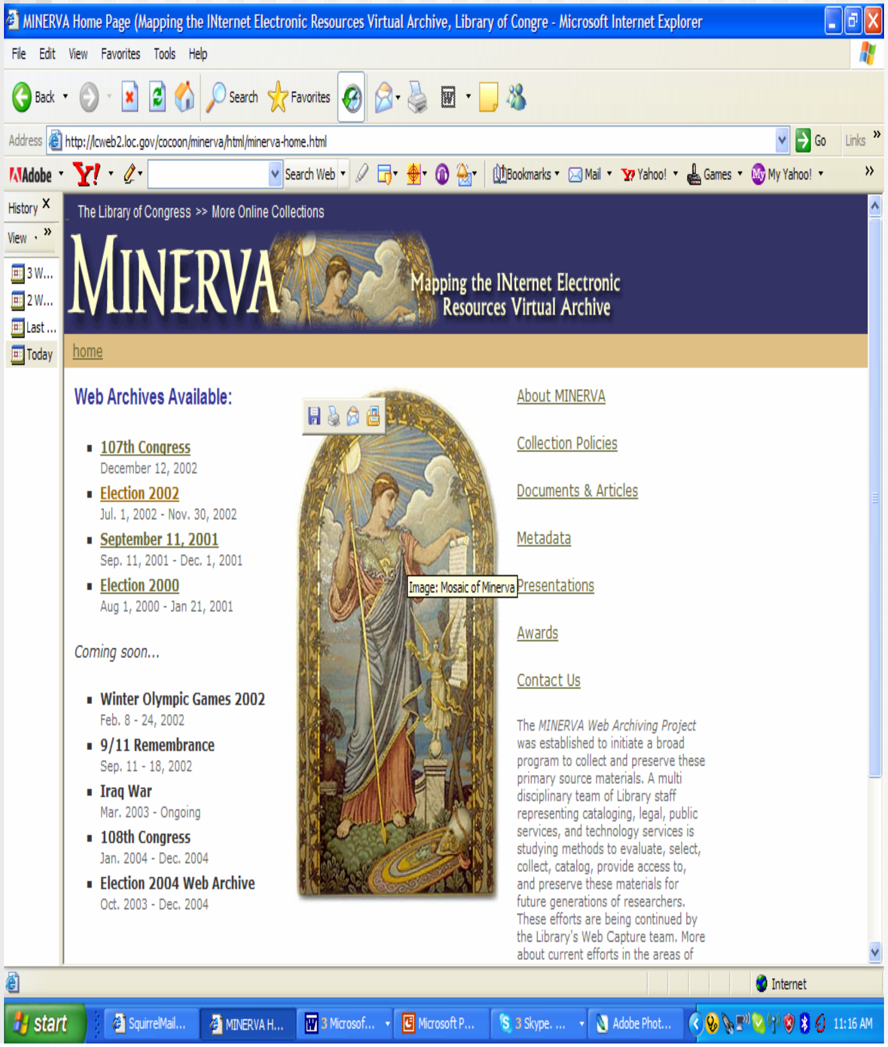
The Internet Archive is building a digital library of Internet sites and other cultural artifacts in digital form. Like a paper library, we provide free access to researchers, historians, scholars, and the general public.

Moving Images 43,535 movies	Live Music Archive 39,630 concerts	Audio 103,243 recordings	Texts 37,352 texts
Browse (by keyword) Upload your own movie	Browse (by band) Upload your own concert	Browse (by keyword) Upload your own recording	Browse (by keyword) Upload your own text
This Just In (more) RSS	This Just In (more) RSS	This Just In (more) RSS	This Just In (more) RSS
Hide-Show-Folders-HAMADA 1 hour ago	Blues Traveler Live at... 37 minutes ago	cbapodcast_016 28 minutes ago	Ela_Rashool_Alaah 16 minutes ago
Curator's Choice (more)	Curator's Choice (more)	Curator's Choice (more)	Curator's Choice (more)
	Live Music Archive Mike Doughty Live at Plush		

(1 item remaining) Downloading picture http://ia310139.us.archive.org/0/items/On-LineD1987/On-LineD1987.gif... Internet

Web Archiving Services

INTERNET ARCHIVE


MINERVA Home Page (Mapping the INternet Electronic Resources Virtual Archive, Library of Congre - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://lcweb2.loc.gov/cocoon/minerva/html/minerva-home.html
Go Links
Adobe Search Web Bookmarks Mail Yahoo! Games My Yahoo!
History View
3 W...
2 W...
Last ...
Today
Image: Mosaic of Minerva
Internet
start SquirrelMail... MINERVA H... 3 Microsof... Microsoft P... 3 Skype. ... Adobe Phot... 11:16 AM

The Library of Congress >> More Online Collections

MINERVA Mapping the INternet Electronic Resources Virtual Archive

home

Web Archives Available:

- 107th Congress
  December 12, 2002
- Election 2002
  Jul. 1, 2002 - Nov. 30, 2002
- September 11, 2001
  Sep. 11, 2001 - Dec. 1, 2001
- Election 2000
  Aug 1, 2000 - Jan 21, 2001

Coming soon...

- Winter Olympic Games 2002
  Feb. 8 - 24, 2002
- 9/11 Remembrance
  Sep. 11 - 18, 2002
- Iraq War
  Mar. 2003 - Ongoing
- 108th Congress
  Jan. 2004 - Dec. 2004
- Election 2004 Web Archive
  Oct. 2003 - Dec. 2004

About MINERVA

Collection Policies

Documents & Articles

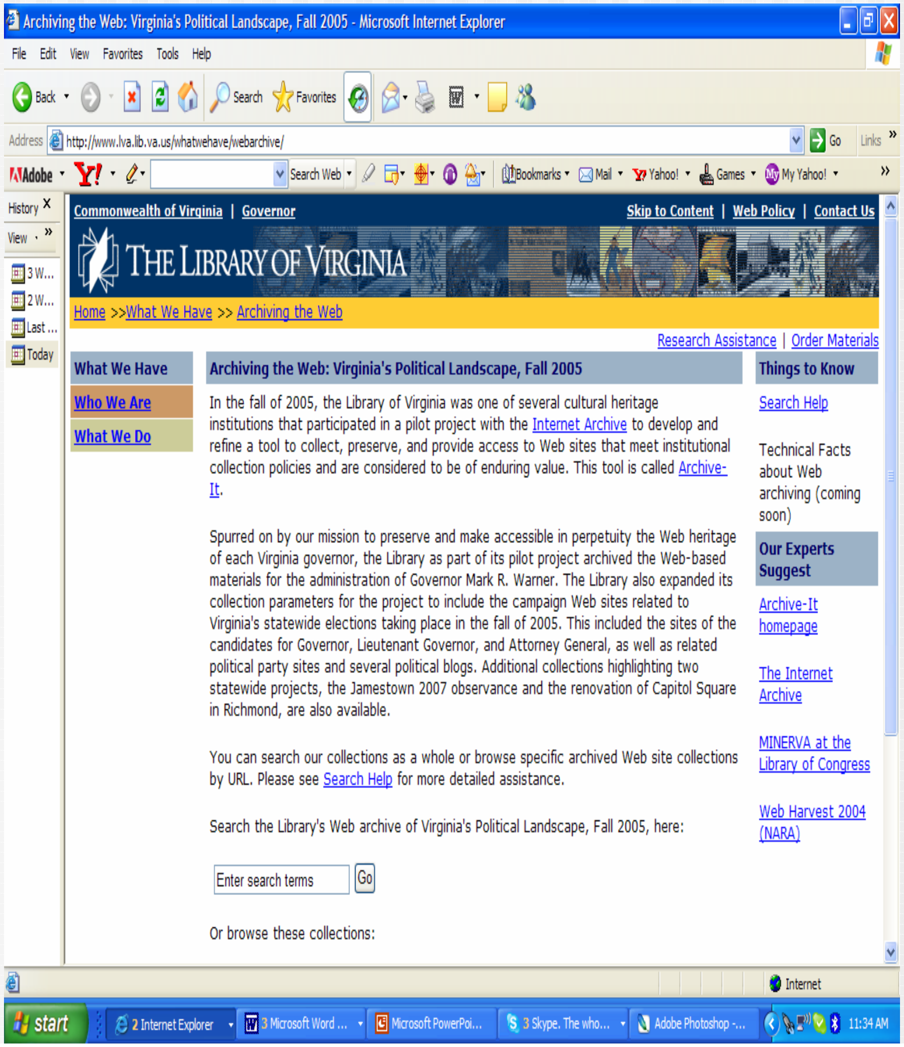
Metadata

Presentations

Awards

Contact Us

The *MINERVA Web Archiving Project* was established to initiate a broad program to collect and preserve these primary source materials. A multi disciplinary team of Library staff representing cataloging, legal, public services, and technology services is studying methods to evaluate, select, collect, catalog, provide access to, and preserve these materials for future generations of researchers. These efforts are being continued by the Library's Web Capture team. More about current efforts in the areas of


Archiving the Web: Virginia's Political Landscape, Fall 2005 - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.lva.lib.va.us/whatwehave/webarchive/
Go Links
Adobe Search Web Bookmarks Mail Yahoo! Games My Yahoo!
History View
3 W...
2 W...
Last ...
Today
Internet
start 2 Internet Explorer 3 Microsoft Word ... Microsoft PowerPoi... 3 Skype. The who... Adobe Photoshop -... 11:34 AM

Commonwealth of Virginia | Governor   Skip to Content | Web Policy | Contact Us

THE LIBRARY OF VIRGINIA

Home >>What We Have >> Archiving the Web

Research Assistance | Order Materials

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:

Enter search terms [Go]

Or browse these collections:

Things to Know

Search Help

Technical Facts about Web archiving (coming soon)

Our Experts Suggest

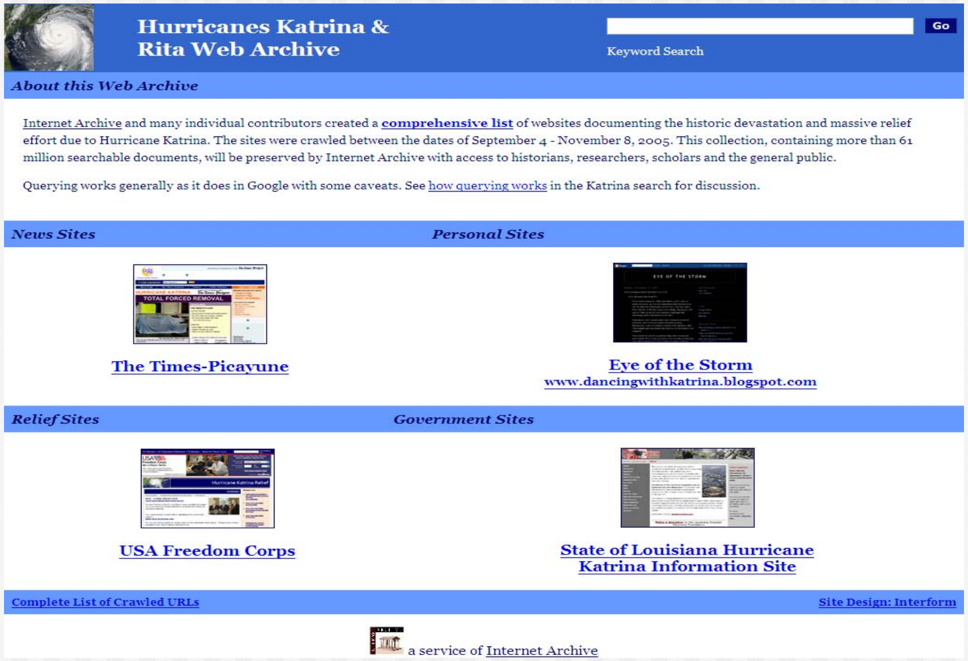
Archive-It homepage

The Internet Archive

MINERVA at the Library of Congress

Web Harvest 2004 (NARA)

Project Overview



Hurricanes Katrina & Rita Web Archive

About this Web Archive

Internet Archive and many individual contributors created a **comprehensive list** of websites documenting the historic devastation and massive relief effort due to Hurricane Katrina. The sites were crawled between the dates of September 4 - November 8, 2005. This collection, containing more than 61 million searchable documents, will be preserved by Internet Archive with access to historians, researchers, scholars and the general public.

Querying works generally as it does in Google with some caveats. See how querying works in the Katrina search for discussion.

News Sites

The Times-Picayune

Personal Sites

Eye of the Storm
www.dancingwithkatrina.blogspot.com

Relief Sites

USA Freedom Corps

Government Sites

State of Louisiana Hurricane Katrina Information Site

Complete List of Crawled URLs

Site Design: Interform

a service of Internet Archive

The ‘Katrina’ Collection

The Goal

- To create an historical record of the devastation and the massive relief effort that followed as it was documented on the Web.

Scope & Timing

- Web content was harvested between September 4 and November 8, 2005
- Collection includes over 61 million unique Web captures, all text searchable, from over 1700 Web sites.

Key Contributors:

- Library of Congress
- Indiana University, CDL/LSU, University California at Berkeley, and others
- Individuals

Available at: http://websearch.archive.org/katrina/

The Katrina Collection - Tools

- **Heritrix**: a web crawler

  http://crawler.archive.org/

- **Wayback Machine**: an address-based access tool used to locate and view archived web pages

  http://archive-access.sourceforge.net/projects/wayback/

- **NutchWAX**: tools for full text search of archival web content.

  http://archive-access.sourceforge.net/projects/nutch/

NutchWax is an extension of other open source projects:

- **Nutch:** open source web search engine software

  http://lucene.apache.org/nutch/

- **Hadoop:** a framework for running applications on large clusters of commodity hardware.

  http://lucene.apache.org/hadoop/

Lessons Learned

- Back-Drop
- Defining the Collection
- Harvest & Access

Back-Drop

- Tight timeline
- No formal submission tools
- No definition of desired collection scope
- No process for recruiting contributors
- No pre-established “permissions to crawl”

Defining the Collection

What Worked

Acquiring Partners & Permissions

- Able to collect seeds rapidly from a broad range of sources
- Received permission to crawl national and regional press and news sites

What Worked

Collection Scope & Quality

Essential contributions came from LoC curators and subject experts at research universities

- 450+ seeds (~300 from LoC, 100+ from CDL/LSU, and others)
- A subset of the seeds were so unique that they would not have been found via a search engine or submitted by the general public

Challenges/Limitations

Acquiring Partners & Permissions

- Requests to contribute were distributed ad hoc, artificially limiting participation
- Partners had difficulty convincing experts to drop current work
- More difficult than usual to reach web site operators to acquire permissions

Challenges/Limitations

Collection Scope & Quality

- Sans permissions, forced to respect robots.txt
- Curatorial partners and the public submitted blind
- Given time constraints cast “a very big net”
  - Broad expansion of seeds using online directories, blog aggregators, and search engines returned mixed results

Harvest & Access

What Worked

Flexible Action

- Allocated hardware and engineering resources on short notice to enable timely capture, preservation and access to the collection
- Helped re-establish some Web presences using captures from the IA general archive harvested prior to the event

What Worked

Open Source Solutions

- Used Heritrix for rapid, continuous harvests during a six week period
- Automated the tracking and capture of seeds
  - some prior to disappearance of a site; and then when/if it re-appeared
- Nutch/NutchWax enabled full text search
  - 61 million unique captures harvested from 1700 sites

Challenges/Limitations

Scalability of Harvesting Tools

- Improved distributed crawl capabilities would have increased the speed, efficiency and effectiveness of crawls

Prioritized/Parameterized Crawling

- “Smart” content discovery
  - Adaptive revisits and/or selective capture, automated link relationship analysis, geographic clustering, topic based clustering &/or content relationships defined by curatorial resources, etc.

Challenges/Limitations

Search & Analysis

- Duplicate management, indexing techniques not yet fully optimized for archived Web content
- Absence of meta data
- Limited options for viewing and analyzing contents

Technology Review

- Heritrix
- Wayback Machine
- NutchWAX

  with Nutch, Hadoop

3 Tasks, 3 Tools

- Collect
  - Heritrix Web Crawler
- Redisplay/Browse
  - Wayback Machine Web Archive Viewer
- Search
  - NutchWAX Web Archive Search Engine

Heritrix

“Open source, Extensible, Web-scale, Archival Quality Web Crawling Software”

http://crawler.archive.org

- Collaboratively developed
  - Internet Archive, IIPC, partner libraries, and others
  - First release (0.2) in January 2004; 10 since
- Katrina crawls used version 1.4
  - Latest version: 1.10, released September 2006
- *‘Heritrix’* means? ***woman who inherits (heiress)***

Heritrix Goals

Heritrix was designed to…

- Crawl for completeness/depth
- Be highly configurable – especially in collection scope
- Offer a web-based control interface
- Respect the robots.txt exclusion directives & META robots tags
- Collect material at a measured, adaptive pace, not to disrupt ordinary web site usage

…within the parameters…

- Available under an open source license (LGPL)
- Use Java, leveraging existing open source libraries
- Scale to 100's of millions to a billion documents
- Run well on Linux

  (but also work wherever Java is available)

Heritrix

Current Release: 1.10.1

Notable Additions & Changes Since Katrina

- Ability to distribute a crawl across multiple independent crawlers
- Per-host/domain/queue-grouping collection quotas
- Improved performance and stability in large crawls
- De-duplication add-on
- Expanded crawl configuration options

Planned Enhancements

- “Smart” crawling enhancements
- Scaling to ongoing, 1billion+ URL crawls

Wayback Machine

http://archive-access.sourceforge.net/projects/wayback/

The Wayback Machine is an open source (LGPL) web archive viewing application in Java designed to

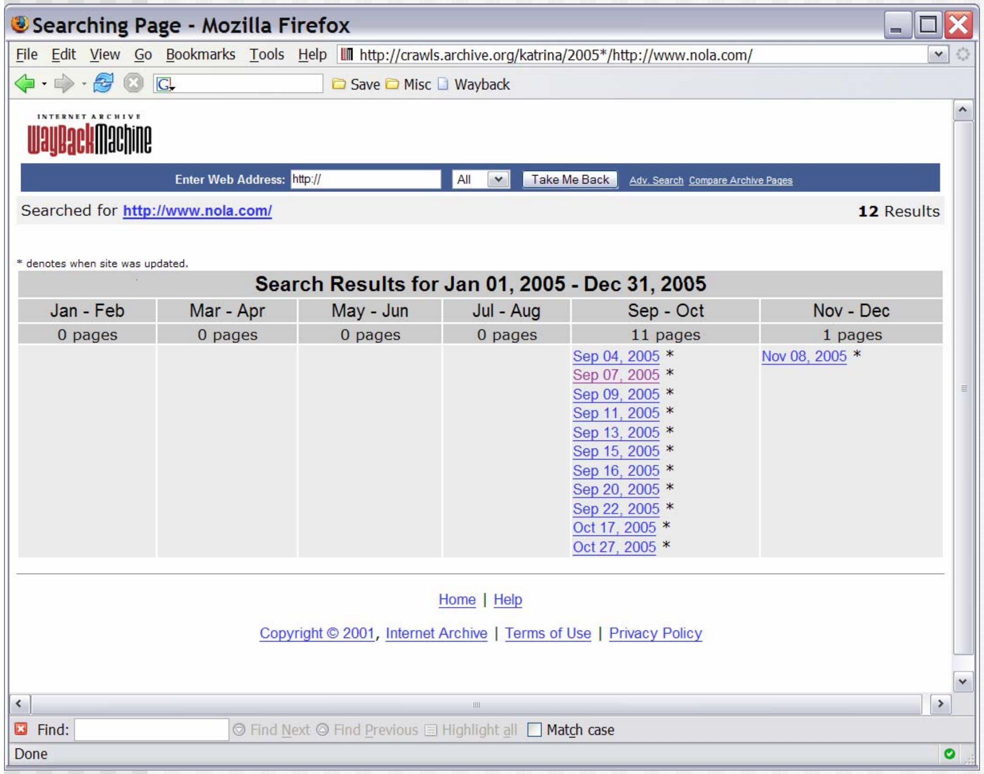
- Display lists of available captures by date
- Allow browsing ‘as it was’ in a natural manner
- Offer multiple UI modes and index options for different classes of users and deployment sizes

Current Release: v0.6.0

Notable Additions & Changes Since Katrina

- First open source release in December 2005
- Timeline and proxy modes

Wayback Machine



Searching Page - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://crawls.archive.org/katrina/2005*/http://www.nola.com/
Save Misc Wayback
INTERNET ARCHIVE
WaybackMachine
Enter Web Address: http://
All
Take Me Back
Adv. Search
Compare Archive Pages

Searched for http://www.nola.com/ **12** Results

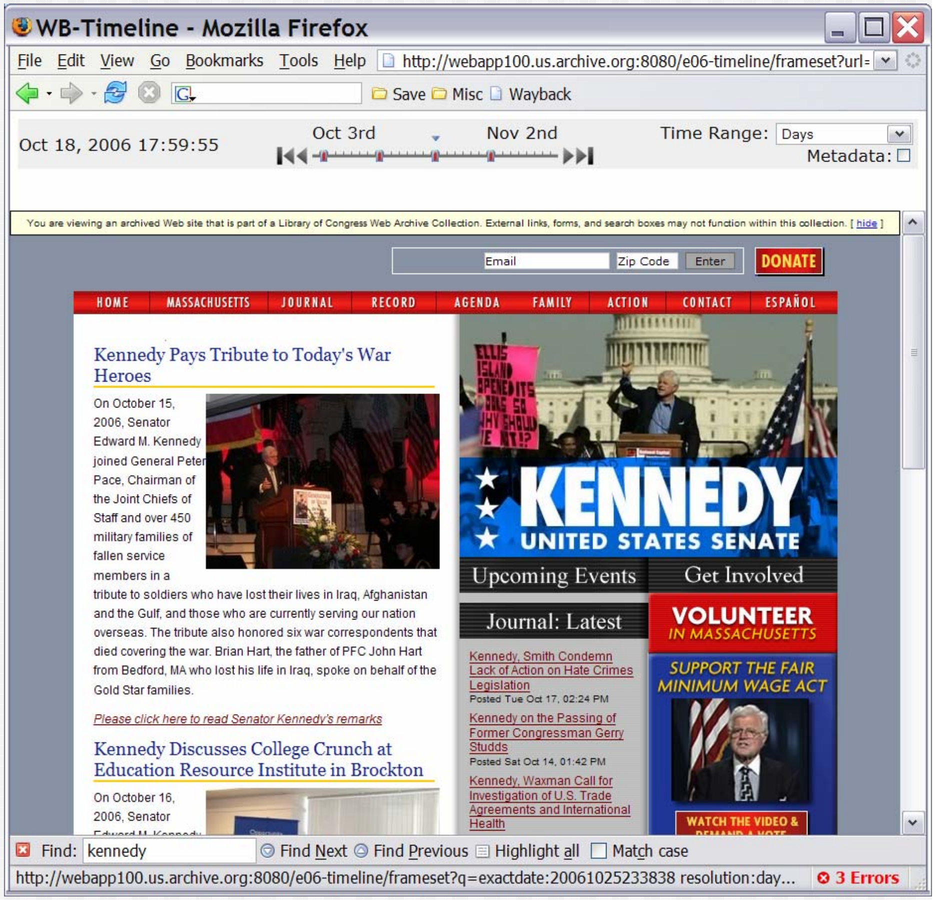
* denotes when site was updated.

Search Results for Jan 01, 2005 - Dec 31, 2005

Jan - Feb	Mar - Apr	May - Jun	Jul - Aug	Sep - Oct	Nov - Dec
0 pages	0 pages	0 pages	0 pages	11 pages	1 pages
				Sep 04, 2005 * Sep 07, 2005 * Sep 09, 2005 * Sep 11, 2005 * Sep 13, 2005 * Sep 15, 2005 * Sep 16, 2005 * Sep 20, 2005 * Sep 22, 2005 * Oct 17, 2005 * Oct 27, 2005 *	Nov 08, 2005 *

Home | Help

Copyright © 2001, Internet Archive | Terms of Use | Privacy Policy



WB-Timeline - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://webapp100.us.archive.org:8080/e06-timeline/frameset?url=
Save Misc Wayback

Oct 18, 2006 17:59:55

Time Range: Days

Metadata: □

You are viewing an archived Web site that is part of a Library of Congress Web Archive Collection. External links, forms, and search boxes may not function within this collection. [hide]

Kennedy Pays Tribute to Today's War Heroes

On October 15, 2006, Senator Edward M. Kennedy joined General Peter Pace, Chairman of the Joint Chiefs of Staff and over 450 military families of fallen service members in a tribute to soldiers who have lost their lives in Iraq, Afghanistan and the Gulf, and those who are currently serving our nation overseas. The tribute also honored six war correspondents that died covering the war. Brian Hart, the father of PFC John Hart from Bedford, MA who lost his life in Iraq, spoke on behalf of the Gold Star families.

Please click here to read Senator Kennedy's remarks

Kennedy Discusses College Crunch at Education Resource Institute in Brockton

On October 16, 2006, Senator Edward M. Kennedy

Kennedy, Smith Condemn Lack of Action on Hate Crimes Legislation
Posted Tue Oct 17, 02:24 PM

Kennedy on the Passing of Former Congressman Gerry Studds
Posted Sat Oct 14, 01:42 PM

Kennedy, Waxman Call for Investigation of U.S. Trade Agreements and International Health

Find: kennedy

http://webapp100.us.archive.org:8080/e06-timeline/frameset?q=exactdate:20061025233838 resolution:day...

3 Errors

NutchWAX, Nutch, Hadoop

http://archive-access.sourceforge.net/projects/nutch/

NutchWAX includes tools to search Web Archive Collections (WACs). NutchWAX is built on two other open source projects:

- **Nutch**, web-search software based on Lucene Java adding Web specifics such as link-graph analysis, parsers for HTML and other formats, and a batch-oriented crawler.
- **Hadoop**, a framework for running applications (such as the indexing of large scale web content) on clusters of commodity hardware. Hadoop…
  - Implements a computational paradigm named map/reduce
  - Provides a reliable distributed file system on the compute nodes
  - Has been tested on clusters of 600 nodes, but there are reports of support for 900+ nodes for both map/reduce and file system storage.
  - IA's largest cluster for full-text indexing has been 34 nodes.

NutchWAX

NutchWAX bundles Nutch/Hadoop with extensions for working with archived content, including:

- Adapted Nutch fetcher step to go against archives rather than open net
- Index-time and query-time plug-ins to allow querying of a records' location in a repository
- Awareness of capture-times and multiple captures per URL
- Sending searchers/browsers to an associated Wayback Machine

Current Release: **v0.6.0**

Notable Additions & Changes Since Katrina

- Moved to a map/reduce version of Nutch in May 2006

Planned Enhancements

- Improved performance and stability for indexing jobs of 100+ million captures

Recommendations

For Spontaneous, Event-Based, Web Harvests

Plan Ahead

Establish a broad network of expert contributors & create a communication plan

- Encourage expert participation from research institutions, universities, museums, historical societies, archives, libraries, etc.
  - Secure commitments to participate, where feasible
- Provide incentives for involvement
- Clarify responsibilities for communication & procedures for initiating action

Plan Ahead

Brainstorm scenarios for event-based, Web harvests

Define criteria that trigger action and standard criteria for collection “quality”

- E.g. Wilma vs. Katrina?, epidemic or isolated threat?, etc.

Create flexible resource pools (hardware, bandwidth, harvesting tools, people, etc.)

Facilitate
Contribution & Collaboration

Use Wiki's to organize participation, share information, encourage interaction amongst contributors

Define desired collection "quality" & provide specific guidelines for submission

Facilitate
Contribution & Collaboration

Notify contributors of duplicate seeds as they attempt to submit

Make seed list/s, crawl status, prioritization, scope and frequency transparent to all contributors

Enable experts and curators to edit/expand seed lists, crawl scope, prioritization and frequency

Thank You!

Kris Carpenter Negulescu, Director

- kcarpenter@archive.org

Gordon Mohr, Chief Technologist

- gojomo@archive.org

The Internet Archive, Web Group