



Web and Twitter Archiving at the Library of Congress

Web Archive Globalization Workshop
June 16, 2011

Nicholas Taylor ([@nullhandle](https://twitter.com/nullhandle))

Web Archiving Team

Library of Congress

why archive the web?

- preserve our nation's history and culture
- identify and preserve at-risk digital content
- develop of tools, models, and methods for digital preservation



“World Wide Web 1997: 2 Terabytes in 63 Inches”

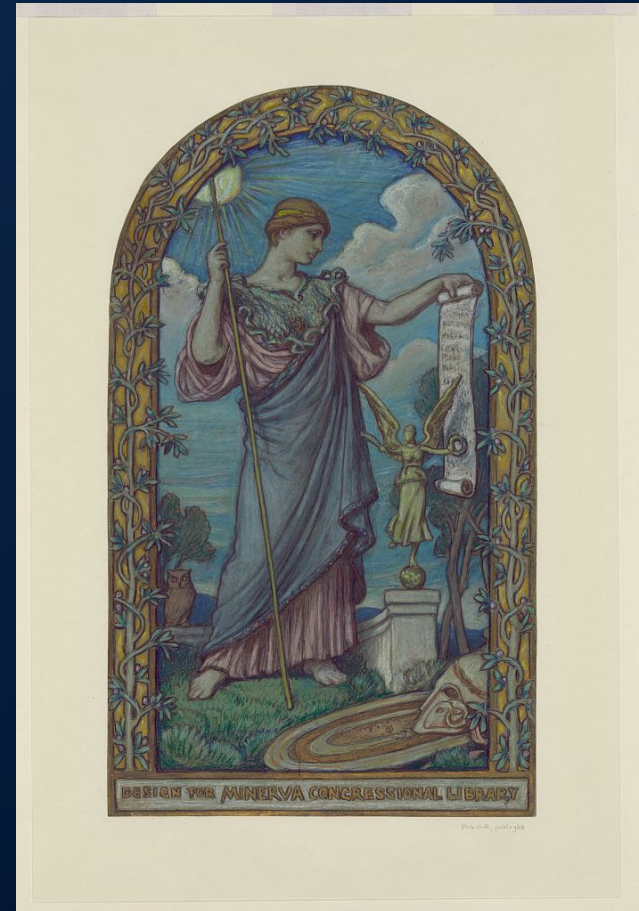
various collection strategies

- entire web domain—Internet Archive
- national domain—Sweden, Denmark, others
- selective (individual URLs) and thematic—Australia
- thematic or event-based—Library of Congress

<http://netpreserve.org/about/archiveList.php>

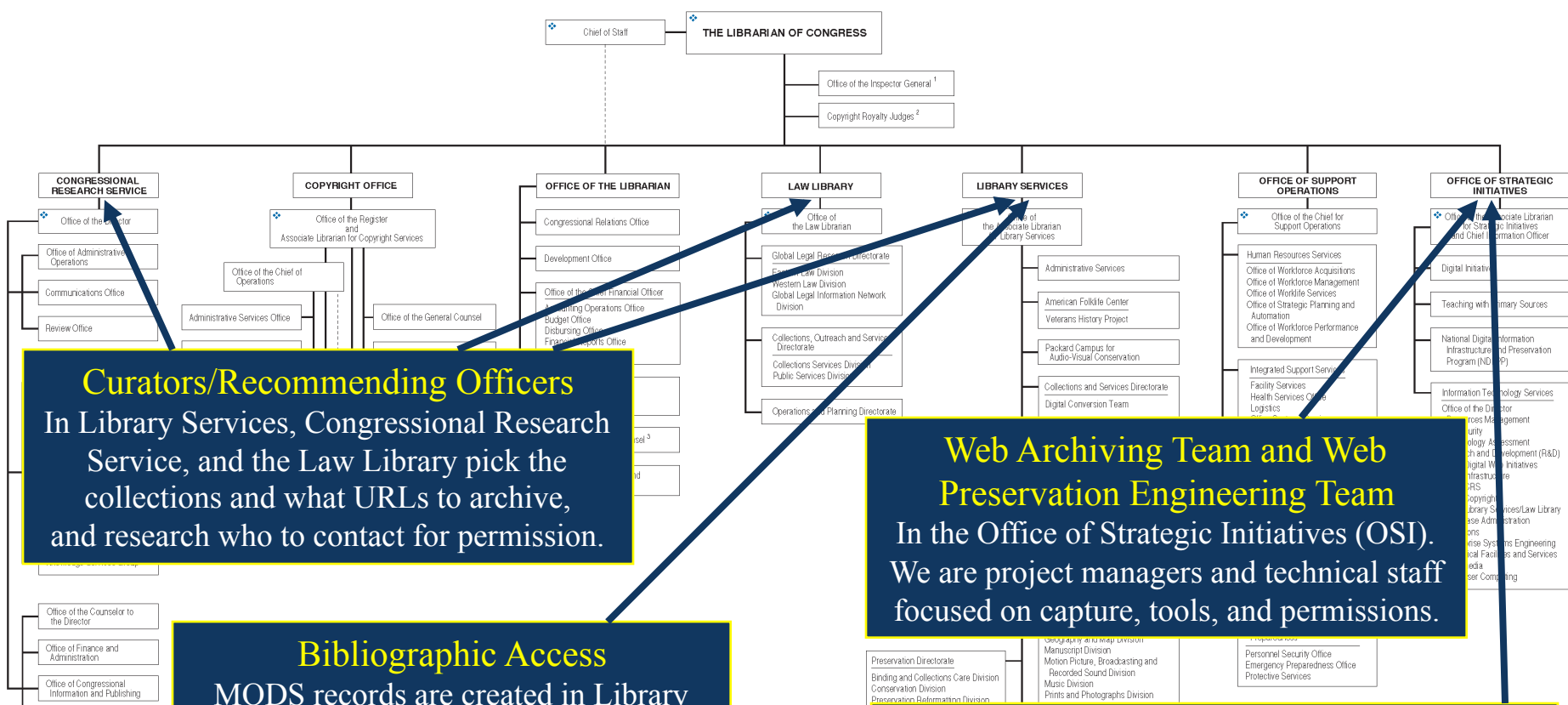
web archiving at LC

- began in 2000 with MINERVA pilot
- identify policy issues, establish best practices, build tools (internally and w/ partners)
- broaden expertise and understanding of Web Archiving within LC
- collect, manage and sustain at-risk digital content



LC Prints & Photographs: [design for Minerva Congressional Library](#)

Y * O



Curators/Recommending Officers
 In Library Services, Congressional Research Service, and the Law Library pick the collections and what URLs to archive, and research who to contact for permission.

Web Archiving Team and Web Preservation Engineering Team
 In the Office of Strategic Initiatives (OSI). We are project managers and technical staff focused on capture, tools, and permissions.

Bibliographic Access
 MODS records are created in Library Services: the Network Development and MARC Standards Office (NetDev) and Acquisitions and Bibliographic Access (ABA) staff do the cataloging.

Information Technology Office and Technical Architecture Team
 Also in OSI. Supports Wayback Machine, Heritrix, repository and tools development, and data transfers. Contractors are also used in this area.

LC collections: over 245 TB + 5 TB/month

- ongoing collections, including:
 - Congress/Legislative websites
 - Legal Blawgs
 - Public Policy Topics
- event-based collections, including:
 - U.S. National Elections—2000, 2002, 2004, 2006, 2008, 2010
 - Iraq War 2003-2009
 - September 11 2001 and September 11 Remembrance 2002
 - Civil War Sesquicentennial
 - Olympics 2002
 - Supreme Court Nominations
 - Papal Transition
 - Case Studies: health care, terrorism, visual image content, organizational Web sites, Crisis in Darfur, “single site”
- Overseas Operations collections, including: Egypt 2008; Brazilian, Indian, Indonesian, Philippine, and Thai Elections; Afghanistan Government; Pakistan Nationalisms

<http://www.loc.gov/webarchiving/collections.html>

web archives access: loc.gov/lcwa

The Library of Congress >> More Online Collections

Library of Congress Web Archives *Minerva*

[LC Web Archives](#)

Web Archives Available:

- [Crisis in Darfur, Sudan, Web Archive, 2006](#)
- [Iraq War 2003 Web Archive](#)
- [Law Library Legal Blawgs Web Archive](#)
- [Library of Congress Manuscript Division Archive of Organizational Web Sites](#)
- [Papal Transition 2005 Web Archive](#)
- [September 11, 2001 Web Archive](#)
- [Single Sites Web Archive](#)
- [United States 107th Congress Web Archive](#)
- [United States 108th Congress Web Archive](#)
- [United States Election 2000 Web Archive](#)
- [United States Election 2002 Web Archive](#)
- [United States Election 2004 Web Archive](#)
- [United States Election 2006 Web Archive](#)
- [United States Election 2008 Web Archive](#)
- [Visual Image Web Sites Archive](#)



The Library of Congress Web Archives (LCWA) is composed of collections of archived web sites selected by subject specialists to represent web-based information on a designated topic. It is part of a continuing effort by the Library to evaluate, select, collect, catalog, provide access to, and preserve digital materials for future generations of researchers. The early development project for Web archives was called MINERVA.

[LC Web Archives](#)

GRESS

essential tools

- capture: [Heritrix](#) (contract crawling w/ IA and in-house)
- replay: [Wayback](#)
- permissions/seed management; capture quality review; reporting; transfer tracking: custom apps built on [LAMP](#) stack
- transfer: [BagIt Library](#) (based on [BagIt spec](#)); *nix ingest/staging/storage/access servers; Internet2 connection

other useful tools


- web archiving workflow management: [NetArchiveSuite](#), [Web Curator Tool](#)
- small-scale web archiving: [HTTrack](#)
- Firefox add-ons: [Firebug](#), [Web Developer](#)

cataloging for access


- collection-level metadata
- site-level bibliographic metadata
 - nominators provide subject heading
 - HTML metadata extraction via [cURL](#)
 - cataloger assigns keywords
- cataloging metadata stored in [MODS](#)
- assisted keyword assignment: [HIVE](#)

collection-level record example

The Library of Congress >> [Go to Library of Congress Authorities](#)



LIBRARY OF CONGRESS ONLINE CATALOG



Help New Search Search History Headings List Titles List Request an Item Account Info Start Over

DATABASE: Library of Congress Online Catalog
YOU SEARCHED: Keyword (match all words) = election 2004 web archive
SEARCH RESULTS: Displaying 1 of 9.

◀ Previous Next ▶

Brief Record Subjects/Content Full Record MARC Tags

United States election 2004 Web archive

Relevance: ●●●●●


LC control no.: 2008700238
LCCN permalink: <http://lcn.loc.gov/2008700238>
Type of material: Loose-leaf, Web site, Database, etc.

Summary: A selective collection of approximately 2,000 Web sites associated with the United States presidential, congressional, and gubernatorial elections. The collection includes Web sites for candidates who appeared on the final state ballots as well as Web sites for political parties at the national level (all registered parties) and Democratic and Republican party sites at the state level; educational and research institutions; advocacy groups; government sites including federal, state and territorial, and election boards; creative expressions and miscellaneous Web sites related to the 2004 elections. This collection also included blogs (or Weblogs) centered on those certified as "Convention Bloggers" by the Democratic and Republican parties prior to the respective national conventions.

Subjects: [United States. Congress --Elections, 2004.](#)
[Governors --United States --Election.](#)
[Political campaigns --United States.](#)
[Elections --United States.](#)
[Political parties --United States.](#)
[Web archives --United States.](#)
[Blogs --United States.](#)
[United States --Politics and government --2001-2009.](#)

LC classification: [JK1968 2004](#)

CALL NUMBER: [Electronic Resource](#)
-- Request in: .Electronic Journal
-- Link(s): <http://hdl.loc.gov/loc.natlib/collnatlib.00000016>



CONGRESS

MODS bibliographic record example

Library of Congress Web Archives *Minerva*

[home](#) >> [overview](#) >> [browse results](#) >> **bibliographic information**

Election 2004 Web Archive

<< [Back](#)

Title: Arizona Secretary of State Home Page

Abstract: Secretary of State, Jan Brewer, Biography

Date Captured: October 9, 2004 - December 20, 2004
[Archived Site](#)

Subject(s): Elections--Arizona
Administrative agencies--Arizona
Elections--United States
United States--Politics and government--2001-

Language(s): English

Genre: web site

Access Condition: Access restricted to on-site users at the Library of Congress.

URL at time of capture: www.azsos.gov/

Citation ID: <http://hdl.loc.gov/loc.natlib/mrva0016.0178>

Record ID: mrva0016.0178

Collection Title: [Election 2004 Web Archive](#)

Wayback Machine Resource Page

The Library of Congress >> More Online Collections

Library of Congress Web Archives *Minerva*

BROWSE | SEARCH | TECHNICAL INFORMATION

[home](#) >> [previous page](#) >> Election 2004 Web Archive Collection Resource Page



Election 2004 Web Archive

<http://www.azsos.gov/>

2004 2003

January	February	March	April	May	June
July	August	September	October	November	December
			09, 15, 22, 25, 29	01, 04, 12, 19, 26	03, 07, 10, 16, 20

15 archived captures for 2004

[« Back to collection page](#)

Resource page by IGNACIO GARCIA

[home](#) >> [previous page](#) >> Election 2004 Web Archive Collection Resource Page

The Library of Congress >> More Online Collections

Contact Us

LIBRARY OF CONGRESS

example of an archived site



Note: External links, forms and search boxes may not function within this collection

Election 2004 Web Archive Collection

This is an archived Web site from the Library of Congress

<http://www.azsos.gov/>

Archived: 12/20/2004 at 19:34:27

[« Back to previous page](#)

[« First \(10/09/2004\)](#) [« Previous](#) #15 of 15 [Next »](#) [Last \(12/20/2004\) »](#)



Jan Brewer
Secretary of State

- About Jan Brewer
- Duties of the Secretary
- Contact Information
- News & Press Releases

Election Services

- Election Info
- Campaign Finance
- Lobbyist Info
- Voter Registration
- Citizens Clean Election Comm.
- Az Independent Redistricting
- County Election Contacts



The Secretary of State's Office welcomes you to the Arizona Secretary of State's Web Site. Our goal at the Secretary of State's office is to give you access to all of our public records in a timely and efficient manner. Let us know how we are doing, and what you would like to see. You can e-mail sosadmin@azsos.gov or our [Web Master](#)

New! Search our site: Go



Arizona - A Golden Rule State

Arizona, a Golden Rule State - Nominate someone today to be a Golden Rule Citizen. Learn more [HERE](#).

search and discovery

- bibliographic metadata search
- (not yet) Memento-enabled
- full-text search based on NutchWAX unfeasible
- Lucene/Solr looks promising



Challenges for

WEB ARCHIVES

challenges for web archives

- technical
 - large, deep, dynamic, interlinked
 - continuous transformation, simultaneously growing and disappearing
- intellectual property laws and regulations
 - legal deposit laws, mandates for preservation, laws that do not address web content
- economic environment
 - few good business models for sustaining web collections
- social environment
 - who is responsible and how is responsibility shared?

capture, replay, and preservation

- capturing websites - [Heritrix](#)
 - “form-fronted” databases (i.e., “deep web”)
 - URLs the crawler can’t see that we want
 - ...and URLs the crawler can see that we don’t
 - web 2.0 and other “new” web technologies
- replaying archived versions - [Wayback](#)
 - non-rigorous website coding
 - live site “leakage”
 - significant interactivity may be lost
- preserving access to our archives
 - billions of files
 - thousands of file types
 - how do we ensure content is accessible in 10, 25, 50 or more years?

scaling capacity

- budgetary pressures
- limited access server disk space
 - competing w/ other big data projects
- new infrastructure for new capabilities

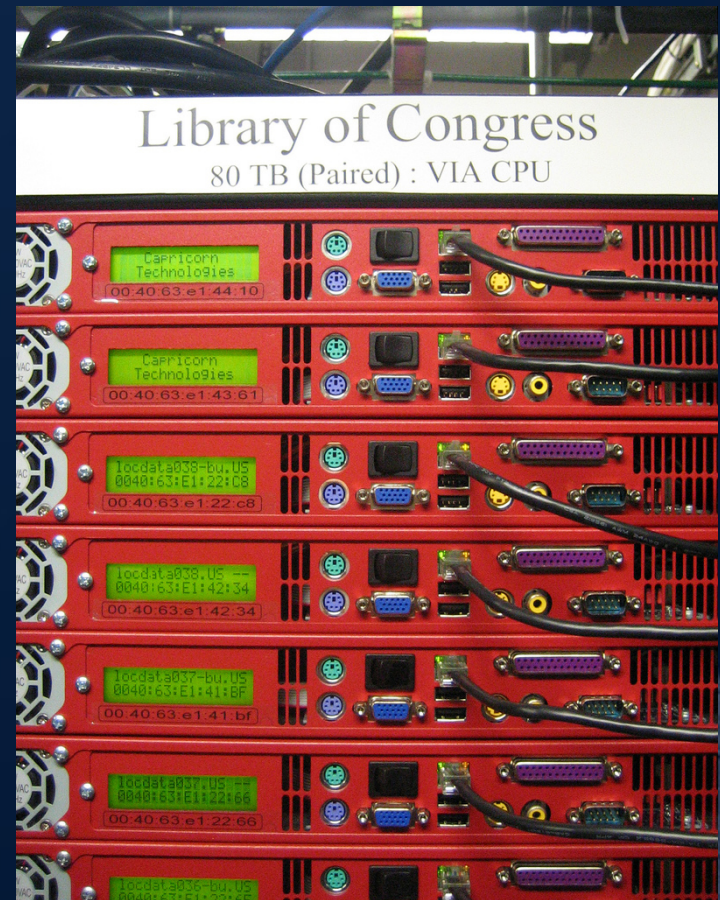


photo by [Henrik Bennetsen](#) under [CC BY-SA 2.0](#)

when the only tool you have is a library...



LC Prints & Photographs: [exterior view of the LC Jefferson Building](#)

...many things look like collections

- archive behaves more like discrete records than web
 - archived sites not contiguously navigable
 - data doesn't readily allow for downloading
- Twitter archive may prompt re-thinking web archive data access

web archiving and U.S. copyright law

- legal deposit requirement only applies to “published works” ([§ 407](#))
- [§ 108](#) of the Copyright Act provides library exceptions
 - doesn't address digital preservation and web archiving



photo by [Gabriel de Urioste](#) under [CC BY 2.0](#)

why not rely on robots.txt?

- unreliable proxy for copyright permissions
- archival crawler \neq search crawler
- LC disregards robots.txt but leaves contact info

```
User-Agent: *
Disallow: /music?
Disallow: /widgets/radio?

Disallow: /affiliate/
Disallow: /affiliate_redirect.php
Disallow: /affiliate_sendto.php
Disallow: /affiliatelink.php
Disallow: /campaignlink.php
Disallow: /delivery.php

Disallow: /music/+noredirect/

Disallow: /harming/humans
Disallow: /ignoring/human/orders
Disallow: /harm/to/self

Allow: /
```

last.fm: [robots.txt](#)

capture and access permissions

- permissions-based approach began in 2002
- permission plans for each collection developed w/ counsel
- permission requirements depend on site type
- more liberal about capture than about offsite access

	capture	access offsite
government	no notice	no notice
advocacy/ policy	notice	permission
news	permission	permission

implications of opt-in permissions

- no response treated as denial
 - very few denials
 - many non-responsive
- case study: September 11, 2001
 - 2300 cataloged, 30000 uncataloged URLs
 - many news sites (“high risk” permissions category)
 - no permissions sought
 - very few takedown requests

the future of permissions

- risk of more liberal approach appears low
- hope to move to more notice-based, opt-out policy
- may affect previously-captured sites as well



photo by [RJ Sangosti](#), [Denver Post](#) under © (fair use)



Challenges for the

TWITTER ARCHIVE

why archive Twitter?

- historical record of communication, news reporting, and social trends
- complements collections and mission



@klerner

Kevin Lerner

This is a night that justifies the Library of Congress archiving all of Twitter.

1 May via [TweetDeck](#)

Retweeted by [keithallynbaker](#) and 100+ others



<http://twitter.com/#!/klerner/status/64895357355704320>

Twitter Archive FAQs

- currently receiving Tweets through [Gnip](#)
- includes only the public archive
 - deletions will propagate to archive
- access limitations
 - 6 month embargo on new Tweets
 - no bulk distribution
- downstream users
 - no commercial use
 - no substantial re-distribution

a (literally) growing challenge

- ~3 years: time it took from 1st Tweet to billionth
- 1 week: time it now takes users to send a billion Tweets
- average Tweets/day in 3/10: 50 million
- average Tweets/day in 3/11: 149 million

<http://blog.twitter.com/2011/03/numbers.html>

tools we're exploring

- [Hadoop](#)
- [ElasticSearch](#)
- [Elephant-bird](#)
- [HBase](#)
- [Hive](#)
- [Pig](#)



photo by [The.Rohit](#) under [CC BY-NC 2.0](#)

questions to consider

- how does archive fit in w/ existing collections?
- how are agreement guidelines interpreted and implemented technically?
- what kind(s) of access can we provide?
- what context do we provide for content?

additional goals

- justify value to Congress and public
- understand and respond to researcher needs
- push the institution beyond existing curatorial models

for more information

- Library of Congress Web Archiving Program:
<http://www.loc.gov/webarchiving/>
- Library of Congress Web Archives:
<http://loc.gov/lcwa/>
- National Digital Information Infrastructure and Preservation Program:
<http://www.digitalpreservation.gov/>
- Library of Congress - Twitter FAQ:
<http://blogs.loc.gov/loc/2010/04/the-library-and-twitter-an-faq/>
- Section 108 Study Group:
<http://www.section108.gov/>

for more information

- “Legal Issues in Building Social Media Collections:”
<http://www.arl.org/bm~doc/mm11sp-okeeffe.pdf>
- “How the Library of Congress is building the Twitter archive:”
<http://radar.oreilly.com/2011/06/library-of-congress-twitter-archive.html>
- “Web Archives: The Future(s):”
http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1830025



questions?

Nicholas Taylor

[@nullhandle](#)

ntay@loc.gov

