Web and Twitter Archiving at the Library of Congress

Web Archive Globalization Workshop
June 16, 2011

Nicholas Taylor (@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 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
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.

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.

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.

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:
  - 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
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.
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
MODS bibliographic record example

Library of Congress Web Archives  Minerva

Election 2004 Web Archive

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://ndl.lcc.gov/loc.natlib/mrva0016.0178

Record ID: mrva0016.0178

Collection Title: Election 2004 Web Archive
## Election 2004 Web Archive

http://www.azsos.gov/

<table>
<thead>
<tr>
<th>Year</th>
<th>2004</th>
<th>2003</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>February</td>
<td>March</td>
</tr>
<tr>
<td>July</td>
<td>August</td>
<td>September</td>
</tr>
</tbody>
</table>

| 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
example of an archived site

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

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 sccadmin@azsos.gov or our Web Master

Search our site: 

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

<table>
<thead>
<tr>
<th></th>
<th>capture</th>
<th>access offsite</th>
</tr>
</thead>
<tbody>
<tr>
<td>government</td>
<td>no notice</td>
<td>no notice</td>
</tr>
<tr>
<td>advocacy/policy</td>
<td>notice</td>
<td>permission</td>
</tr>
<tr>
<td>news</td>
<td>permission</td>
<td>permission</td>
</tr>
</tbody>
</table>
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

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

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
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/
• 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-
• “Web Archives: The Future(s):”
questions?

Nicholas Taylor
@nullhandle
ntay@loc.gov