Cooperative Archiving: Event Harvesting in Perspective

Abbie Grotke, The Library of Congress
Benefits of Cooperative Archiving

- React more quickly to rapidly unfolding events
- Archive more content
- Variety of partners bring their own expertise to table: subject, technical, languages
- Learn from others, share with others
Examples

- September 11 Web Archive
- Hurricanes Katrina and Rita
- End of Term Government Archive (2008)
- Earthquake in Haiti
- Jasmine Revolution
- North Africa and Middle East (Arab Spring)
- Japanese Earthquake
- Olympics 2012
End of Term 2008 Project

Focus: US Government Websites

- Institutions: Library of Congress (LoC), California Digital Library (CDL), University of North Texas (UNT), Government Printing Office (GPO), Internet Archive (IA)
- Crawled: August 2008 – August 2009
- 2500+ seeds
- Bookend snapshots, weekly, monthly, quarterly crawls performed pre and post elections, pre and post inauguration
- ~25 TBs

<table>
<thead>
<tr>
<th>News from the Library of Congress</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Contact:</strong> Jennifer Gavin, Public Affairs Office (202) 707-1940; Abbie Grotke, Office of Strategic Initiatives (202) 707-2833</td>
</tr>
</tbody>
</table>

**August 14, 2008**

Library Partnership Preserves End-of-Term Government Web Sites

The Library of Congress, the California Digital Library, the University of North Texas Libraries, the Internet Archive, and the U.S. Government Printing Office today announced a collaborative project to preserve public United States Government web sites at the end of the current presidential administration ending January 19, 2009. This harvest is intended to document federal agencies’ online archive during the transition of government and to enhance the existing collections of the five partner institutions.

As part of this collaboration, the Internet Archive will undertake a comprehensive crawl of the .gov domain. The Library of Congress has been preserving congressional Web sites on a monthly basis since December 2003 and will focus on development of this archive for the project. The University of North Texas and the California Digital Library will focus on in-depth crawls of specific government agencies. The project will also call upon government information specialists -- including librarians, political and social science researchers, and academics -- to assist in the selection and prioritization of web sites to be included in the collection, as well as identifying the frequency and depth of the act of collecting. The Government Printing Office will lend expertise to the curation process along with libraries in its Federal Depository Library Program. A tool has been designed by the project team and developed by the University of North Texas to facilitate the collaborative work of these specialists, and will be made available to participants in August 2008.

"Digital government information is considered at-risk, with an estimated life span of 44 days for a web site. This collection will provide an historical record of value to the American people," said Director of Program Management Martha Anderson of the Library of Congress’ National Digital Information Infrastructure and Preservation Program (NDIIPP).
UNT Nomination Tool

End of Term Presidential Harvest 2008

Project Home  About This Project

Number of URLs Nominated: 4622
Number of Nominators: 29

Search by URL
http://
submit

Search for an existing URL in the system.

Browse URLs

COM
0 1 2 3 4 5 6 7 8 9
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

GOV
0 1 2 3 4 5 6 7 8 9
End of Term 2012


➡ Institutions: Library of Congress (LOC), National Archives and Records Administration (NARA), California Digital Library (CDL), Harvard Libraries, Harvard Kennedy School (HKS) Library and Knowledge Services, University of North Texas (UNT), Government Printing Office (GPO), Internet Archive (IA)

➡ In planning stage

➡ Crawl Start: Fall, 2011
What’s Coming for 2012…

➡ Outreach to gov doc experts to help seed 2012 end of term collection
➡ Outreach to researchers whose specialty is in analysing elections to address candidates for federal offices
Jasmine Revolution and Middle East Project Statistics

Jasmine Revolution - Tunisia 2011
- Institutions: Library of Congress, Bibliothèque nationale de France (BnF)
- Started: Jan 19, 2011
- 513 seeds
- daily, weekly, monthly, bi-monthly crawling
- 17.7 million urls, 1.7 TB
- status: winding down (currently just bi-monthly crawling, with a few seeds at weekly)

North Africa & the Middle East 2011
- Institutions: Library of Congress, BnF, British Library, American University in Cairo, Stanford University
- Started: Jan 27, 2011
- 2,020 seeds
- daily, weekly, monthly, bi-monthly crawling
- status: ongoing
- 35.5 million urls, 2 TB
Giant protest to kick off in Egypt

Thousands begin to assemble in downtown Cairo for the "million-man-march" aimed at forcing president Mubarak to resign.

Last Modified: 01 Feb 2011 07:04 GMT

Read More

- Live blog Feb 1 - Egypt protests
- Live Messages from Egypt
- New cabinet sworn in
- Chaos reigns at Cairo airport
- Egyptian financial crisis looms
- Al Jazeera equipment seized
Total internet blackout in Egypt

Country’s last working internet service provider has been disconnected, completing a complete shut down of web access.
Last Modified: 01 Feb 2011 01:56 GMT
Read More
Annual Meeting, Washington DC

July 20, 2011
Comité de Solidarité avec la Lutte du Peuple Egyptien est sur Facebook
Inscrivez-vous sur Facebook pour communiquer avec Comité de Solidarité avec la Lutte du Peuple Egyptien.
Managing Scope…

- The Web challenges our state and national boundaries and policies
- What’s in? What’s out?
- Need to define consistent selection and scoping criteria while the project and events develop.
- Tools such as the UNT Nomination Tool can help organize many people working on a project and the many URLs
- Who should take care of what is in between institutional boundaries - or everywhere?
- What is the risk? What is the value?