2008 DOT GOV HARVEST PRESERVING ACCESS

Cathy N. Hartman
Mark E. Phillips

FDLC
Oct 21, 2008

UNIVERSITY OF NORTH TEXAS LIBRARIES

Outline

- Project History
- Tool Building
- Partner Activities
- Future Work

Project History

- Collaborating Institutions:
  - Library of Congress
  - Internet Archive
  - California Digital Library
  - University of North Texas
  - US Government Printing Office

Project History

- First Meeting – Canberra, Australia
  - Early April 2008, at the National Library of Australia, International Internet Preservation Consortium (IIPC) – formed the partnership and discussed implications and possible roles for each institution.
  - Agreed from the beginning to share all content with any partner who wished a copy.

Project History

- Monthly meetings since that time – conference calls and one face-to-face meeting.
  - Defined roles
  - Released an announcement
  - Sought help from specialists to nominate URLs for harvesting
  - Shared technology planning
  - Developed URL Nomination Tool

Tool Building

- URL Nomination Tool
  - Allows for combining multiple seed lists
  - Allows for collaboration with subject experts
  - Helps create future seed lists
  - Helps to define overall scope of project

Pause for Vocabulary

- **Seed List –** List of URLs fed to the crawler for harvesting.
- **Crawler–** Software which downloads file, parses text to extract URLs, adds URLs to list and repeats
- **Scope –** Whether a URL should be included or not
- **Crawl –** Running a crawler on a given seed list
- **SURT –** Sort-friendly URI Reordering Transform

List of URLs

1890scholars.program.usda.gov
2001.cancer.gov
2001.nci.nih.gov
acc.nos.noaa.gov
access.usgs.gov
access.wa.gov
accessamerica.gov
accesstospace.gsfc.nasa.gov
acrim.jpl.nasa.gov
acs.oes.ca.gov
acweb.fsl.noaa.gov
adc.gsfc.nasa.gov

List of SURTs

gov.accessamerica

gov.ca.oes.acs

gov.cancer.2001

gov.nasa.gsfc.accesstospace

gov.nasa.gsfc.adc

gov.nasa.jpl.acrim

gov.nih.nci.2001

gov.noaa.fls.acweb

gov.noaa.nos.acc

gov.usda.program.1890scholars

gov.usgs.access

gov.wa.access

Back to the tool…

- Tool Requirements
  - Ingest seed lists from different sources
  - Keep track of who nominated seed
  - Record known metadata for seed
  - Allow people to help with nomination
    - Search
    - Browse
    - “easy to use”
  - Create seed lists for crawls

Tool Concepts

- URL – Single instance of metadata in system
  - URL
  - Attribute – (metadata element)
  - Value – (metadata value)
  - Nominator ID
  - Project ID
  - Timestamp
- Nominator
  - Nominator Email Address
  - Nominator Name
  - Nominator Institution
- Project
  - Project Metadata

Batch Ingest

- Administrator can import CSV files with URLs and associated metadata with batch importer
- An ingest needs to be associated with a Nominator and a Project.
- Arbitrary metadata is recognized and added to the system.

Nomination – In Scope/Out of Scope

- On batch import a URL is given a positive nomination +1
- A user of the Nomination Tool has the ability to nominate a URL as in scope (+1) or out of scope (-1)
- Nominations are calculated to give a possible measure of importance for a project.

EOT 2008 Project Metadata

- Metadata fields defined for EOT 2008 Harvest
  - Branch
  - Department/Agency Name
  - Title
  - Comment
- Nominators don't need to register but Name, Email and Institution are required.

Nomination Tool Demo

- URL Nomination Tool
- URL Nomination Tool - Admin

You can give us a hand…

- For more information about the project and partners, visit: http://www.loc.gov/today/pr/2008/08-139.html

- To sign up to participate, send an email to eotproject@loc.gov for more information.

Partner Roles

- **Internet Archive** – Broad Crawls
- **Library of Congress** – In-depth Legislative branch crawls
- **University of North Texas** – Sites/Agencies that meet current UNT interests and collections, as well as several "deep web" sites.
- **California Digital Library** – Executive/Judicial branch emphasis with one site per crawl scope
- **Government Printing Office** - Support

Project Schedule - 2008

- **August 14**: Project Announced via Press Release
- **End of August**: Call for volunteer nominators went out. Nominators began prioritization of URLs.
- **September 15**: Internet Archive began first broad crawl of everything that was in the nomination tool at the project start.
- **October 17**: UNT began crawl of prioritized URLs of topic areas interest to their collection development.

Project Schedule 2008 – cont.

- **October 29**: LC expanded monthly Congressional Crawl
- **November 1 – January 19**: CDL begins weekly crawl of all prioritized URLs. Will continue weekly until the Inauguration.
- **November 4: Election Day -** UNT begins second crawl of prioritized URLs of topic areas interest to their collection development.
- **November 26:** LC expanded monthly Congressional Crawl
- **December 31**: LC expanded monthly Congressional Crawl

Project Schedule 2009

- **January 20:** Inauguration Day
- **January 21:** Internet Archive begins second broad crawl - de-duplication will be on.
- **UNT begins** third crawl of prioritized URLs of topic areas interest to their collection development.
- **January 28:** LC expanded monthly Congressional Crawl begins
- **Late February/March:** Public access via Internet Archive's website

Near-Future Work

- Centralizing web data into a single collection at the Internet Archive
- Providing WayBack access to content
- Providing search access to content
- Distributing collection among partners
  (25-35 TB projected)
- Investigation of browse by Agency/Branch

Other Future Work

- Extracting topical collections from crawl data
- Providing programmatic access for data-mining
- Research in calculating "size" of collection in relation to real world measures
  - Number of pages of text collected
  - Number of 8x10 in equivalent images collected
  - Hours of Audio
  - Hours of Video
  - Physical library space requirements to hold collection if in physical format.

Questions?

cathy.hartman@unt.edu

mark.phillips@unt.edu