WEB ARCHIVING

Tools for the Capture of Digital Assets on Websites

March 27, 2006

Kelly Eubank

Why Capture Websites?

- Websites now the primary way that North Carolina state agencies communicate with the public
- Over 80% of publications disseminated through the Web
- Important records on the Web:
  - Minutes, speeches (video), policies, images
- Websites have become an important part of agency history

What is Web Capture?

- Web crawler or “spider” collects web content
- Starts at predetermined list of URLs
- Makes a copy of web page, including all objects that are part of the web page
- Follows hyperlinks and captures additional web pages, as long as part of acceptable domain list
- Content captured is “clickable content”
  - must be a link for spider to find
  - must not require input from user

Website Archiving Activities to Date

- Website guidelines and capture from 2001
- http://www.ah.dcr.state.nc.us/records/e_records/default.htm#web
- CEP from IMLS grant
- WAW from OCLC, NDIIPP partnership
- Archive-It from Internet Archive

Capturing Electronic Publications

- Developed by the University of Illinois through an IMLS grant
- From September 2004-2005, NC State Library participated in a pilot program to capture websites.
- Open-source software
- 200 state government sites, monthly
- Currently, 273 gigabytes
- No search mechanism

Web Archives Workbench

- Developed by OCLC through National Digital Information Infrastructure Preservation program
- Based on the Arizona Model for Web site Harvesting
- Consists of 4 tools: Discovery tool, Properties tool, Analysis tool, and Packager tool
- Beta Testing phase—to be completed in 2008

Archive-It Project

- 3 collections of 100 websites each
- 10 million documents total, up to .5 terabytes of data

Scope Continued

- Crawled 3 collections—Cabinet, Council of State, Boards and Commissions
- 274 web addresses
- Initial search was daily, scaled to weekly after 3 weeks.
- End of Project– 9.5 million objects
- Archive-it captured 54 different formats

Archive-it captures at least 4 important types of archival documents

- The Web contains valuable archival records in varying file formats.
- Minutes—Microsoft Word, PDF
- Speeches--Streaming Video, MP3, .WAV
- Images--.JPEG, .GIF, .PNG
- Policies--.HTML, .PDF

Speeches/Video

- A search in Archive-it allows you to specify file types.
- Search “governor and .rm”
- Search: N.C. Project Green
- There are countless other examples. During our trial period we captured over 700 instances of video.

Discussion of Capture

- Relative versus absolute links:
  - http://www.ncagr.com
- Java Script:
  - http://www.ncstatefair.org/
- Some files are redirected to the Live web
- Community Colleges
  - http://www.ncccs.cc.nc.us/

Discussion of Capture

- Archive-It cannot capture Streaming Video
- Archive-It cannot capture dynamic database driven sites
  - http://www.ncfarmfresh.com/
- Archive-It cannot capture password protected sites.

Lessons Learned For IA

- One hop off
  - "out of scope materials"
  - "inappropriate materials"
    - Search for "lottery"
  - Tremendous amount of material would have to be masked.
- Re-do analysis to go into production
  - Turn off of one hop off feature.
- Cannot search across collections
  - Clumsy search
  - Broken links

Analysis of Cabinet Collection

Cabinet	Domains	urls	bytes
In Scope	**251**	**563,169**	**59,397,010,697**
Out of Scope	3,622	124,290	5,142,890,476
Total	3,873	687,459	64,539,901,173
Percentage out of Scope	94%	18%	8%

Analysis of 3 Methodologies

- **CEP—**on-site solution
  - One spider for each URL CVS format
  - Control over the crawler
  - Need institutional support
  - Need IT support.
  - Practitioner needs background in programming.

  WAW—hosted solution
  - Same crawler. Assume it will work similarly to Internet Archive.
  - Crawls specific web documents.
  - Control over the crawler but
- **Internet Archive—**hosted solution
  - captures a moment in time
  - crawls everything within the domain.
  - No control over crawlers

Website Philosophy

- Arizona Model vs. “Newspaper” theory
  - Arizona
    - Capture identified “series” hosted on website
    - Crawler only goes to the specified site
  - Newspaper
    - Website contains information that might be valuable or interesting to future resource

Going into Production

- 2006—Archive-It tool
- IA turned off one hop off feature
- 300 “active” seeds
- Combining Collections into one Collection
- Will continue to test WAW

Contact Information

Kelly Eubank

Electronic Records Archivist

North Carolina Archives and History

Telephone: (919) 807-7355

Email: kelly.eubank@ncmail.net

Web:

http://www.ah.dcr.state.nc.us/records/default.htm