Digital Library Federation, November 8, 2006

Web Archiving Update

Jennifer Marill
Office of Strategic Initiatives, Library of Congress
jmarill@loc.gov, 202-707-2868

www.loc.gov/webcapture/
Agenda

• LC Web collections
  – Highlights and observations
• Some current activities of the Library
  – Process and policy issues
  – Tool development
  – Storage and transfer
  – Partnerships and collaborations
• Directions for 2007
LC Collections

• Event
  – Elections 2000, 02, 04, 06, Supreme Court nominations, Papal transition, 9/11…

• Topic/Subject
  – Iraq War, Crisis in Darfur…

• Region/Domain
  – Congressional (.gov) . . .

• Genre
  – Legal blogs, still image content …
LC Collections – cont.

- Partnership collections
  - GMU/CUNY—September 11 Digital Archive
  - University of Arizona—Terrorism Web sites
  - CDL & Internet Archive--Hurricane Katrina
- Almost 60 TB of data

More: www.loc.gov/webcapture/projects.html
September 2, 2005

President George W. Bush delivers a speech to the Supreme Court Nominations:

"One of the most consequential decisions a President makes is the appointment of a Justice to the Supreme Court. When a President chooses a Justice, he's selecting someone who will apply the Constitution and uphold the rule of law. The decisions of the Supreme Court affect the life of every American.

And a nominee to that Court must be a person of superb intellect and the highest integrity, a person who will faithfully apply the Constitution and keep up a strong, principled voice of judicial restraint."

October 3, 2005

President George W. Bush announces Judge John Roberts as Supreme Court Nominee:

"I am pleased to announce Judge John Roberts as my nominee to be an Associate Justice of the Supreme Court. Judge Roberts is a brilliant legal mind, a man of integrity, and a tireless advocate of the Constitution."

October 31, 2005

President George W. Bush announces Judge Samuel A. Alito Jr. as a Supreme Court Nominee:

"I am pleased to announce Judge Samuel A. Alito Jr. as my nominee to be an Associate Justice of the Supreme Court. Judge Alito is a man of integrity, a tireless advocate of the Constitution, and a true conservative."

The White House website also features articles and updates on these events.
Election 2000: Vote Swap Sites

voteswap.com

Vote Swapping Portal

Are you considering voting for Ralph Nader but afraid to because George Bush might win this election? Third party candidates play an important role in setting discussing important issues and solutions that the major party candidates avoid in their effort to pander to the voters.

Are you an Al Gore supporter that is concerned that the Nader vote might cause George Bush to win? In 1968 Richard Nixon defeated Hubert Humphrey when Eugene McCarthy ran as a third party candidate. In 1980, Ronald Reagan was elected when Anderson ran as a third party candidate.

Voting for the candidates who represent best your views is desirable. Having this vote elect the least desirable candidate is not. Here is how you can have your cake and eat it too.

The presidential election consists of 50+ separate elections that determine the outcome of the election. Some of these elections are close and the Nader vote could make a difference. Some are so large that the major candidates are not even bothering to campaign there.

If you are in a close state and considering voting for Nader or undecided because you are afraid of electing George Bush, then you can swap your vote with someone from another state that will pledge to vote for Ralph Nader if you would vote for Al Gore.

Both Al Gore and Ralph Nader will be winners. The vote count for Ralph Nader will go up and hopefully reach a level in which he could receive matching funds to mount a more serious challenge or at least cause the major parties to change. Al Gore could defeat George Bush.

There are several sites available that will match you with someone else who is willing to swap your vote with theirs and notify you when they find someone to swap your vote with.

Vote swapping is not the same as vote pairing. With vote pairing, a Bush voter finds a Gore voter and both agree to vote for Nader instead of canceling each other’s vote (See VotePact.com). It requires more trust on both parties. For example, if the Bush voter doesn’t keep his or her pledge, then Gore loses a vote. Vote swapping requires less trust since Bush cannot benefit by either party failing to keep their pledge. In addition, there is no benefit for a Bush supporter to falsely register.

Swap your vote on these sites:
Election 2002: Blog Collecting Begins

Sunday | November 03, 2002

MN polls all over the place

Actually, the title of this post is misleading. Most polls have shown Mondale with a small but decent lead (5 points or so). However, one poll, by the Pioneer Press/Minnesota Public Radio, shows Coleman with a 46-41 lead.

So to recap:

- *Zogby*
  - Mondale: 50
  - Coleman: 45

- *Star Tribune Minnesota Poll* (released today)
  - Mondale: 48
  - Coleman: 41

- *Star Tribune* (last Wednesday)
  - Mondale: 47
  - Coleman: 39

And the crazy-ass poll from the Pioneer Press and NPR:

- Mondale: 41
- Coleman: 46

Hmm, two separate indie polls show identical 46-41 results, which jive (when accounting for margin of error) with another indie poll conducted last week. The last poll has the numbers completely reversed, and that poll surveyed 300 less likely voters than the ST poll (929 vs. 625).

I like Mondale's chances.

Aside from that, here's the day's funniest quote:

[A's political strategists] have to sit back and consider nothing more sophisticated than our guts. And my gut tells me Coleman has lots and lots of political momentum.

Who said the quote? Tara Mason, a spokesperson for the Coleman campaign. So believe it or not, the Coleman...
Statement about Paul Wellstone

Paul Wellstone was one of a kind. He was a man of principle and conviction, in a world that has too little of either. He was dedicated to helping the little guy, in a business dominated by the big guys. We who had the privilege of working with him hope that he will be remembered as he lived every day: as a champion for people.

His family was the center of his life and it breaks our hearts that his wife of 39 years and his daughter Marcia were with him. Our prayers are with Mark and David and the grandchildren he and Sheila cherished so much.
Election 2004: Blogging and Online Donations

Howard DEAN for America

Make a Contribution

Help Governor Dean take his message to every community in America by generously contributing today. Make a one-time contribution, or contribute monthly as a Dean for America Sustainer.

Contribute MONTHLY
MAKE A RECURRING MONTHLY CONTRIBUTION

Contribute
MAKE A ONE-TIME CONTRIBUTION TODAY

Contribute
PRINTABLE FORM

Or contribute using the toll-free number:
Election 2004: Meetups

National Dean in 2004 Meetup Day

in 19 Days

WHAT
Meetup to discuss Howard Dean's campaign to take back America. Get the latest campaign news and organize your community for Dean.

WHO
Dean Supporters Worldwide. So far, 188,609 have signed up.

Join other Dean Supporters near you!

Dean in 2004 Meetups can happen in up to 612 cities worldwide on the same day. Enter your location to find the one near you:

Country: United States
Zip Code

continue
Election 2004: Campaign Site Companies

Welcome to Campaign Site Builder - Mozilla Firefox

You are viewing an archived website that is part of a Library of Congress Web Archive Collection. External links, forms, and search boxes may not function within this collection.

CampaignSiteBuilder.com

FREE SPECIAL REPORT
Avoiding the Five Biggest Campaign Website Mistakes

40 FREE Campaign Templates

FREE GETTING STARTED CAMPAIGN LESSONS

FREE 30-DAY Trial
Click Here!

110% MONEY-BACK GUARANTEE

Free 30-Day Trial Click Here!

Build your Campaign Website in 10 minutes!

- Accept Credit Card Contributions
- No Training or Software Required
- Personalized Email Accounts
- Unlimited 24/7/365 Editing
- Hosting Included
- Events Calendar
- Unique Domain Names
- No Hidden Costs or Fees
- Volunteer Sign-up Form
- And Much, Much More

Watch an Online DEMO
Click Here!
2006 Candidate Site Observations

• More complex sites with more bells and whistles, some hosted on external sites:
  – Photos hosted on flickr
  – Stores
  – Myspace pages
  – Video sites
  – Blogs everywhere - some as campaign sites

• More targeted companies hosting candidates sites
Welcome to The Foley for Congress Online Campaign.

Here you can join our team as we take our campaign across the 16th District of Florida from the beaches of the east coast, across Lake Okeechobee to the Gulf shores of Southwest Florida.

I need friends like you on this campaign and urge you to volunteer, show your support by placing a sign in your yard or bumper sticker on your car or contributing to our campaign.

We have a strong record of accomplishment with innovative ideas for Florida’s families and I look forward to earning your support and help.

Sincerely,

Mark Foley
Election 2006: The Wave of the Future...
The Web Capture Process

Collection Planning → Legal Review → Selection → Notification/Permissions → Technical Review → Crawl & QA → Cataloging → Interface Development → Access → Store & Manage
Project with Library Services Staff

- Define roles and responsibilities for life cycle management of archived web content
- Work on clear and consistent intellectual property protocols for crawling
- Explore how technical aspects of Web sites affect selection criteria
- What content should we now be collecting? How long should we collect it?
Project with Library Services Staff, cont.

- Document staff resource requirements, time and workflow
- Expand staff participation
- Train larger group of LS staff
- Learn by doing
- Inform and educate other Library staff
Tool Challenges

• Tools to select, acquire and display are not available off-the-shelf
  – Small community of users

• Tools will always be playing catch-up to current Web standards, file formats
  – Continuum for what can be captured

• Tool suite: selection, permissions, crawling, quality review, storage and preservation, access and display
Selection and Permissions: Web Collection Management System

• Module 1 will support:
  – Web site selection and nomination for multiple collections
  – Site review and limited metadata collection
  – Permissions gathering process
  – Technical review prior to crawling
  – Setting crawl parameters
  – Creating crawl specification and submitting a “work order” to crawler.
Capturing Streaming Media

- Plug-in for the post crawl process
- Phase 1
  - Review Heritrix's activity and errors logs
  - Extract URLs we assume will lead to video files
- Phase 2
  - Download the video files and bundle them into ARCs
  - Add metadata gleaned from Heritrix and open source application MPlayer
- Phase 3
  - Store the ARCs in relation to the other ARCs in the collection
Capturing Streaming Media, cont.
Access and Display

• Access tools necessary so crawled content is viewable
  – Multi-dimensional nature of archived web sites
  – Fixing links in content to point back into the archive rather than out to the live Internet

• Current tools do the basics, but more work needed
  – LC public collections use IA/Alexa Wayback Machine, under restricted license, with no search features
  – IIPC working on NutchWAX: integrates Nutch/Lucene search tools with Web Archive EXtension
  – IA rewriting Wayback Machine as open-source
## 2006 U.S. Election Crawl


<table>
<thead>
<tr>
<th>Jan - Feb</th>
<th>Mar - Apr</th>
<th>May - Jun</th>
<th>Jul - Aug</th>
<th>Sep - Oct</th>
<th>Nov - Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 pages</td>
<td>0 pages</td>
<td>0 pages</td>
<td>4 pages</td>
<td>16 pages</td>
<td>0 pages</td>
</tr>
</tbody>
</table>

* denotes when site was updated.

**Search Results for Jan 01, 2006 - Nov 04, 2006**

- Aug 23, 2006 *
- Aug 29, 2006 *
- Aug 30, 2006 *
- Aug 30, 2006 *
Refugees International: Search: Darfur 2005: 96-year-old Flees Janjaweed

... Refugees International: Search: Darfur 2005: 96-year-old Flees Janjaweed Search: Home About Us Donate Where We Work Issues Contact Us Photo Gallery > Darfur 2005: 96-year-old Flees Janjaweed ...

http://www.refugeesinternational.org/content/photodetail/6149/ - 2005-02-10 02:44:20 AM - text/html - other versions - more from
www.refugeesinternational.org - explain

Khartoum's Janjaweed militia allies burning children alive, August 2, 2004 :: sudanre news.org :: Sud

... Khartoum's Janjaweed militia allies burning children alive, August 2, 2004 :: sudanre news.org :: Sudan Research, Analysis, and Advocacy Khartoum's Janjaweed ...


Sudan: Darfur Destroyed: SUMMARY

... other Africans, and therefore liable to sympathize with their complaints. "Come to the Janjaweed... The Janjaweed is the government. The Janjaweed is Omar Bashir." 113 In December 2003, Janjaweed moved into the army barracks in Murree. Hassan, a twenty-two-year-old student...


http://www.darfurpeaceanddevelopment.org/research_papers.htm

... a single day as recently as March 2004. One farmer from Sidi, 40-year-old Ahmad, said "the Janjaweed come with horses and camels and..."
Welcome

Welcome to the Black Caucus of the American Library Association, Inc. website. This website is designed to serve African American library and information professionals seeking to learn and share in the exchange of information and resources for preserving and providing access to our cultural heritage. To learn more about participating in BCALA events and activities please take a look around our website.

Quick Links

JCLC Award Nomination Form
E.J. Josey Scholarship Award
Student Resources
HBCU Library Alliance
6th National Conference (Call for Participation)
Cardin Launches New TV and Radio Ads
10.03/2006

Contact: Oren Shur, 410-366-1900


The TV ad, which is airing on all four Washington networks, highlights key differences between Rep. Cardin and Lt. Gov. Steele. Rep. Cardin says he will fight President Bush's plan to privatize Social Security and will support funding embryonic stem cell research, while Steele will not.


The radio ad features Prince George's County State's Attorney Glenn Ivey (D) hailing Rep. Cardin’s long record of standing up for Marylanders on the issues that matter most. It will air on the following D.C. area radio stations: WKYS, WMMJ, WOL, WYCB, WPGC, WYCB, WHUR.

“With the serious issues we face in Maryland, Ben Cardin is who we need fighting for us in the Senate,” Ivey says. “To give every child in every community the chance to succeed, we need leaders like Ben Cardin.”
Storage and Transfer

• San Diego Supercomputer Center (SDSC)
  – One-year project to test a trusted third party to store large quantities of data
  – 6 TB of data transferred by shipment and over Internet 2 to SDSC
  – Testing remote management, including the ability of LC users to upload, download, delete, modify, and generate data reports

• At the Library
  – Transferring data into archival storage
  – How often are checksums, reports and inventory needed?

• Practice and policies will emerge from efforts
Partnerships and Collaborations - NDIIPP

• University of California Digital Library
  – The Web at Risk: A Distributed Approach to Preserving our Nation’s Political Cultural Heritage
    • Collecting CA government information
    • Tool suite
    • Partnership building

• University of Illinois at Urbana-Champaign, OCLC
  – The ECHO DEPository - how digital resources will be identified, archived and preserved
    • AZ State Library selection rationale using an archival approach
    • OCLC tool suite
    • Repository evaluation
Partnerships and Collaborations, cont.

- Internet Archive: 20th Century Archive Project
  - Further development of NutchWAX suite for indexing and searching large-scale Web collections
- San Diego Supercomputer Center
  - Testing remote management of Library’s data
- Information sharing with other US government agencies
  - Government Printing Office
  - National Archives and Records Administration
International Internet Preservation Consortium (IIPC)

• Collect and preserve a rich body of Internet content from around the world
• To foster the development and use of common tools, techniques and standards
• Share experiences and best practices with libraries everywhere
• 12 members during the charter period July 2003 - July 2006
IIPC Charter Members

- National Library of France, Lead
- National Library of Italy, Florence
- Royal Library, Denmark
- Helsinki University Library, National Library of Finland
- University and National Library of Iceland
- Internet Archive, USA
- Library and Archives Canada
- National Library of Norway
- National Library of Sweden
- National Library of Australia
- British Library, UK
- Library of Congress, USA
IIPC Future

- Next consortium phase 2007-2009
- Continue sharing information and facilitating tool development
- Open to any library, archive or heritage/research organization with experience in or commitment to Web archiving

More information in early 2007: netpreserve.org
Directions for 2007

• Collecting
  – Should there be a core set of regularly collected events and/or sites?
  – How do we build infrastructure and what are the triggers to support “spontaneous” events, such as terrorist attacks, natural disasters, etc.
  – Monitor Section 108 Study Group developments
Directions for 2007, cont.

- Improving selection tool
- Full-text indexing of collections
- Survey of access needs
- Large-scale collection storage

More: www.loc.gov/webcapture