The 3 M’s: MINERVA, MODS, and METS

Allene Hayes (LC)
Rebecca Guenther (LC)
Leslie Myrick (NYU)

DLF -- New Orleans  April 20, 2004
Topics of Discussion

- MINERVA http://www.loc.gov/MINERVA/
- MODS http://www.loc.gov/standards/mods/
- METS http://www.loc.gov/standards/mets/
Mission Statement

• MINERVA collects and preserves Web sites and Web pages for the Library

Goal is to preserve primary source materials for future generations
Why Archive Web sites?

• "The Internet is as important as the print media for documenting these events."

• "Why? Because the Internet is immediate, far-reaching, and reaches a variety of audiences. You have everything from self-styled experts to known experts commenting and giving their viewpoint."
On the Web, Research Work Proves Ephemeral
Electronic Archivists Are Playing Catch-Up in Trying to Keep Documents From Landing in History's Dustbin

By Rick Weiss
Washington Post Staff Writer
Monday, November 24, 2003, Page A08

It was in the mundane course of getting a scientific paper published that physician Robert Dellavalle came to the unsettling realization that the world was dissolving before his eyes.

The world, that is, of footnotes, references and Web pages.

Dellavalle, a dermatologist with the Veterans Affairs Medical Center in Denver, had co-written a research report featuring dozens of footnotes -- many of which referred not to books or journal articles but, as is increasingly the case these days, to Web sites that he and his colleagues had used to substantiate their findings.

Problem was, it took about two years for the article to wind its way to publication. And by that time, many of the sites they had cited had moved to other locations on the Internet or disappeared altogether, rendering useless all those Web addresses -- also known as uniform resource locators (URLs) -- they had provided in their footnotes.

"Every time we checked, some were gone and others had moved," said Dellavalle, who is on the faculty at the University of Colorado Health Sciences Center. "We thought, 'This is an interesting phenomenon itself. We should look at this.'"

He and his co-workers have done just that, and what they have found is not reassuring to those who value having a permanent record of scientific progress. In research described in the journal Science last month, the team looked at footnotes from scientific articles in three major journals -- the New England Journal of Medicine, Science and Nature -- at three months, 15 months and 27 months after
Our Core Team

- Project Manager
- 2 Reference Specialists
- 2 Digital Conversion Specialists
- Cataloging Specialist
- Network Development and MARC Standards Office representative
- and soon, we hope, a digital rights specialist...
Partners

- Internet Archive
- Alexa Internet
- WebArchivist.org
  - University of Washington, Center for Communications and Civic Engagement
  - State University of New York, Institute of Technology (SUNY-IT)
Collections

**FY00**

LC Prototype: sampling of sites
Election 2000: 767 sites

**FY01**
September 11th: 30,000+ sites
FY02
2002 Winter Olympics: 70 sites
9/11 Remembrance: 1,800 sites
Election 2002: 3,000+ sites

FY03-04
107th Congress: 588 sites
War on Iraq: 288 sites
Election 2004: 134 sites & growing
The Prototype

• Initial test Mid 2000
• LC crawled about 30 sites using a desktop Web crawler (HTTrack)
• Provided an initial investigation into the processes involved in Web capture
• Sites were cataloged
• Access is on campus only
The MINERVA Process

Collection Planning

Legal Review

Selection

Notification/Permissions

Technical Review

Crawl & QA

Cataloging

Interface Development

Access

The MINERVA Process

Collection Planning

Legal Review

Selection

Notification/Permissions

Technical Review

Crawl & QA

Cataloging

Interface Development

Access
The Process is Evolving

Collection Planning
Legal Review
Notification/Permissions
Legal Authority
Cataloging
In-house expertise
Access
Templates & Tools
Interface Development
Crawl & QA
Automation & Tools?
Alternate Crawl Contractors
Preservation?
Non-Event based collecting
Select + create metadata up front
International Internet Preservation Consortium
In-house expertise
Templates & Tools
OSI
Preservation?

Selecting
International
Internet Preservation
Consortium

Non-Event based collecting
Select + create metadata up front
Collection planning

• Identify the “event” (thematic)
• Secure the funding
• Develop task order
  – Scope of collection
  – Acquisition parameters
    • Size of collection
    • Collection period (start and end dates)
    • Web site URLs – define categories for selection
    • Frequency, depth, breadth
Legal Review

- Bibliographic Services Agreement
- Task Orders
- Notifications and permissions
  - Fair use argument for event-based collecting
    - Notice of crawl; Permission to display offsite
    - Permission to crawl and display: foreign and “creative” sites
  - Modification of 407 mandatory deposit regulation
MINERVA reference specialists work with Recommending Officers to:

- Develop collection policy statement for event-based collections
- Identify specific Web sites to collect
- Identify contact information of site owner/producer
- Enter into database using nomination form
Notification/Permissions

Tasks currently shared by all members of the team:

- Email appropriate notification and permissions
- Respond to phone & email queries
- Find alt addresses and resend if no response
- Track responses in database
- Provide access condition statements for catalog record, inform ITS of restrictions
Technical Analysis

• Identify potential technical challenges:
  – Macromedia Flash™ introductions
  – Log-ins, either free or pay sites
  – Dynamically-generated Web pages
  – Dynamic menu
  – Databases

• Determine best point of entry (eg. start URL)
Crawl and QC

- Access to crawl 24 hrs after completed
- Quality assurance using status and error reports
- Troubleshoot problems with crawl
- Modifying the crawl to be more focused, refined, specific
CATALOGING

• Prototype
  – MARC records in ILS

• Collection Level
  – MARC records in ILS

• MODS Item Level
  – MODS records on MINERVA site, but created with import to ILS in mind
What is MINERVA?
- About MINERVA

Collections
- Election 2002 Web Archive (Jul 1, 2002 - Nov 30, 2002)
- September 11, 2001 (Sept 11, 2001 - Dec 1, 2001)
- Election 2000 (Aug 1, 2000 - Jan 21, 2001)
- Prototype (Library of Congress staff access only)

Coming Soon ...
- Winter Olympics 2002 (Feb 8, 2002 - Feb 24, 2002)
- September 11th Remembrance (Sept 11-18, 2002)
- The 107th Congress (Dec 2002)

Papers, Articles & Reports
- Collecting and Preserving the Web: The Minerva Prototype (RLG DigiNews)
- Report to the Library (William Y. Arns, Cornell Univ.)
  - Interim Report (PDF: 217K / 32 pages)
  - Final Report (PDF: 169K / 20 pages)
  - Download Acrobat Reader (for PDF files)

Presentations

Awards
- September 11 Archive:
  - Site of the Year (Yahoo! Internet Life)
  - Internet Scout Report

Election 2000 Archive:
- Internet Scout Report
Minerva Home Page - Browse by Subject - Browse by URL

0-9

Bush/Cheney (http://www.georgewbush.com)


History and Politics Out Loud (http://www.hpol.org)
LC Control Number: 00530048
Type of Material: Computer File
Main Title: Bush-Cheney [computer file].
Variant Title: Also known as: Bush-Cheney 2000
Also known as: georgewbush.com
Title as viewed on Aug. 21, 2000: George W. Bush for President
Published/Created: Austin, Tex. : Bush for President, 2000-
Related Names: Bush for President, Inc.
Summary: Presents information about Texas Governor George W. Bush (b. 1946) and his campaign to become the U.S. President. Contains a biographical sketch, a schedule of appearances, and news about the campaign. Offers access to details about issues and speeches. Discusses how to participate in the campaign.
Notes: Title from home page as viewed on Oct. 19, 2000.
In English and Spanish.
archived 2000
Subjects: Bush, George W. (George Walker), 1946-
Cheney, Richard B.
Governors--Texas--Biography.
Presidential candidates--United States--Biography.
Presidents--United States--Election--2000.
Political campaigns--United States.
LC Classification: F391.4.B87
Dewey Class No.: 324.973 13
Language Code: engspa
What is MINERVA?

- About MINERVA

Collections

- Election 2002 Web Archive (Jul 1, 2002 - Nov 30, 2002)
- September 11, 2001 (Sept 11, 2001 - Dec 1, 2001)
- Election 2000 (Aug 1, 2000 - Jan 21, 2001)
- Prototype (Library of Congress staff access only)

In production...

- Winter Olympics 2002 (Feb 8, 2002 - Feb 24, 2002)
- September 11th Remembrance (Sept 11-18, 2002)
- The 107th Congress (Dec 2002)
Election 2000 an Internet library.

**LC Control Number:** 2001561915

**Type of Material:** Book (Print, Microform, Electronic, etc.)

**Main Title:** Election 2000 [computer file] : an Internet library.

**Variant Title:** HTML source code title: Alexa Wayback machine, election 2000

**Published/Created:** San Francisco : Alexa Internet, [2001]

**Related Names:** Alexa Internet (Firm)
Library of Congress

**Summary:** Commissioned by the Library of Congress, presents a digital archive detailing the United States 2000 Presidential campaign and election with materials covering the period from August 2000 to January 2001.

**Contents:** Conservation district sites -- Democratic Party sites -- Government sites -- Green Party sites -- Humor & criticism sites -- League of Women Voters sites -- Libertarian Party sites -- News sites -- Other election 2000 sites -- Presidential candidate sites -- Reform Party sites -- Republican Party sites -- U.S. Congressman sites -- U.S. Senator sites -- Vote trading sites.

**Notes:** Title from home page as viewed on July 18, 2001.
"A publicly accessible repository of digital materials covering the U.S. Presidential Election of 2000."

**Subjects:** Presidents--United States--Election--2000
Political campaigns--United States
Contested elections--United States

**LC Classification:** JK526

**Dewey Class No.:** 324.973 13

**Geog. Area Code:** n-us---

**Quality Code:** pcc

**Electronic File Info.:** http://web.archive.org/collections/e2k.html
Travel back in time...
Use the Wayback Machine to view Election 2000 Web sites from the past.

The Presidential Candidate Sites

**George W. Bush**
For President
www.georgewbush.com

See www.georgewbush.com on Election Day:  
**Tues. Nov. 7, 2000**

See all archived dates...

**Gore Lieberman**
www.gorelieberman.com

See www.gorelieberman.com on Election Day:  
**Tues. Nov. 7, 2000**

See all archived dates...

**Ralph Nader For President**
www.votenader.com

See www.votenader.com on Election Day:  
**Tues. Nov. 7, 2000**

This collection was commissioned by the Library of Congress to archive digital materials covering the Election of 2000. It contains 800 gigabytes of data gathered from 8/1/2000 to 1/21/2001. For more information see the Press Release or Help.

See a directory of the 797 sites in this collection...
- Conservation District Sites (39)
- Democratic Party Sites (86)
- Government Sites (6)
- Green Party Sites (11)
- Humor & Criticism Sites (24)
- League of Women Voters Sites (11)
- Libertarian Party Sites (45)
- News Sites (38)
- Other Election 2000 Sites (167)
- Presidential and Party Candidate Sites (52)
- Reform Party Sites (63)
- Republican Party Sites (60)
- U.S. House Sites (51)
- U.S. Senate Sites (130)
- Vote Trading Sites (14)
# Democratic Party Sites

<table>
<thead>
<tr>
<th>13thdistrictdems.org</th>
<th>fsudemocrats.org</th>
<th>ohioedems.org</th>
</tr>
</thead>
<tbody>
<tr>
<td>3rdcd.vademocrats.org</td>
<td>hawaidemocrats.org</td>
<td>okdemocrats.org</td>
</tr>
<tr>
<td>8thcd.vademocrats.org</td>
<td>housedems2000.com</td>
<td>padems.com</td>
</tr>
<tr>
<td>akdemocrats.org</td>
<td>housedflcaucus.campaignoffice.com</td>
<td>salinecodemocrats.org</td>
</tr>
<tr>
<td>aladems.org</td>
<td>iaho-democrats.org</td>
<td>senatorflcaucus.campaignoffice.com</td>
</tr>
<tr>
<td>alaska.net</td>
<td>ildems.org</td>
<td>sbdemocrats.org</td>
</tr>
<tr>
<td>allencountydemocrats.org</td>
<td>indems.org</td>
<td>scdp.org</td>
</tr>
<tr>
<td>arkdems.org</td>
<td>iowademocrats.org</td>
<td>sddemocrats.org</td>
</tr>
<tr>
<td>arlingtondemocrats.org</td>
<td>kansassenedemocrats.org</td>
<td>sddp.org</td>
</tr>
<tr>
<td>azdem.org</td>
<td>kansashousedemocrats.org</td>
<td>skdemocrats.org</td>
</tr>
<tr>
<td>beachdemocrats.org</td>
<td>kansasyoungdemocrats.org</td>
<td>socialdemocrats.org</td>
</tr>
<tr>
<td>bexardemocrats.org</td>
<td>ksdp.org</td>
<td>stonewalldemocrats.org</td>
</tr>
<tr>
<td>charlestowndemocrats.org</td>
<td>kydems.campaignoffice.com</td>
<td>tejanodemocrats.org</td>
</tr>
<tr>
<td>chesapeakeodemocrats.org</td>
<td>lademo.org</td>
<td>tennesseedems.campaignoffice.com</td>
</tr>
<tr>
<td>cobbdemocrats.org</td>
<td>lubbockdemocrats.org</td>
<td>tndemocrats.org</td>
</tr>
<tr>
<td>coloradodemocrats.org</td>
<td>mainedems.org</td>
<td>traviscountydemocrats.org</td>
</tr>
<tr>
<td>columbiam.arkdemocrats.org</td>
<td>massdems.org</td>
<td>txdemocrats.org</td>
</tr>
<tr>
<td>ctdems.org</td>
<td>mi-democrats.com</td>
<td>union.arkdemocrats.org</td>
</tr>
<tr>
<td>dcdemocrats.org</td>
<td>missouridems.org</td>
<td>utdemocrats.org</td>
</tr>
<tr>
<td>deldemocrats.org</td>
<td>mtdemocrats.org</td>
<td>vademocrats.org</td>
</tr>
<tr>
<td>democrats.org</td>
<td>muscatinedemocrats.org</td>
<td>vayoungdemocrats.org</td>
</tr>
<tr>
<td>democratsabroad.org</td>
<td>ncdp.org</td>
<td>vtdemocrats.org</td>
</tr>
<tr>
<td>dems2000.com</td>
<td>nddempnl.campaignoffice.com</td>
<td>washingtoncountydemocrats.org</td>
</tr>
<tr>
<td>dfi.org</td>
<td>nebraskademocrats.org</td>
<td>web.democrats.org</td>
</tr>
<tr>
<td>dpnm.org</td>
<td>nh-democrats.org</td>
<td>windemocrats.org</td>
</tr>
</tbody>
</table>
Lesson Learned

• Some type of cataloging/descriptive metadata needs to be done to be able to search through the collection
Election 2002 Web Archive Search

Search Path: New Search

Search Options

Office
- Governor
- House
- Senate

Party
- Democratic
- Green
- Libertarian
- Other Third Party
- Reform
- Republican

State
- AK
- AL
- AR
- AZ
- CA
- CO
- CT
- DE
- FL
- GA
- GU
- HI
- ID
- IL
- IN
- KS
- KY
- LA
- MA
- MD
- ME
- MI
- MN
- MO
- MS
- MT
- NC
- ND
- NE
- NH
- NJ
- NM
- NV
- NY
- OH
- OK
- OR
- PA
- RI
- SC
- SD
- TN
- TX
- UT
- VA
- VI
- VT
- WA
- WI
- WV
- WY

Candidate Name
- 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

All Candidates

Search Instructions

Select one of the terms in the Search menu on the left to list candidates matching the selected category. The first 10 matching candidates will be displayed in the Results section on the right side of the screen. Use the page navigation below this list to view additional candidates.

Narrowing Your Search Selecting another category in the Search menu on the left will list candidates matching the additional term and previously selected terms. Candidates matching the refined terms will be presented in the Results section.

Viewing a Catalog Record Select a name to view the catalog record for that candidate. The record will open in a new browser window. Access to archived sites is through the catalog records.

Using the Search Path The Search Path, found directly above the Search menu, can be used to navigate and identify the layers of categories as you use the archive. Each time you select an additional category, it will be added to the Search Path. Previously selected categories remain in the Search Path and can be selected to return to previous iterations of the search.

Start a New Search Click on "New Search" in the Search Path.
MODS

- Metadata Object Description Schema
- Descriptive metadata standard
- Uses XML Schema
- A derivative of MARC using language-based tags but MARC semantics
- Element set is compatible with existing descriptions in large library databases
- Particularly applicable to digital objects
- Hierarchy allows for rich description, especially of complex digital objects
Why MODS

- Rich but not too rich descriptive metadata format; richer/more hierarchical than Dublin Core and simpler than MARC 21/MARCXML
- Alternative for emerging initiatives
  - Z39.50 Next Generation specified format
  - extension schema to METS
  - to represent metadata for harvesting (OAI)
  - As an interoperable format for convergence between MARC and non-MARC XML descriptions
Status of MODS

• Open listserv collaboration of self-selected possible implementors, LC coordinated (1st half 2002)
• First comment and use period: June – December 2002
• MODS version 3.0 now available
• Registration submitted to NISO and going through approval process
Fields Used in Election 2002

- Title
- Name (structured form)
- Abstract
- Date captured
- Genre (value always “Web site”)
- Physical description (file formats)
- Identifier (base URL)
- Language
- Access conditions/rights management
- Subject (keyword or LCSH if possible)
## Election 2002 Web Archive Search

**Search Path:** New Search  Category [Candidate] > Name - U (5)  Home

### Search Options
- **Office**
  - Governor
  - House

- **Party**
  - Democratic
  - Libertarian
  - Republican

- All Candidates

### Results
<table>
<thead>
<tr>
<th>Name</th>
<th>Geographic Area</th>
<th>Party</th>
<th>Office</th>
</tr>
</thead>
<tbody>
<tr>
<td>Udall, Mark</td>
<td>CO-2</td>
<td>Democratic</td>
<td>House</td>
</tr>
<tr>
<td>Udall, Steve</td>
<td>AZ-1</td>
<td>Democratic</td>
<td>House</td>
</tr>
<tr>
<td>Ulmer, Fran</td>
<td>AK</td>
<td>Democratic</td>
<td>Governor</td>
</tr>
<tr>
<td>Umphress, Dennis</td>
<td>CA-16</td>
<td>Libertarian</td>
<td>House</td>
</tr>
<tr>
<td>Upton, Fred</td>
<td>MI-6</td>
<td>Republican</td>
<td>House</td>
</tr>
</tbody>
</table>

**Number of records:** 5

1
Sample MODS record for Election 2002 Web site

- Fran Ulmer Web site record (XML)
Title: FranUlmer.com -- Home Page

Alternative Title: Fran Ulmer, Democratic Party candidate for Governor, Alaska, 2002.

Name: Ulmer, Fran

Abstract: Web site promoting the candidacy of Fran Ulmer, Democratic Party candidate for Governor, Alaska, 2002. Includes candidate biography, issue position statements, campaign contact information, privacy policy and campaign news press releases. Site features enable visitors to sign up for campaign email list, volunteer, make campaign contributions and follow links to other internet locations.

Date Captured: July 2, 2002 - November 30, 2002

 Archived Site

Subjects: Elections--Alaska
Governors--Alaska--Election
Democratic Party (AK)

Language: English

Genre: Web Site

Access Condition: None

Active Site: http://www.franulmer.com/

Collection Title: Election 2002 Web Archive
Election 2002 Web Archive Resource Page

Each date listed below represents one archived version of «http://www.franulmer.com/». Clicking any date will display the web site record as it appeared at that time. Following other links from that page will return the closest matches available from the Election 2002 Web Archive. The record will open in the same window. Please use the browser back button to return to this page if you wish to visit other archived dates.

194 archived captures for http://www.franulmer.com/

* denotes when site was updated.

<table>
<thead>
<tr>
<th>Search Results for 2002</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan-Feb</td>
</tr>
<tr>
<td>0 pages</td>
</tr>
<tr>
<td>Jul 02, 2002 *</td>
</tr>
<tr>
<td>Jul 03, 2002 *</td>
</tr>
<tr>
<td>Jul 03, 2002</td>
</tr>
<tr>
<td>Jul 04, 2002</td>
</tr>
<tr>
<td>Jul 05, 2002</td>
</tr>
<tr>
<td>Jul 07, 2002</td>
</tr>
<tr>
<td>Jul 09, 2002</td>
</tr>
<tr>
<td>Jul 10, 2002</td>
</tr>
<tr>
<td>Jul 11, 2002</td>
</tr>
<tr>
<td>Jul 12, 2002 *</td>
</tr>
<tr>
<td>Jul 13, 2002 *</td>
</tr>
</tbody>
</table>
107th Congress

- Enhanced descriptive metadata (MODS) created at LC
- Include authoritative forms of names, subject headings, and classification numbers
- Registered handles for archived sites
- Plan to use LC created tools for METS creation, record input and search/browse
  - Example: http://www.loc.gov/ihas
MODS record

- Joint Economic Committee XML record
Lessons learned (cataloging)

- More accurate, usable cataloging if we do in-house
  - We understand the way the data will be used
- Good metadata results in good searching
- Simple records can be created in minimal time using MODS
- Flexibility of XML allows for options in display
Interface & Access

• IA & WebArchivist.org interfaces: Election 2000, September 11, Election 2002
• LC moving to refine, improve, develop interfaces in-house
• Access to available collections through MINERVA home page and through ILS collection level records
• Transferring access to archive to on-site
METS/MODS for Minerva

• Experiment with METS objects for Web sites
• MODS for descriptive metadata
  – Hierarchy in relatedItem powerful for multiple captures and linked pages
  – Works well with METS structMap
  – relatedItem type="host" link to aggregated resource
  – relatedItem type="constituent" for linked pages
  – Allows for descriptive metadata at lower level and facilitates display of object
• Developing tools for capture of some technical metadata
• Developing METS profile for Web sites