A Partnership Born of Urgency and Civic Responsibility
Preserving Access to Government Websites Through the CyberCemetery

Starr Hoffman
Librarian for Digital Collections
University of North Texas Libraries
22 April 2010
2010 AGA Regional Professional Development Conference

Presentation Overview
• Intro: What is the CyberCemetery?
• Purpose: Why create a CyberCemetery?
• Development
• Archiving Process
• Technical Details
• Users by Country
• Types of Content
• Using the CyberCemetery
• Other Resources
• Conclusion

What is the CyberCemetery?
• online archive of websites from U.S. government agencies or commissions that are no longer operating

http://govinfo.library.unt.edu

What is the CyberCemetery?

http://govinfo.library.unt.edu
What is the CyberCemetery?

- online archive of websites from U.S. government agencies or commissions that are no longer operating
- maintained by the University of North Texas Libraries
- freely accessible world-wide

http://govinfo.library.unt.edu

CyberCemetery vs. Dot Gov Harvest

- Partners: UNT, GPO, NARA
- archive of websites from U.S. government agencies or commissions that are no longer operating
- “dead” websites (no longer hosted or maintained by the government)
- currently live & useable
- purpose:
  + to preserve “dead” government websites and provide permanent public access

- Partners: LC, IA, UNT, others
- archive of government website “snapshots” from key time periods (i.e., before/after an administration change)
- will include snapshots of many still-live websites archived, but not currently “live”
- purpose:
  + to preserve a record of government web presence during specific time periods and administrations
  + to track changes in government websites over time

Why Create the CyberCemetery?

- At-Risk Information:
  - 1990’s: U.S. government information moved online
  - much of it born-digital
  - often edited or removed without warning

- Federal Depository Library Program (FDLP)
  - mission:
    - to provide free, permanent public access to government information
    - online information complicates this mission
    - administered by the U.S. Government Printing Office (GPO)
    - UNT = federal depository library

Development

- 1995
  - report from Government Printing Office (GPO):
    - need to preserve electronic government publications

- 1997
  - UNT & GPO discuss a partnership
  - UNT archives ACIR website
    - (Advisory Commission on Intergovernmental Relations)
Development

1999
- UNT/GPO partnership expanded
  - permanent public access
  - multiple government websites
  - government agency or commission which is no longer operating
    - (and/or has issued a final report)
- Collection named “CyberCemetery”
  - websites from “dead” government agencies and commissions

Development

2006
- UNT/GPO partnership expanded
  - U.S. National Archives and Records Administration (NARA)

Archiving Process

Identify at-risk government agencies and commissions
- read/listen to the news
- online queries targeting keywords (i.e., “final report”)
- read government-related websites and blogs
- referrals from other librarians
- contacted by GPO
- contacted directly by the agency/commission

Archiving Process

Evaluate the website
- official government website
  - agency or commission must:
    - be closing
    - issued a final report
    - other indication that the website is at-risk
Archiving Process

Evaluate the website (continued)
+ Questions for website administrator:
  1. What operating system was used to host this website?
  2. What webserver software was used for the hosting of this website?
  3. Are server-side includes (SSI) used in this website?
  4. Was this website static HTML or a dynamic site?
  5. If dynamic, what scripting languages were used for this website (php, perl, python)?
  6. Was a database used for this website?
     1. If so, what database was used for this website?
     2. What methods were used to connect to the database?
  7. Is there streaming media associated with this website?
  8. Are there proprietary content types used in this website?
  9. Are there any comments you would like to add?

Archiving Process

Harvest the website
+ Past method: HTTrack
  - http://www.httrack.com/
  - user interface:
    - UNT’s Digital Collections website
+ Current method: Heritrix
  - http://crawler.archive.org/
  - ARC files
    - website in a single file: 100 – 600MB
  - user interface:
    - Internet Archive’s Wayback Machine

Archiving Process

Harvesting alternative: Donated content
+ directly receive files from agency or commission
  + Why donated content?
    - If content cannot be accessed by harvesting
    - flash video, large amounts of media
  + Why not donated content?
    - Content could be altered
    - Harvesting = exact copy of online published content

Archiving Process

Link Checking
+ Automated:
  - Xenu Link Checker
    - http://home.snafu.de/tilman/xenulink.html
  - compare reports of original and archived sites
+ Manual:
  - manually navigate original and archived sites
Archiving Process

Archive Preparation (previous method)
- add text “Archive”
- 8 point, Times New Roman font
- added to top/center of each page
- manually disable contact links
- “mail to” links
- submit-able forms

(Heritrix makes these preparations unnecessary.)

Archiving Process

Load to UNT Server
- Upload archived website
- Add navigation
- Notify GPO (or agency/commission) that archived version is live

Technical Details

Equipment
- Four servers (three as backup)
- Four node fail-over clustered configuration
- SAN volume
- 27.2GB of content on 40GB server

Environment
- Library basement
- 38 Fahrenheit (3 Celsius)
- 50% humidity

Backup
- Full backups to magnetic tape
- Performed each weekend
- Shipped to offsite storage company
- Iron Mountain
- http://www.ironmountain.com
Where Are Our Users?

- United States
- Other/Latin America
- Continental Europe
- Asia
- Canada
- United Kingdom

9%
5%
12%
16%
10%

Types of Content

- web files (HTML, XML)
- text documents (.txt, .pdf, .doc)
- spreadsheets & statistical information (.xls)
- presentations (.ppt)
- media files:
  - images & photographs (.jpg, .gif, .png, .tiff)
  - audio (.mp3)
  - video (.wm, .mov, .sp)

Using the CyberCemetery

http://digital.library.unt.edu/explore/collections/GDCC/

Navigating

- browse by:
  - title
  - date of expiration
  - government branch
Navigating

- main search box
  - all CyberCemetery content at once
  - National Partnership for Reinventing Government
  - Office of Technology Assessment
  - 9/11 Commission

Other Resources

- Congressional Research Reports
  - research specialists at Library of Congress
  - topics relevant to pending legislation
  - high-quality, non-biased information
  - created for members of Congress
  - not typically publicly available
  - +10,000 reports available

http://digital.library.unt.edu/
Ask us!

- [http://www.library.unt.edu/govinfo](http://www.library.unt.edu/govinfo)
  - phone: (940) 565-2870, main desk
  - email: govinfo@unt.edu

Government Documents Dept. Service Desk Hours

<table>
<thead>
<tr>
<th>Day</th>
<th>Open</th>
<th>Close</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunday</td>
<td>1:00 PM</td>
<td>9:00 PM</td>
</tr>
<tr>
<td>Monday - Thursday</td>
<td>9:00 AM</td>
<td>9:00 PM</td>
</tr>
<tr>
<td>Friday &amp; Saturday</td>
<td>11:00 AM</td>
<td>6:00 PM</td>
</tr>
</tbody>
</table>

Conclusion

- permanent public access
- archived government information
- freely, globally available
- partnership:
  - University of North Texas Libraries
  - U.S. Government Printing Office
  - National Archives and Records Administration

Contact Information

- [http://govinfo.library.unt.edu](http://govinfo.library.unt.edu)
- [http://digital.library.unt.edu/explore/collections/GDCC/](http://digital.library.unt.edu/explore/collections/GDCC/)

download this presentation:
  - [http://geekyartistlibrarian.wordpress.com](http://geekyartistlibrarian.wordpress.com)

- Starr Hoffman
  Librarian for Digital Collections
  Government Documents Department
  University of North Texas Libraries
  starr.hoffman@unt.edu
  [http://geekyartistlibrarian.wordpress.com](http://geekyartistlibrarian.wordpress.com)
  940.565.4150