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?

The CyberCemetery is an archive of government web sites that have ceased operation. Usually these sites are federal government agencies and commissions that have issued a final report. This collection features over 3000 sites that cover topics supporting the university's curriculum and special program strengths. This collection supports the mission of the UNT Libraries by acquiring, preserving and providing knowledge in the form of government websites.

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.

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

1997 - present

- **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

2008 - present

- **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)
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)
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
Evaluate the website

- official government website
- agency or commission must:
  - be closing
  - issued a final report
  - other indication that the website is at-risk
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?
   1. If dynamic, what scripting languages were used for this website (php, perl, python)?
   2. 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?
5. Is there streaming media associated with this website?
6. Are there proprietary content types used in this website?
7. Are there any comments you would like to add?
Archiving Process

Harvest the website

- Past method: HTTrack
  - user interface:
    - UNT’s Digital Collections website

- Current method: Heritrix
  - 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](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°F Fahrenheit (3°C Celsius)
  - 50% humidity
Technical Details

• 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: 79%
- Asia: 10%
- Canada: 5%
- Continental Europe: 4%
- Other/Unknown: 2%
- United Kingdom: 1%
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, .rp)
Using the CyberCemetery

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

The CyberCemetery is an archive of government web sites that have ceased operation (usually web sites of defunct government agencies and commissions that have issued a final report). This collection features a variety of topics indicative of the broad nature of government information. In particular, this collection features web sites that cover topics supporting the university's curriculum and particular program strengths.

Background

Foreseeing the potential preservation problems created by federal agencies' ventures into electronic publishing, UNT became the second Depository Library in the nation to join the Federal Depository Library Program Content Partnerships Program. This program attempts to ensure permanent public access to electronic federal information. As a participant, the UNT Depository Library was designated the host of the permanent online collection of the National Advisory Commission on Intergovernmental Relations (NACIR).

We've since expanded our Content Partnership with the federal government to include dozens of other defunct federal agency web sites. This electronic repository is popularly known as the CyberCemetery. In recognition of our work in this area, the UNT Libraries were designated an Affiliated Archives of the National Archives in 2006. Under this agreement, the UNT Libraries will continue to preserve and provide access to the records of defunct government web sites, while the National Archives and Records Administration (NARA) will legally accession the records as part of the Archives of the United States and will join the UNT Libraries and the Government Printing Office (GPO) in ensuring the preservation of these valuable records. We are currently the only non-military educational institution to be designated an Affiliated Archives.

Source: Kelly, Norely (Specialist: Our History: Celebrating 65 Years of Public Service. Accessed March 31, 2010.)
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 publically available
  - +10,000 reports available

http://digital.library.unt.edu/
Other Resources

- UNT Digital Library
  - digitizing our “legacy” collection of government documents
    - “A-Z Digitization Project”
  - FCC Record
    - (FCC Report = future project)
  - U.S. Agricultural Experiment Station Record
  - OTA documents
  - ACIR documents

http://digital.library.unt.edu/
Other Resources

- get updates via RSS
- example feed:
  - feed://digital.library.unt.edu/explore/collections/ATOZ/feed/

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(s)</th>
<th>Opens</th>
<th>Closes</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>4:00 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