Preserving Access to Government Websites

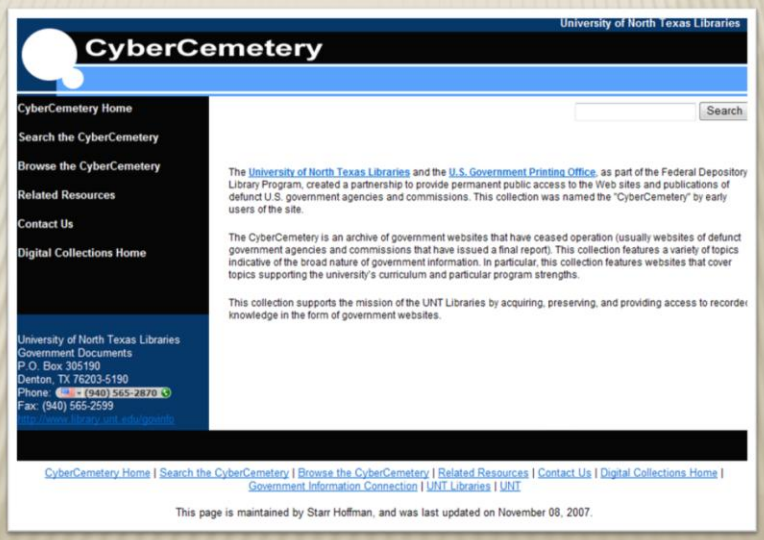
Development and Practice in the CyberCemetery

Starr Hoffman
Librarian for Digital Collections
University of North Texas Libraries
13 August 2008

PRESENTATION CONTENTS

- Intro: *What is the CyberCemetery?*
- Purpose: *Why create a CyberCemetery?*
- Development
- Archiving Process
- Technical Details
- Users by Country
- Conclusion

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

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

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)

  - administered by the U.S. Government Printing Office (GPO)
  - **mission:** to provide free, permanent public access to government information
  - online information complicates this mission
  - University of North Texas is a federal depository library

DEVELOPMENT

- 1995
  - Government Printing Office (GPO) publishes report stating need to preserve electronic government publications

- 1997
  - University of North Texas (UNT) talks to GPO about forming a partnership
  - UNT archives website of the Advisory Commission on Intergovernmental Relations

DEVELOPMENT

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

DEVELOPMENT

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

ARCHIVING PROCESS

1. Identify at-risk government agencies and commissions
   - read/listen to the news
   - read government-related websites and blogs
   - get referrals from other librarians
   - contacted by GPO
   - contacted directly by the agency/commission

ARCHIVING PROCESS

2. **Evaluate the website**
   - must be an official government website
   - The agency or commission must:
     - Be closing
     - Have issued a final report
     - Other indication that the website is at-risk

ARCHIVING PROCESS

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

ARCHIVING PROCESS

3. Harvest the website
   - Past method: HTTrack
     - http://www.httrack.com/
   - Future method: Heritrix
     - http://crawler.archive.org/
     - ARC files
     - website in a single file: 100 – 600MB

ARCHIVING PROCESS

3. 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

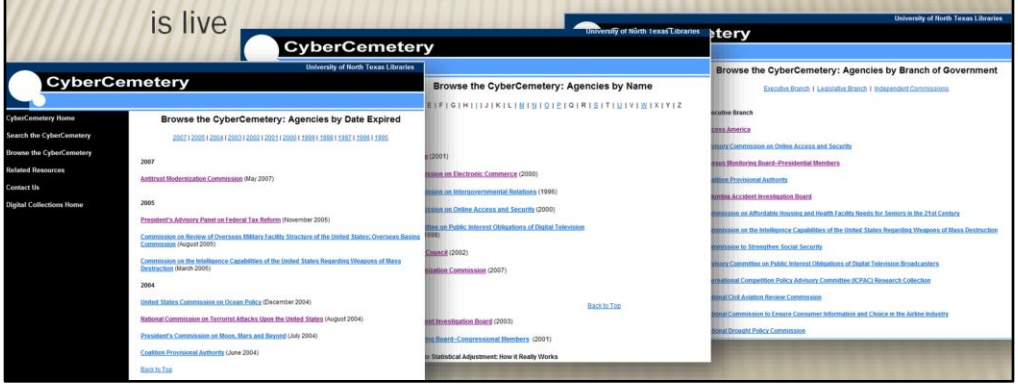
4. Link Checking
   - Automated:
     - Xenu Link Checker
     - http://home.snafu.de/tilman/xenulink.html
     - compare reports of original and archived sites
   - Manual:
     - manually navigate original and archived sites

ARCHIVING PROCESS

5. Archive Preparation
   - 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 will make these preparations unnecessary)

ARCHIVING PROCESS

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



TECHNICAL DETAILS

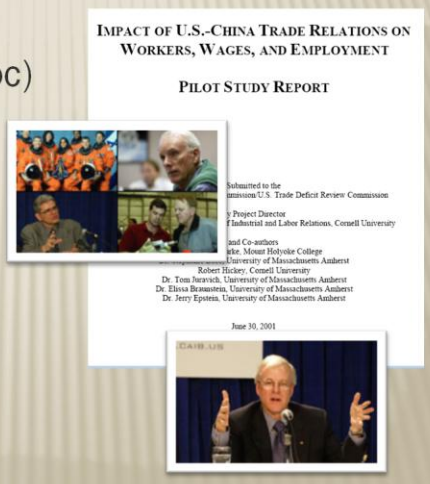
- Equipment
  - Four servers (three as backup)
  - Four node fail-over clustered configuration
  - SAN volume
  - 27.2GB of content on 40GB server
- Environment
  - Library basement
  - 38° Fahrenheit (3° Celsius)
  - 50% humidity

TECHNICAL DETAILS

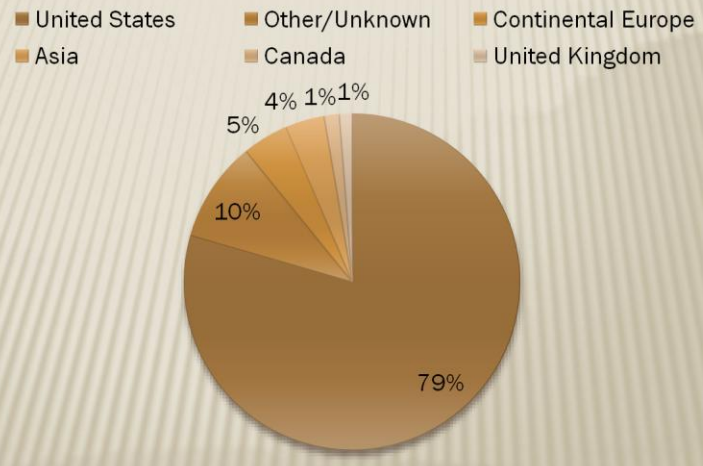
- Backup
  - full backups to magnetic tape
  - performed each weekend
  - shipped to offsite storage company
    - Iron Mountain
    - http://www.ironmountain.com

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)



WHERE ARE OUR USERS?



CONCLUSION

- provides **permanent public access**
- archive of government information
- **freely, globally available**
- 45 websites and growing
- **partnership** between:
  - University of North Texas Libraries
  - U.S. Government Printing Office
  - National Archives and Records Administration

FOR FURTHER INFORMATION:

- http://govinfo.library.unt.edu
- http://www.library.unt.edu/govinfo/digital-collections

- Starr Hoffman

  Librarian for Digital Collections
  University of North Texas Libraries
  starr.hoffman@unt.edu
  http://geekyartistlibrarian.blogspot.com