CyberCemetery

Preserving At-Risk Government Web Content
Background of CyberCemetery

Robbie Sittel
The University of North Texas

• A public university established in 1890 and located in Denton, TX

• UNT became a congressionally-designated selective depository library in January 1948.

• In 1948, the college served a student body of 4,668 with 265 faculty and 14 librarians.

• In 2015, UNT serves over 36,000 students with 6,700 faculty and staff, including 57 faculty librarians.
Preserving electronic information

• 1993 Congress passes the U.S. Government Printing Office Electronic Information Access Enhancement Act (Public Law 103-40)
• 1994 UNT named as GPO Gateway Access Library
• 1995 GPO publishes report stating need to preserve electronic government publications
• 1997 UNT signs MOU with GPO and acquires its first website of a defunct government agency – Advisory Commission on Intergovernmental Relations (ACIR)
• 1999 GPO and UNT expands agreement to include additional websites and the CyberCemetery is born
Partnerships in preserving electronic information

- 1994: GPO Gateway Access Library
- 1999: MOU with GPO supporting CyberCemetery
- 2000: FDLP Content Partner
- 2000: Texas Agency Content Partnership
- 2006: Affiliated Archives
- 2011: expanded services to the Texas State Library & Archives Commission
- 2015: FIPNet (Federal Information Preservation Network)
What is the CyberCemetery?

- An online archive of websites from U.S. government agencies or commissions that are no longer active or in operation
- A final point-in-time snapshot of the online presence of the agency
- A way to preserve sun setting, final or at-risk digital-born materials
What the CyberCemetery is NOT

• Point in time snapshots of government websites
• “Live” websites
• Examples of versioning
• A record of a government website under a specific administrator or at a specific time
Why the CyberCemetery?

• Freely accessible resource
• A natural extension of the mission of the Federal Depository Library Program: to provide free, permanent access to government information
• Preserves at-risk, born-digital material
• Directly supports the mission of the UNT Libraries
• Reaches a broad user population on the UNT campus and around the world
• Provided the beginnings of what is now an ever-expanding digital collection
Harvesting and Technology

Lauren Ko
Early CyberCemetery

- First site: The U.S. Advisory Commission on Intergovernmental Relations (1996)
- Transfer of files via physical media (e.g. zip drive)
- Other early sites captured with HTTrack
Expanding Websites

• Sites have grown
• Frequently there is a database backend
• More subdomains
• Embedded media, etc.
Formation of the IIPC

- International Internet Preservation Consortium (IIPC) formed 2003
- International community built around Web Archiving
- Membership of national libraries, Internet Archive, increasing number of university libraries
- Develop tools, share workflows, standardize practices
- Developed WARC standard
WARC

• WARC (Web ARChive) file format
• Format became an ISO standard (ISO 28500) in 2009
• Store downloaded resource plus info
• WARC files are made of WARC records concatenated together (often compressed)
• For a downloaded resource we write 3 records: response, request, metadata
• For a CyberCemetery archived site, we package the WARC files together for storage in our repository
Heritrix

- Crawler (written in Java) developed by the Internet Archive
- Configure job for each Web crawl
- Writes WARC s
OpenWayback

• Web application (Java) initially developed by Internet Archive as Wayback Machine
• Now developed by IIPC as OpenWayback
• Replays archived website’s WARC files
• Give it index of resources in WARC files and an index of where to find those files
• Rewrites URLs within page to pull from the archive
• Inserts message about being an archived website
Archiving Process for CyberCemetery

- Get a request from closing agency or commission
- Run test crawl with Heritrix
- Stage the archived site in OpenWayback for QA
- Reconfigure/add seeds for supplemental crawl
- More QA
- Use this configuration to run final crawl
- Do another round of QA and supplemental crawling
- Process into our repository
- Put into our production OpenWayback
- Make a metadata record for UNT Digital Library CyberCemetery collection
Scoping a Crawl

- Configure a crawl with list of scope rules
- Get everything on certain subdomains
- Ignore certain paths
- Hops from seed: try to get everything relevant
- Are external links in page relevant?
- External media
- Example: http://www.ed.gov/about/bdscomm/list/mathpanel/index.htm
  vs.
  http://www.jamestown2007.org/
Preventing Terrorism: Assessing the Nation’s Progress

In its report to the American people in 2004, the 9/11 Commission warned that “the greatest danger of another catastrophic attack in the United States will materialize if the world’s most dangerous terrorists acquire the world’s most dangerous weapons.”

Congress heeded that warning in 2007 when it established the Commission on the Prevention of Weapons of Mass Destruction Proliferation and Terrorism. Republicans and Democrats alike recognize that stopping terrorists from obtaining nuclear and biological weapons must be an urgent national priority.

The mandate of our Commission is to build on the work of the 9/11 Commission and complete a critical task: to assess our nation’s progress in preventing the proliferation of weapons of mass destruction and terrorism, and to provide a roadmap to greater security with concrete recommendations for improvement.

We’re examining the government’s current policies and programs, identifying gaps in our government’s prevention strategy and recommending ways to close them when we issue our report in December.

Since May, we have interviewed over 200 experts inside and outside of government. We have met with counterterrorism and intelligence officials here at home and abroad who are working to stop proliferation and terrorism.

Seventeen years after the tragic events of September 11, 2001, the threat of a terrorist attack on the United States is still very real. We are not yet safe. There is more that can and must be done. The security of the United States—and the world—depends on it.
Web Archiving Issues

• Web archiving mostly works but...

• Difficult to capture third party media (e.g. YouTube, Vimeo)
  • Help for this: feed extra URLs to Heritrix with helper tools like Umbra or other download helpers

• URLs not rewritten at replay time in OpenWayback

• Third party embedded media doesn’t replay

• Search/forms no longer work
Access to CyberCemetery

- http://cybercemetery.unt.edu
- Records in UNT Digital Library
- Some agencies/commissions want to retain original domain
The CyberCemetery is integrated into the same digital Library infrastructure used by the UNT Digital Library. This includes other Federal Government Documents. Including technical reports, historical publications, and other government information collected from the Web.
The CyberCemetery is a collection in the UNT Digital Library.

Each collection provides contextual information about the collection as well as access to stats, information about APIs for the collection and links to recent items added to the collection.
Users can search and browse the metadata records for each of the CyberCemetery sites via the collections “Explore” interface.

Facets on the left side of the screen help users refine their search.

Currently there are 100 sites in the CyberCemetery collection.
After a site has been crawled, a screenshot of the site is taken and a descriptive metadata record for the entire site is created.

This metadata record includes creation dates, harvesting dates and other information important to the acquisition of the site.

A unique ARK identifier is assigned to the site when it is ingested into the UNT Digital Library, this is used to track and record preservation metadata about the site over time.
All CyberCemetery sites are added to the Coda repository.

The Coda repository manages the archival WARC files and acts as the UNT Libraries Digital Preservation Environment.

Each site is packaged as a BagIt Bag and serialized with the UNTL Archival Information Package METS Profile.

Once in Coda, the digital object is replicated to another instance of repository that located at UNT’s Discovery Park research campus north of Denton.
PREMIS Events are logged for ingest, fixityCheck, and replication events.

All associate with the original digital object using the unique ARK identifiers.

A copy of this digital object is used as the input content for the OpenWayback system that provides end user access to a site in the CyberCemeter.
Other Web Archiving Activities

- The CyberCemetery complements the other Web Archiving activities at UNT
- Members of the International Internet Preservation Consortium (Steering Committee Members)
- Worked with the Internet Archive to extract and acquire the 110TB Web archive of .gov content from 1996-2013
Preserving Government Information

• The CyberCemetry is one of the many projects the UNT Libraries is involved with to preserve government information and make it available to the public

• Born-Digital publications such as governmental Websites continue to be some of the most “at risk” resources.
Thank You

Questions?

- Robbie Sittel – roberta.sittel@unt.edu
- Lauren Ko – lauren.ko@unt.edu
- Mark Phillips – mark.phillips@unt.edu