Web-Archiving: Collecting and Preserving Important Web-based National Resources

Mark Phillips
Dr. Daniel Gelaw Alemneh
University of North Texas
UNT Libraries
February 18, 2015
The Web is the platform for communication for the 21st century.
Each year more and more content is created digitally.
And published to the world via the Web.
Ensuring the long-term preservation and access to these resources offer new challenges and opportunities to memory organizations such as libraries, museums, and archives.
Background:
Web archiving is the process of collecting portions of the World Wide Web to ensure the information is preserved in an archive for future researchers, historians, and the public.
It involves concepts that are core to libraries and archives.
Selection: Defining the scope of the Web for preservation.
National Domain
Federal Domain
State/Municipal Web
Subject Based
Event Based
Acquisition: Harvesting selected content by using Web crawlers such as Heritrix
Access: Enabling interested users to accessing Web resources from the past.
Preservation: Ensuring the long-term access to these Web resources through systematic planning and curation.
These concepts enable memory organizations to align Web archiving with traditional collections, processes and missions.
Examples:
UNT Libraries CyberCemetery
CyberCemetery

What's Inside this Collection.

About This Collection

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 Libraries, part of the Federal Depository Library Program (FDLP), were designated the host of the permanent online collection of the defunct Advisory Commission on Intergovernmental Relations (ACIR).
Archive of Federal agencies, commissions, committees and initiatives that are no longer active.
Started in 1997 with the Advisory Commission on Intergovernmental Relations (ACIR)
Today contains over 100 Websites
UNT Libraries is an Affiliated Archives with the U.S. National Archives and Records Administration (NARA) because of the CyberCemetery
End of Term Presidential Web Archives
Good Evening

Welcome to the White House

Social Security Administration Announces
New Online Retirement Planner

The President & Vice President:
Their accomplishments, their families, and how to send them electronic mail --

Interactive Citizens' Handbook: Your guide to information about the Federal government

White House History and Tours:
Past Presidents and First Families, Art in the President's House and Tours -- Tour

Commonly Requested Federal Services:
Direct access to Federal Services

What's New:
What's happening at the White House
- Expanding Opportunities for America's Senior Citizens
- A National Call to Action to Close the Digital Divide

Site News:
Featured sections of our site
- White House Easter Egg Roll
- China Trade Relations Working
Secretary Snow Discusses Economy on "Ask the White House"
Treasury Secretary John Snow on Tuesday discussed the economy on "Ask the White House". Read the Full Transcript full story

Steady Job Growth Continues in States Across the Country
June’s employment data indicate that the President’s economic policies are creating jobs and driving steady economic growth, but there is still more work to be done. full story

President Bush Welcomes President Lagos of Chile
President Bush and Chilean President Lagos discussed trade issues on Monday at the White House. full story

Press Briefing by Scott McClellan
Watch Monday's Press Briefing with White House Press Secretary Scott McClellan. full story

President Bush Welcomes Malaysian Prime Minister to White House
President George W. Bush waves to the troops at Fort Campbell, Ky., Tuesday, Nov. 25, 2008, after being introduced on stage during a Thanksgiving visit. Said the President, “Over the past seven years, folks from this base have done exactly what they were trained to do.” “You have taken the battle of the terrorists overseas so we do not have to face them here in the United States. You have helped counter the hateful ideology of tyranny and terror with a more hopeful vision of justice and liberty. You’re part of the great ideological struggle of our time. With the soldiers of Fort Campbell out front, the forces of freedom and liberty will prevail.” White House photo by Eric Draper

President Bush Pardons “Pumpkin and Pecan” in National Thanksgiving Turkey Ceremony

President Bush on Wednesday said, “This is my final Thanksgiving as the President. Over the past eight years, I have been given many reasons to be thankful. I'm thankful to our men and women in uniform and I am incredibly proud to have been their Commander-in-Chief. I am thankful for the armies of compassion volunteers who feed the hungry and shelter the poor. I am thankful...”
America has never been about what can be done for us. It’s about what can be done by us, together, through the hard and frustrating but necessary work of self-government. That’s the principle we were founded on.

— President Barack Obama, November 7, 2012

JOIN PRESIDENT OBAMA

TOP NEWS

November 11, 11:55am
Thank an American Hero

November 10, 5:45am
Weekly Address: Extending Middle Class Tax Cuts to Grow the Economy
Collaboration between institutions throughout the United State to capture the Federal (.gov) Web at the end of each presidential term.
Selection, harvesting, archiving distributed to meet local collection needs.
Centralized access
UNT Libraries has acquired archive for 2008 and 2012 resulting in a 16TB and 30TB dataset
Extracted .gov dataset from Internet Archive
Collaboration between UNT and the Internet Archive in 2014 to extract the .gov data from the global Wayback instance at the Internet Archive
109 TB dataset
Using Web Archives in scholarship, teaching and research
Collaborating Internationally
International Internet Preservation Consortium
(netpreserve.org)
A GLOBAL NETWORK OF EXPERTS ARCHIVING THE WEB FOR FUTURE GENERATIONS.

Learn More About IIPC

WHY ARCHIVE THE WEB?

Information published on the World Wide Web today will be the primary resources for future
51 member institutions
Members from 25 countries: North America, Europe, Asia, Africa
National Libraries
National Archives
Research Libraries
Government Agencies
Private Companies
Common goal to preserve the Web
Collaboratively develop tools for the community such as the Heritrix Web crawler and the Open Wayback system for replay of Web archives.
Lead work on the formalization of the WARC File Format (ISO 28500:2009)
Web archiving is an important component in preserving a national heritage for future generations.
Thank you.
አመሰግናለሁ

Mark Phillips
mark.phillips@unt.edu

Dr. Daniel Gelaw Alemneh
daniel.alemneh@unt.edu