Web Archiving and Distributed Preservation

The UNT Libraries is among an elite group of research institutes and U.S. libraries internationally recognized for enterprising work in Web harvesting and archiving. Featuring a robust, state-of-the-art digital archive infrastructure, the Libraries can access, store, manage and transfer large quantities of digital information to accommodate a variety of complex research needs. Since 1997 the UNT Libraries has been at the forefront of broad harvest web captures and has played an instrumental role in capturing and organizing the US Government's web presence in order to preserve a historical record of its administrations. As web-based publishing increases, the Libraries' projects have significant consequences for the general public, scholars, historians and librarians worldwide — both for the preservation of important historical documents that would otherwise be lost and for setting precedents in capturing, classifying and accessing massive Web collections — hundreds of millions of files —using the latest tools and strategies.

- **International leader of innovative web harvesting and archival research**
- **The UNT Libraries hosts federal and state Web sites (since 1997) with a 20 TB Web archive**
- **Robust, state-of-the-art digital archive infrastructure with flexible storage capacity to handle complex, massive data**
- **Technology applications include link analysis and specialized programs to understand and visualize connections between government agencies**
- **Sophisticated classification software identifies metrics and enables expert characterization of materials for extensive collection building**
- **Develop advanced, distributed preservation models for small and large archives and protocols for the submission, dissemination and transfer of data**
- **Founding member of the National Digital Stewardship Alliance (NDSA); and elected member of the Steering Committee of the International Internet Preservation Consortium (IIPC)**

Special Collections

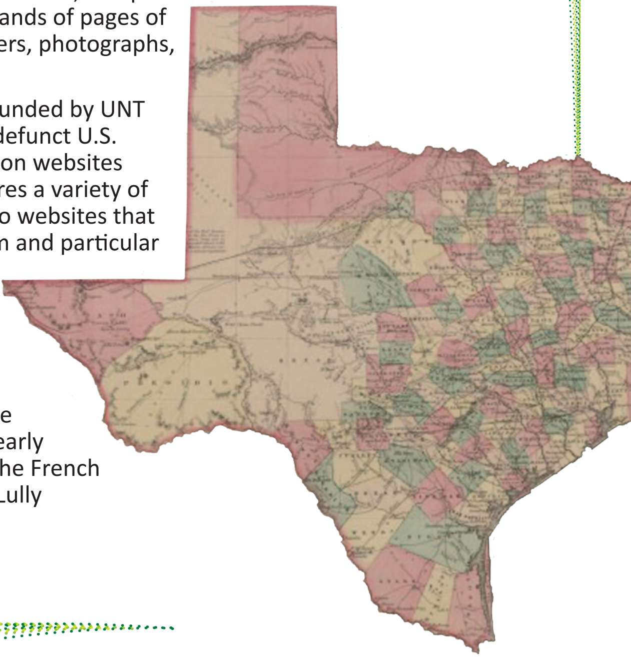
The University Libraries house collections of over 6 million catalogued items in a variety of formats, located in five libraries in five separate facilities, and provide electronic access through its website, including:

The **Congressional Research Service Reports Archives** is a large, searchable database of documents that primarily address national security, foreign policy and related topics.

The **Portal to Texas History** is a gateway to Texas history, from prehistory to the present day, where one can explore material from Texas libraries, museums, archives, historical societies, genealogical societies, and private family collections, including thousands of pages of digitized historical Texas newspapers, photographs, and rare, historical maps.

The **CyberCemetery collection**, founded by UNT Libraries in 1997 as an archive of defunct U.S. government agency and commission websites that have ceased operation, features a variety of government-related topics but also websites that support the university's curriculum and particular programs.

The UNT Libraries offer the **largest audio archive in the Southwest** as well as the Jean-Baptiste Lully Collection — a multimedia thematic catalog of the UNT Music Library's collection of early editions of operas and ballets by the French Baroque composer Jean-Baptiste Lully (1632-1687).



Representative Faculty

Martin Halbert, Dean of UNT Libraries; and Associate Professor of Library and Information Sciences: *digital preservation; metadata aggregation and organization; and scholarly portal design and digital curation*

Cathy Hartmann, Associate Dean of UNT Libraries: *digital preservation and collections; metadata; Web archiving and user access to digital assets; and electronic government information*

Rada Milhalcea, Associate Professor of Computer Science and Engineering: *natural language processing and information retrieval*

William Moen, Director of the Texas Center for Digital Knowledge; Associate Dean of the College of Information; and Associate Professor of Library and Information Sciences: *metadata; digital repositories; interoperability; and technical standards*

Mark Phillips, Assistant Dean of UNT Libraries: *web archiving; repository architectures; digital preservation; and the development of local infrastructure for digital content*

External Partners and Collaborators

Library of Congress

www.loc.gov/index.html

The Library of Congress is the nation's oldest federal cultural institution and serves as the research arm of Congress. It is also the largest library in the world, with millions of books, recordings, photographs, maps and manuscripts in its collections. Its mission is to support the Congress in fulfilling its constitutional duties and to further the progress of knowledge and creativity for the benefit of the American people.

Internet Archive

www.archive.org

The Archive is a non-profit library organization that offers access for researchers, historians, scholars, and the general public to historical collections that exist in digital format. It includes texts, audio, moving images, and software as well as archived web pages, and provides specialized services for persons with disabilities.

Institute of Museum and Library Services

www.imls.gov

The Institute is the primary source of federal support for the nation's 123,000 libraries and 17,500 museums, with a mission to create strong libraries and museums that connect people to information and ideas. The Institute works to sustain heritage, culture, and knowledge; enhance learning and innovation; and support professional development.

IIPC: International Internet Preservation Consortium

netpreserve.org/about/index.php

Open to libraries, archives, museums and cultural heritage institutions around the world, the IIPC acquires, preserves, and makes accessible knowledge and information from the Internet for future generations, promoting global exchange and international relations.

MetaArchive Cooperative

www.metaarchive.org

The MetaArchive is a digital repository and collaborative network of libraries, archives, and other cultural memory organizations that aims to provide low cost, high impact solutions for the care and preservation of digital materials.

TxCDK: Texas Center for Digital Knowledge

txcdk.unt.edu

TxCDK is a research, development, and consulting service enterprise that brings together researchers from multiple disciplines to enhance the relationships between individual knowledge workers and the technology-based environments in which they work.

Contributing Research Cluster:

KDDI: Knowledge Discovery from Digital Information
kddi.unt.edu