



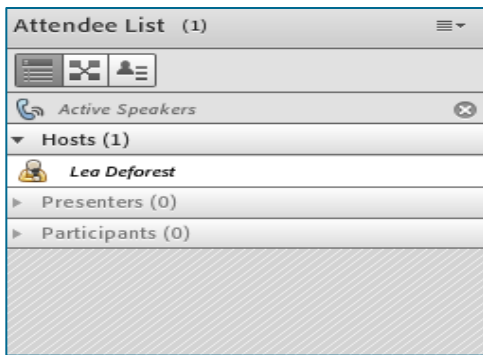
Web Archiving in the Library: Policies, Procedures, and Program Integration

Part 3 of 3:

Introduction to Web Archiving Texas (WATX19) Webinar Series

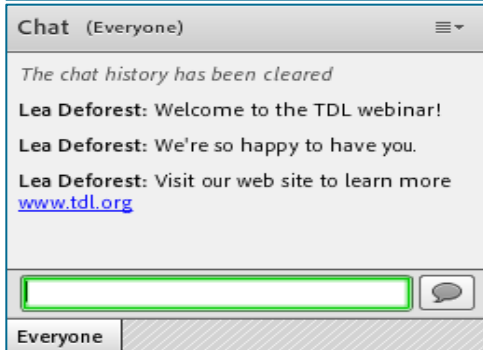
tdl.org

Using Adobe Connect



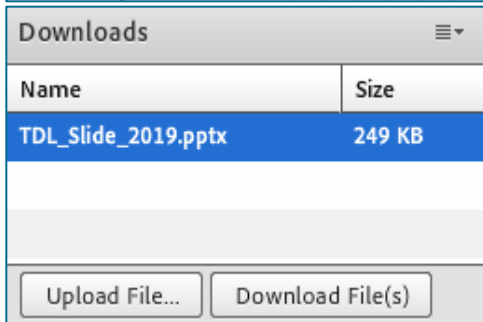
The Attendee List pod shows a header with a menu icon and three icons (list, grid, people). Below is a section for 'Active Speakers' with a close button. Underneath is a 'Hosts (1)' section listing 'Lea Deforest' with a user icon. Below that are sections for 'Presenters (0)' and 'Participants (0)', both currently empty.

TOP LEFT | Attendees Pod Hosts, Presenters, Participants



The Chat pod is titled 'Chat (Everyone)' with a menu icon. It shows a message: 'The chat history has been cleared'. Below are three messages from 'Lea Deforest': 'Welcome to the TDL webinar!', 'We're so happy to have you.', and 'Visit our web site to learn more www.tdl.org'. At the bottom is a text input field with a green border and a speech bubble icon.

MIDDLE LEFT | Chat Pod Questions, comments, links in "everyone"; direct message individuals



The Downloads pod has a header with a menu icon. Below is a table with two columns: 'Name' and 'Size'. The table contains one row: 'TDL_Slide_2019.pptx' with a size of '249 KB'. At the bottom are two buttons: 'Upload File...' and 'Download File(s)'.

Name	Size
TDL_Slide_2019.pptx	249 KB

BOTTOM LEFT | Download Pod Files shared by TDL and/or our presenters

Overview

Records retention & policy mandates as well as program expansion - John Bondurant, TAMU

Mission alignment & codification as well as staffing & resource challenges - Anna Lamphear, UT Austin

Rights & Permissions - Leanna Barcelona, Baylor

Scoping collections - Mark Phillips, UNT

Q&A as time allows - please use chat to ask questions



Web Archiving @ Texas A&M University

University Libraries Vision:

- Web Archiving as Records Retention
 - Required by Texas state law
 - Web archiving originally administered by University Archivist in 2013
 - Administered by Digital Archivist since 2015
- More records are born-digital and web based
 - University President and Regents Offices
 - Annual Budgets
 - Annual Reports
 - Clery Act
 - Registrar/Catalog

Library & Collections

- Selected web crawls for select Cushing Library Collections since 2013
- 2015 end of term crawl for library.tamu.edu domain



Expanding TAMU Libraries' Web Archiving



TAMU Libraries website link to [TAMU Archive-It page](#)

- Library Web Governance Team
 - Proposed 2016, approved 2018
 - Requested regular crawls of library.tamu.edu domains
- [Texas A&M University Libraries Web Archiving Methods and Collection Guidelines](#)
 - Based on UTSA and UW Madison web archiving documents
 - Broader University Coverage
 - [Texas A&M University Research Centers and Institutes](#)
 - Expanded Cushing Library Collections
 - [Science Fiction and Fantasy Research Collection](#)
- Web archive URL submission form
 - Still being reviewed by University Libraries' Executive Team

Challenges to Codifying a Web Archiving Program in the Library



<https://nynl.net/archive.net/media/congressional-library-366c5d>

Defining the mission
Maximizing expertise
Managing finite resources



Defining the mission

<http://netpreserve.org/web-archiving/collection-development-policies/>



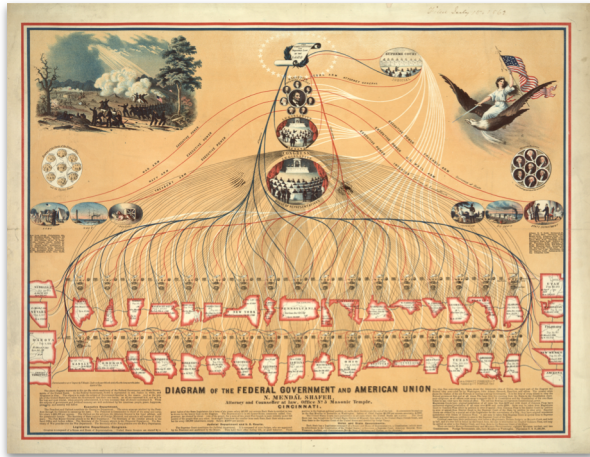
https://upload.wikimedia.org/wikipedia/commons/thumb/5/59/United_States_Constitution.jpg/199px-United_States_Constitution.jpg

tdl.org



Maximizing expertise

<https://catalog.lib.utexas.edu/record=b6982419~S29>



https://upload.wikimedia.org/wikipedia/commons/4/45/Diagram_of_the_Federal_Government_and_American_Union.jpg

tdl.org



Managing finite resources

<https://library.unt.edu/digital-projects-unit/web-archiving/software-processes/#crawl-scope>



<https://upload.wikimedia.org/wikipedia/commons/0/00/Exictbonusarmy.jpg>

tdl.org



Permissions in Web-Archiving: Things to Consider

Presentation Outline

- Public Domain and the Internet
- Robots.txt files
- Internet Archive/Archive-It
- Library of Congress and British Library statements
- Mandated Web-Archiving/Legal Depositing



Public domain and the Internet



“‘Public domain’ refers to material for which copyright has expired or where an author has specifically indicated that the material is in the public domain.

Material that is available to the public via the Internet or other means is not in public domain simply by reason of its being publicly available.”

Public domain and the Internet (cont.)



“It is reasonable to conclude that if a person has made material available on the web, there is an implied licence to make a copy for personal use if there is no statement to the contrary.”

Taken from Southern Cross University: <https://www.scu.edu.au/copyright/key-copyright-concepts/public-domain-and-the-internet/>

robots.txt

A robots.txt file tells crawlers which pages or files the crawler can or can't request from your site.

But, these files can be overridden by CSS and stylesheets.



Archive-It: Web Archiving Lifecycle Model



Risk Management Assessment: “In developing a web archiving program, many institutions consider the level of risk related to copyright they are willing to accept and how they will manage this risk. Whether and how institutions decide to seek permission from site owners before archiving is one of the clearest examples of risk management policy making in action.”

Archive-It: Web Archiving Lifecycle Model (cont.)

“The Archive-It service does not take a stand on copyright, and follows the Oakland Archive Policy, established in 2002, striving to work collaboratively with content providers. The service will honor requests to remove content from public access.”



The Oakland Archive Policy



“Recommendations for Managing Removal Requests And Preserving Archival Integrity”

Removal requests typically fall under five categories laid out in the policy.

The policy then provides responses to each of the five categories/scenarios.

The Cobweb: Can the Internet be archived?

New Yorker Article -- “The Library of Congress has something like an opt-in policy; the Internet Archive has an opt-out policy. The Wayback Machine collects every Web page it can find, unless that page is blocked.”



[Library of Congress](#)



For Site Owners

The Library notifies each site owner that we would like to include their content in the archive (with the exception of government websites) prior to archiving. In some cases, the email asks permission to archive or to provide off-site access to researchers.

[For Site Owners »](#)

[tdl.org](#)

The British Library

Web crawling politeness and protocols

The web crawling software is also programmed with politeness rules and parameters designed to ensure that there is no harmful impact upon the performance of the target website. For example, they include a limit on how many levels are crawled or how much content is requested from an individual website. Also, when multiple requests for different pages and files are issued to the same website, the software is programmed to leave an interval between each request, to safeguard against using up too much bandwidth and overloading the website.

The web crawling software uses standard automated protocols to identify itself and to inform the publisher's webmaster (via information called a "user-agent string" submitted to the web server's log of server requests) on each occasion that a page is crawled. The website owner can choose whether or not to use this information, but is not required to take any action such as changing the website's "robots.txt" permission file.

Where the web crawling software encounters a login facility, it cannot access any material behind the login facility without the appropriate password or access credentials.



Mandated Web-Archiving And Legal Deposit



International Internet Preservation Consortium has a [list](#) of countries with legal deposit laws that extend to websites

Examples:

Mandated: Canada, Spain, Japan, U.K.

Mandate does not include websites: U.S., Poland, Singapore

Links from Presentation

Southern Cross University Information on Copyright:

<https://www.scu.edu.au/copyright/key-copyright-concepts/public-domain-and-the-internet/>

Archive-It Life Cycle Model: <https://archive-it.org/blog/learn-more/publications/web-archiving-life-cycle-model/>

The Oakland Archive Policy: <http://groups.ischool.berkeley.edu/archive/aps/removal-policy>

New Yorker article: <https://www.newyorker.com/magazine/2015/01/26/cobweb>

Library of Congress: <https://www.loc.gov/programs/web-archiving/about-this-program/>

British Library: <https://www.bl.uk/legal-deposit/web-archiving>

International Internet Preservation Consortium on Legal Deposits:

<http://netpreserve.org/web-archiving/legal-deposit/>



Web Archive Collection Scope

tdl.org

Collection Scope Document



A good way to get started talking about building a web archive

What are you going to collect?

Does it relate to other collections?

Who is this collection for?

What is included and what is not included?

Who are your users?

How does this fit into the mission of your organization?

Collection Scope

Especially good for topical and subject-based collections.

Who else is collecting in this area?

How will you handle permissions?

Will you follow robots.txt?

What kind of access will you provide?

What is your plan for preservation?



Building Collection Scope Documents

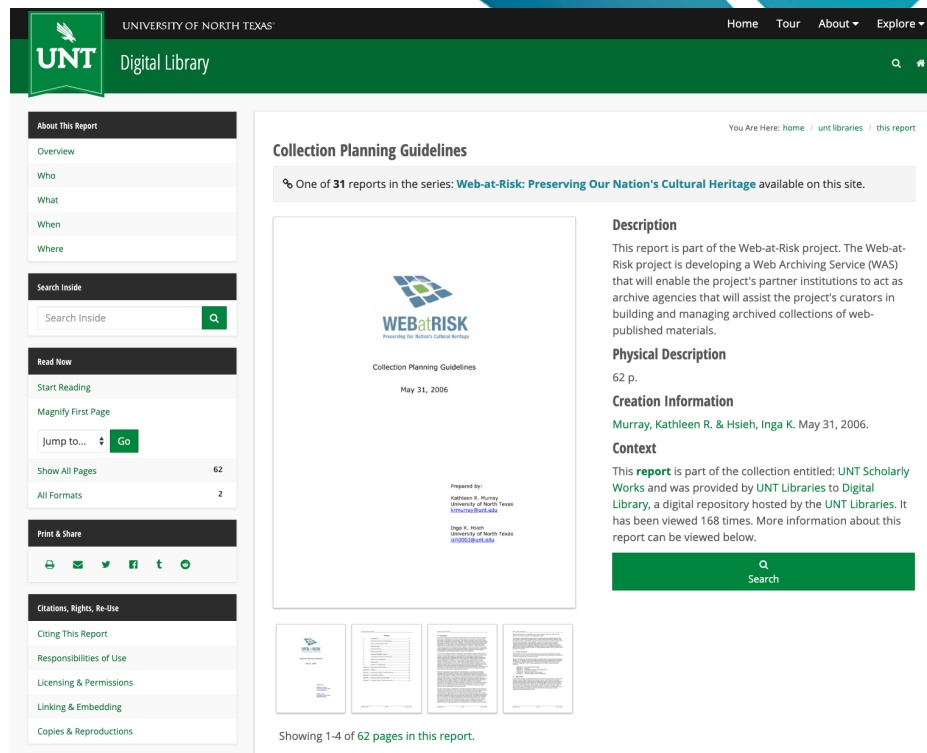
You don't have to start from scratch.

Collection Planning Guidelines

Kathleen R. Murray & Inga K. Hsieh (2006)

Web @ Risk project (NDIIPP)

https://digital.library.unt.edu/ark:/67531/meta_dc33006/



The screenshot displays the UNT Digital Library interface. The header includes the UNT logo and navigation links. The left sidebar contains sections for 'About This Report' (Overview, Who, What, When, Where), 'Search Inside' (with a search bar), 'Read Now' (Start Reading, Magnify First Page, Jump to... with a Go button, Show All Pages: 62, All Formats: 2), 'Print & Share' (with social media icons), and 'Citations, Rights, Re-Use' (Citing This Report, Responsibilities of Use, Licensing & Permissions, Linking & Embedding, Copies & Reproductions). The main content area is titled 'Collection Planning Guidelines' and includes a breadcrumb trail: 'You Are Here: home / unt libraries / this report'. It states that one of 31 reports in the series 'Web-at-Risk: Preserving Our Nation's Cultural Heritage' is available. The report cover is shown, featuring the 'WEBatRisk' logo and the title 'Collection Planning Guidelines' dated May 31, 2006. The cover also lists the authors: Kathleen R. Murray (University of North Texas) and Inga K. Hsieh (University of North Texas). To the right of the cover, the 'Description' section explains that the report is part of the Web-at-Risk project, which is developing a Web Archiving Service (WAS) to assist curators in building and managing archived collections. The 'Physical Description' is 62 pages. The 'Creation Information' lists the authors and the date (May 31, 2006). The 'Context' section notes that the report is part of the 'UNT Scholarly Works' collection, provided by UNT Libraries to Digital Library, and has been viewed 168 times. A search bar is located at the bottom right of the main content area.

UNIVERSITY OF NORTH TEXAS
UNT Digital Library

Home Tour About Explore

You Are Here: home / unt libraries / this report

Collection Planning Guidelines

One of 31 reports in the series: [Web-at-Risk: Preserving Our Nation's Cultural Heritage](#) available on this site.

Description
This report is part of the Web-at-Risk project. The Web-at-Risk project is developing a Web Archiving Service (WAS) that will enable the project's partner institutions to act as archive agencies that will assist the project's curators in building and managing archived collections of web-published materials.

Physical Description
62 p.

Creation Information
[Murray, Kathleen R. & Hsieh, Inga K.](#) May 31, 2006.

Context
This **report** is part of the collection entitled: [UNT Scholarly Works](#) and was provided by UNT Libraries to Digital Library, a digital repository hosted by the UNT Libraries. It has been viewed 168 times. More information about this report can be viewed below.

Showing 1-4 of 62 pages in this report.



Collection Planning Guidelines

May 31, 2006

Prepared by:

Kathleen R. Murray
University of North Texas
krmurray@unt.edu

Inga K. Hsieh
University of North Texas
ikh0003@unt.edu



The good stuff starts on page 18.

“3 Creating a Web Collection Plan”

8 Sections with 3-5 sub-sections.

tdl.org

Web Collection Plan - Outline



Section 1. Mission & Scope

Section 2. Selection Activities

Section 3. Web Site Acquisition

Section 4. Descriptive Metadata Requirements

Section 5. Preservation & Access Requirements

Section 6. Maintenance & Weeding

Section 7. Preservation

Section 8. Appendices

Subsections

Each section is well documented with what kinds of information you might want to include in that section.

Some of these don't apply in the same way that they did in 2006.

But still is a good way to get started.

4 Mission & Scope

Web collection plans begin with articulating the mission that guides collection development, describing the user groups, or Designated Community, served by the collection, and stating the information need(s) the collection will address. Web collections will generally consist of web sites united by a common subject, theme, or event. For example, discipline-related web sites included in curriculum subject guides support an academic library's mission to provide materials in support of faculty and student scholarship and learning.

4.1 Contents

Section 1. Mission & Scope
A. Mission Statement
B. User Group(s)
C. Collection Subject, Theme, or Event
D. Curator(s)

4.2 What to Address

4.2.1 Mission Statement

Articulate the mission under the umbrella of which the collection is being developed. For many collections this will be the mission statement of the library. For others, web collection development may be more appropriately positioned under mission of the organization or institution.

4.2.2 User Group(s)

Define the user groups for the web collection. In many cases there will be more than one user group that will use a collection, for example faculty, students, and the general public. For web collections, a complete understanding of user groups is important so that the unique characteristics and needs of each one can influence the range of collection development activities, which include identifying what to collect and the metadata required for information discovery. Be as detailed as appropriate regarding each user group's demographic characteristics and their use of web-published materials.

Consider assessing the user information needs that could be addressed by web-published materials. Understanding how users currently use web-published materials to carry out their organizational or professional responsibilities might be helpful. Various methods can be used for this, including surveys, focus groups, and interviews. This should help identify gaps in existing collections and prioritize materials targeted for web collection development.

4.2.3 Collection Subject, Theme, or Event

State the subject area or theme that unites the web sites in the web collection. In some cases, web sites in a collection may be related to a common event, such as the Olympic Games or a national election. Describe how the collection supports the mission of the library, organization, or institution.

4.2.4 Curator(s)

Identify the curator(s) of the collection. Include a description of each curator's responsibilities within their organization or institution and their contact information.

Web Collection Plan Overview: Considerations for Project Curators

Kathleen R. Murray & Inga K. Hsieh
(2006)

Companion for the previous
document.

Provides additional guidance for
curators wanting to develop
collection plans.

<https://digital.library.unt.edu/ark:/67531/metadc33004/>

tdl.org

The screenshot displays the UNT Digital Library interface. The top navigation bar includes the UNT logo, the text 'UNIVERSITY OF NORTH TEXAS Digital Library', and links for Home, Tour, About, and Explore. A search icon is also present. Below the navigation bar, a sidebar on the left contains sections for 'About This Report' (Overview, Who, What, When, Where), 'Search Inside' (with a search input field), 'Read Now' (Start Reading, Magnify First Page, Jump to... with a Go button, Show All Pages: 93, All Formats: 2), 'Print & Share' (with various social media and print icons), and 'Citations, Rights, Re-use' (Citing This Report, Responsibilities of Use, Licensing & Permissions, Linking & Embedding). The main content area features the title 'Web Collection Plan Overview: Considerations for Project Curators' and a sub-header indicating it is one of 31 reports in the series 'Web-at-Risk: Preserving Our Nation's Cultural Heritage'. The report cover image shows the 'WEBatRISK' logo and the title. To the right of the cover, a 'Description' section explains the report's purpose as a companion document to the Web Collection Plan Template, dated August 24, 2006. It also includes a 'Physical Description' (93 p.) and 'Creation Information' (Murray, Kathleen R. & Hsieh, Inga K. August 24, 2006). A 'Context' section notes that the report is part of the 'UNT Scholarly Works' collection and has been viewed 316 times. At the bottom right, there is a green search bar with a magnifying glass icon and the word 'Search'.



WATX19 @ Baylor

November 6th

Last chance to register!

<https://www.tdl.org/2019/04/watx19/>



Thanks and Q&A