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

TOP LEFT | Attendees Pod  Hosts, Presenters, Participants

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

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

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
Maximizing expertise

https://catalog.lib.utexas.edu/record=b6982419~S29
Managing finite resources

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

tdl.org
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:


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
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?

tdl.org
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

Web @ Risk project (NDIIPP)
https://digital.library.unt.edu/ark:/67531/meta
dc33006/
The good stuff starts on page 18.
“3 Creating a Web Collection Plan”
8 Sections with 3-5 sub-sections.
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.
Web Collection Plan Overview: Considerations for Project Curators

Companion for the previous document.

Provides additional guidance for curators wanting to develop collection plans.

https://digital.library.unt.edu/ark:/67531/metadc33004/
WATX19 @ Baylor
November 6th
Last chance to register!

https://www.tdl.org/2019/04/watx19/
Thanks and Q&A