UTSA Libraries
Special Collections

Web Archiving Methods and Collection Guidelines

Purpose:

The University of Texas at San Antonio Special Collections began partnering with the Internet Archive's Archive-It web archiving service in September of 2009 to preserve web content that is of enduring value to both South Texas and UTSA. Archive-It allows our staff to capture relevant web content and ensure its long-term access through the Internet Archive's global web archive, the Wayback Machine. The Archive-It service selectively crawls websites, taking a snapshot of the page, and storing a copy in the Internet Archive. The archived web pages are then made publicly accessible on UTSA Special Collection's Archive-It partner page.

Method:

UTSA Special Collections staff:

- Identify a topic/subject/theme for a collection.
- Select relevant, specific websites to crawl
- Administer mechanics of crawls and add descriptions (metadata).
- Determine frequency of content changes/updates and set frequency of crawl accordingly (weekly, monthly, annually).

Types of Web Content Collected:

- Official University of Texas at San Antonio websites
- University-affiliated websites (including organization websites, Facebook, Twitter, Flickr, etc.)
- Websites relevant to Texas and the Southwest (Border Studies, Gender Studies, South Texas and San Antonio History)
- Web pages (HTML, CSS, and embedded content such as images, video, and PDFs)

Types of Web Content Generally Not Collected:

- Websites created by individual students
- Password protected sites
- Databases
- Calendars
- Public websites that have robots.txt exclusion requests

Types of Web Content that may not be adequately captured:

- JavaScript-based content
- YouTube videos
- Streaming audio/video

Copyright:

UTSA does not claim copyright to any of the materials within the archive. It is the sole responsibility of the user to determine the copyright status of archival collections before publishing materials.

Frequently Asked Questions (FAQs)

Q: How does UTSA Special Collections decide which websites to archive?

A: Special Collections staff use a variety of measures to determine whether or not a website is appropriate for archiving, and the frequency with which particular websites are crawled.

Is the website:

- Relevant to our thematic collections?
- Of appropriate size and scope?
- Already part of a larger seed?
- Already crawled by the Wayback Machine? If so, do we agree with the frequency of the Wayback Machine crawls?
- Updated frequently?
- Protected by robots.txt exclusions? If so, do we have the right to crawl the content, or should we contact the site owner for permission to crawl?
- Password protected?
- Dynamic, i.e. does functionality depend upon user input (databases, JavaScript) that the crawler cannot capture?
- Structured such that calendars can be avoided, as these cannot be easily crawled (is there a "\calendar" page we can exclude from the crawl)?

Q: Does Archive-It capture the date and time when a web page is updated?

A: No. The Archive-It web crawler can only take a snapshot of a website at a given time. Comparison of a website with previous crawls of that website may show that content has changed between crawls. Currently, there is no way to exactly determine when and how a particular website modifies its content.

Q: How do I know I am looking at an archived web page, and not the live web?

A: Some archived websites contain internal links that are not captured by the Archive-It web crawler and will divert to the live web. Websites archived by the UTSA Special Collections will always have a yellow banner at the top of the page with following:

'You are viewing an archived web page, collected at the request of University of Texas, San Antonio using Archive-It. This page was captured on time/date, and is part of the University of Texas at San Antonio [collection name] collection. The information on this web page may be out of date. See All versions of this archived page.'

Q: How do I make sure that my website is included in the archive?

A: If you are a member of the UTSA or South Texas community and your web content is not currently being archived by the Internet Archive or the Archive-It Program, you can contact Special Collections staff to inquire about having your website included in UTSA web archive.

Q: How do I remove my web page or photograph from the archive?

A: Web content in our collections is part of the Internet Archive. We are not able to remove content from past web crawls. If you would like for your website to not be archived by the UTSA Special Collections' web crawler in the future, please contact Special Collections staff.

Q: Why am I directed to the Internet Archive's 'Not in Archive' page for internal links within some websites?

A: Some web pages may only have a portion of their website archived by the Archive-It web crawler. Web content may be excluded from the archive either because of robots.txt exclusion requests, if portions of the website are hosted on a different web domain, or if the crawler was not able to copy the entire site.