



A Practical Approach to Web Archiving: Applying archival theory to websites and social media pages

Andrea Weddle, CA
Jeremy Floyd, CA
Adam Northam

Texas A&M University-Commerce

May 23, 2013
Society of Southwest Archivists Conference

How do you Archive the Web? An Overview of Web Archiving

Adam Northam
Digital Collections Librarian
Texas A&M University-Commerce
Adam.Northam@tamuc.edu
@atomic_red77



Credit: U.S. Geological Survey
Department of the Interior/USGS
U.S. Geological Survey/photo by J.K. Nakata

“The Internet
is the world's
largest library.
It's just that all
the books are
on the floor.”

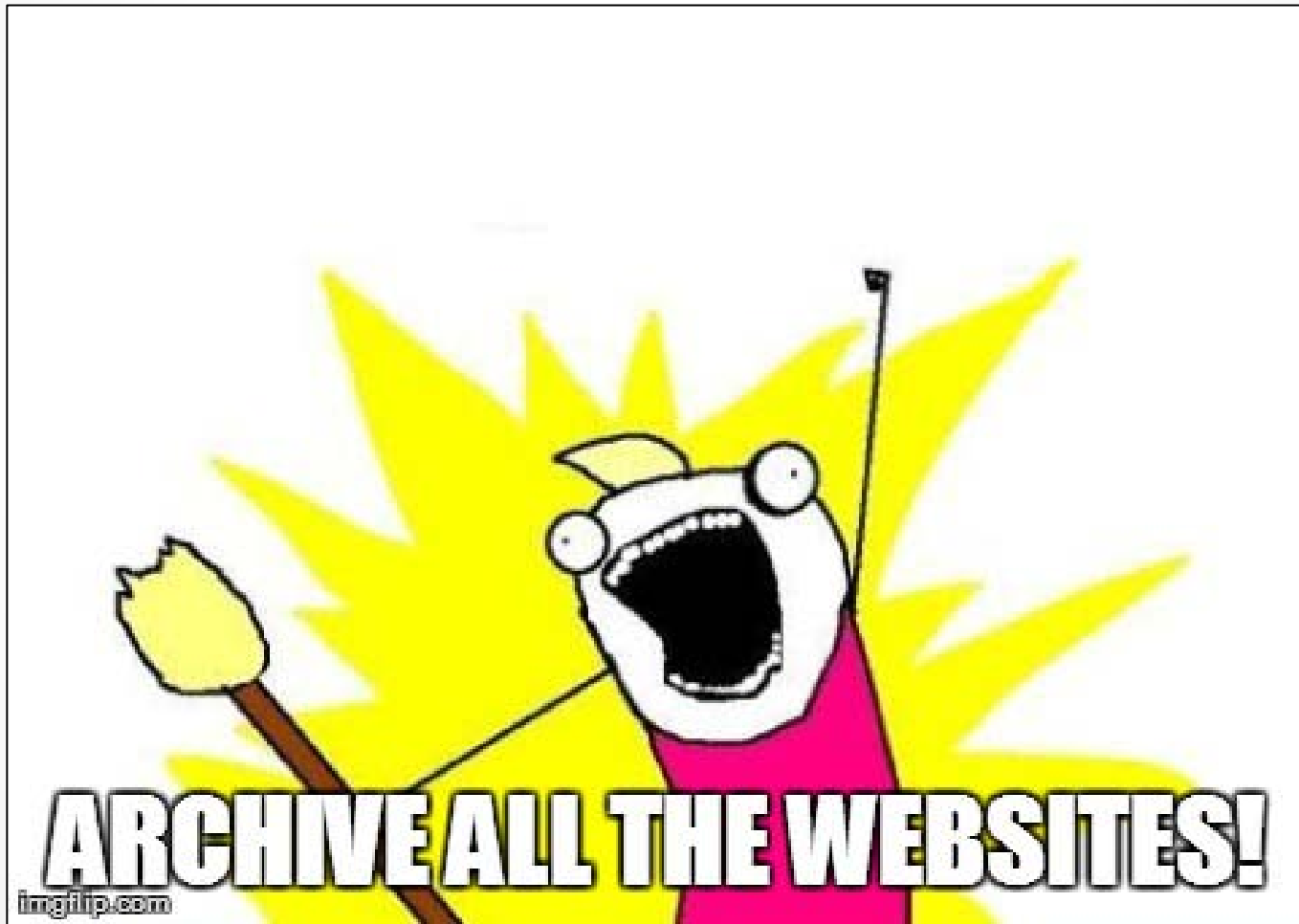
--John Allen Paulos

You Can't Put the Web in a Filing Cabinet



Image from 123rf.com

...but You Can Archive the Web



What's Next?



Two Roads to Choose

2013



DIY: Things to Consider

- Open source (free) software
- Likely more customizable
- IT Infrastructure
- Level of available expertise
- Plan for storage and backup of archived data

Open Source Web Archiving Solutions



HTTrack

- Free software
- Installed on a local computer; not a server
- Saves captured websites to a local hard drive
- <http://www.httrack.com/>

Open Source Web Archiving Solutions



GNU Wget

- Free software package
- Can follow links in HTML, XHTML, CSS
- Runs on most UNIX-like operating systems as well as Microsoft Windows
- Supports WARC output
- <http://www.gnu.org/software/wget/>

Hosted: Things To Consider

- Higher upfront cost
- Not as customizable
- Hosted solutions provide much more infrastructure
- Usually requires less tech expertise and provides more formal training
- Hosted solutions provide all storage, backup, maintenance of data

Hosted Web Archiving Solutions



Web Harvester

- OCLC's entry into web archiving
- Uses proprietary technology
- Integrates with other OCLC tools
- Requires Subscription to OCLC's cataloging service, ContentDM license, and ContentDM hosting services
- <http://www.oclc.org/web-harvester/overview.en.html>

Hosted Web Archiving Solutions

- California Digital Library
- Free to join for all UC departments and organizations (charged only for storage)
- Fee based subscription service for all other institutions

- Utilizes:
 - Heritrix web crawler for capture
 - Wayback for display
 - Nutchwax search engine



- 56 public archives
- 21 partners
- 4407 web sites
- 616,585,489 documents
- 32.3 TB of data

Hosted Web Archiving Solutions



- Subscription Service
- Heritrix web crawler
- Nutchwax search engine
- Wayback Machine browser
 - All developed and maintained by the Internet Archive



- 248 partner organizations
- 5,635,350,2081 URLs in 2,091 collections
- Partners in 45 states and 15 countries including, university libraries, state archives, historical societies, federal institutions, NGOs, public libraries, and museums

Saving Digital Ephemera: Strategies for Preserving University Websites, Blogs, and Social Media

Jeremy Floyd
Archivist

Texas A&M University – Commerce

jeremyjfloyd@gmail.com

[@jjamesfloyd](#)

Texas A&M University – Commerce partnered with Archive-It

Explore Collecting Organizations

[Show All Organizations](#)



Texas A&M University - Commerce

Texas A&M University - Commerce is a public four year teaching and research institution located in Northeast Texas. Founded in 1889 by William Mayo as East Texas...



Maryland State Archives

As the historical agency for Maryland, the State Archives is the depository for records of permanent value. Today, the Internet is the face of government. Through this...

David M. Rubenstein Rare Book & Manuscript Library

Duke University

The David M. Rubenstein Rare Book and Manuscript Library at Duke University collects and preserves materials spanning a broad range of subjects areas in support of research and...



Archive-It
A service of the [Internet Archive](#)

2011-2012 Archive-It [r20130208-0046]
Archiving the internet for future generations
Collect it, manage it, search it... Archive-It

[Home](#)

[Learn More](#)

[About Archive-It](#)
[News/Press](#)

[Contact Us](#)



Selecting Seed URLs

University Websites

<http://www.tamuc.edu/>
<http://web.tamuc.edu/>
<http://catalog.tamuc.edu/>
<http://pride.tamuc.edu/>
<http://www.tamu-commercedining.com/>
<http://tamuc.orgsync.com/>
<http://www.lionathletics.com/>

Facebook

<http://www.facebook.com/tamuccommerce/>
<http://www.facebook.com/TAMUCLibraries/>
<http://www.facebook.com/pages/AM-Commerce-Lion-Athletics/242136009137926?ref=ts/>
<http://www.facebook.com/TAMUCspirit/>
<http://www.facebook.com/tamucalumni/>

Twitter

http://twitter.com/TAMU_Commerce/
http://twitter.com/Lion_Athletics/
<http://twitter.com/ketrradio/>
<http://twitter.com/TheEastTexan/>
<http://twitter.com/LionsAfterDark/>
http://twitter.com/TAMUC_News/
<http://twitter.com/LionSafety/>
<http://twitter.com/TAMUCalumni/>
http://twitter.com/TAMUC_Mesquite/

Youtube

<http://www.youtube.com/user/LionsMedia/>

University News and Media

<http://www.ketr.org/>
<http://TheEastTexanOnline.com>

Managing Scope and Frequency of Crawls

Host Constraints Crawl Limits Expand Scope

Use this area to put constraints on individual hosts (similar to domains). For each host, you can completely block content from being archived at all, can block just certain urls from that host from being archived, or put a maximum limit on the number of documents that can be archived from that host. View additional information on [using Host Constraints](#).

[Add Hosts](#)

View only hosts containing [Filter](#) [Clear](#) 3 Record(s)

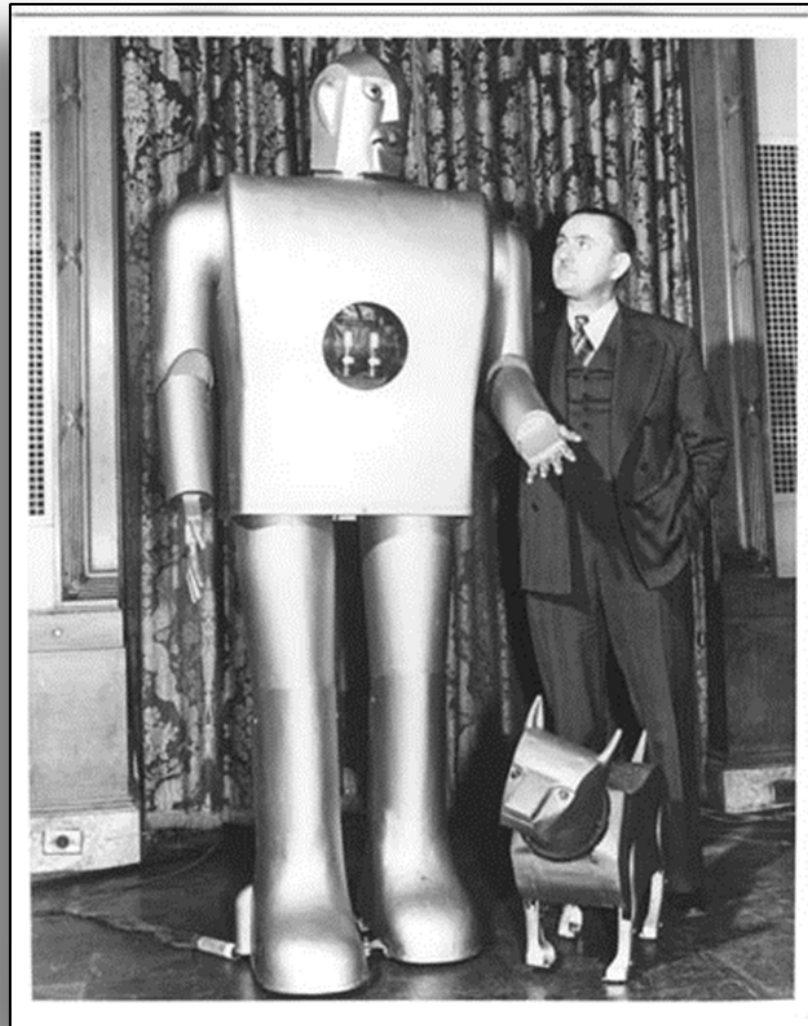
	Host	Block All	Block URLs if	Doc Limit	Ignore Robots.txt
x	fbcfn.net	<input type="checkbox"/>		click me	<input checked="" type="checkbox"/>
x	twitter.com	<input type="checkbox"/>	URL Contains ?lang=	click me	<input type="checkbox"/>
x	www.facebook.com	<input type="checkbox"/>	Reg. Expr. ^https?://...\.facebook\.com.* Reg. Expr. ^https?://www\.facebook\.com/(profile\.php people).*	8,000	<input checked="" type="checkbox"/>

Host Constraints Crawl Limits Expand Scope

Enter document or data limits below. Entries will be automatically saved. Data limits must be listed in megabytes only. Please note there are 1024 megabytes in 1 gigabyte.

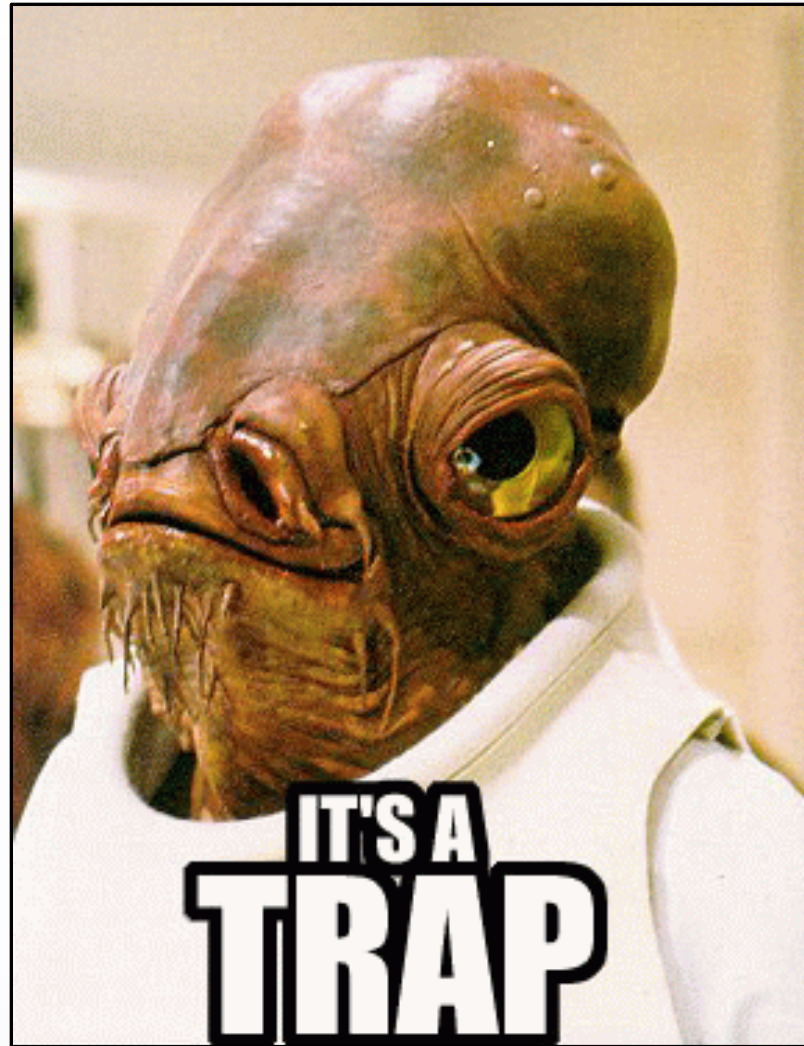
	Document Limit	Data Limit (MB)	Archive PDFs Only	Time Limit
One-Time	<input type="text"/>	<input type="text"/>	<input type="checkbox"/>	<input type="radio"/> 1 day <input checked="" type="radio"/> 3 days <input type="radio"/> 5 days <input type="radio"/> 7 days
Twice Daily	<input type="text"/>	<input type="text"/>	<input type="checkbox"/>	<input checked="" type="radio"/> 12 hours
Daily	<input type="text"/>	<input type="text"/>	<input type="checkbox"/>	<input checked="" type="radio"/> 1 day
Weekly	<input type="text"/>	<input type="text"/>	<input type="checkbox"/>	<input checked="" type="radio"/> 3 days
Monthly	<input type="text"/>	<input type="text"/>	<input type="checkbox"/>	<input checked="" type="radio"/> 3 days <input type="radio"/> 5 days
Bi-monthly	<input type="text"/>	<input type="text"/>	<input type="checkbox"/>	<input checked="" type="radio"/> 3 days <input type="radio"/> 5 days <input type="radio"/> 7 days
Quarterly	<input type="text"/>	<input type="text"/>	<input type="checkbox"/>	<input checked="" type="radio"/> 3 days <input type="radio"/> 5 days <input type="radio"/> 7 days
Semiannual	<input type="text"/>	<input type="text"/>	<input type="checkbox"/>	<input checked="" type="radio"/> 5 days <input type="radio"/> 7 days
Annual	<input type="text"/>	<input type="text"/>	<input type="checkbox"/>	<input checked="" type="radio"/> 5 days <input type="radio"/> 7 days

robots.txt



“Robots- Electro and Sparko” 1940. still image. *Computer History Museum*
< <http://www.computerhistory.org/collections/accession/102693536> >

Crawler Traps



"It's A Trap" 2010. Know Your Meme
<<http://knowyourmeme.com/memes/its-a-trap>>

Adding Descriptive Metadata

Dublin Core

Title Add
Grab Title Remove
A&M-Commerce Lions (Lion_Athletics) on Twitter Remove

Creator Add
Texas A&M University - Commerce, Athletics Remove

Subject Add
social media Remove
twitter Remove
athletics Remove

Description Add
University athletics twitter feed Remove

Publisher Add

Contributor Add

Date Add

Type Add

Format Add

Identifier Add

Source Add

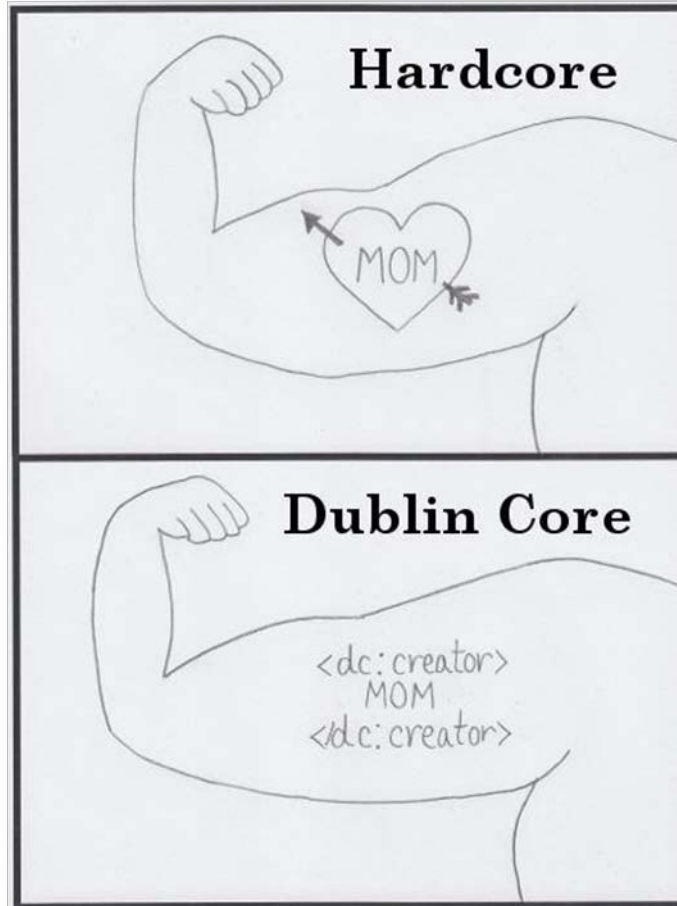
Relation Add

Coverage Add

Rights Add

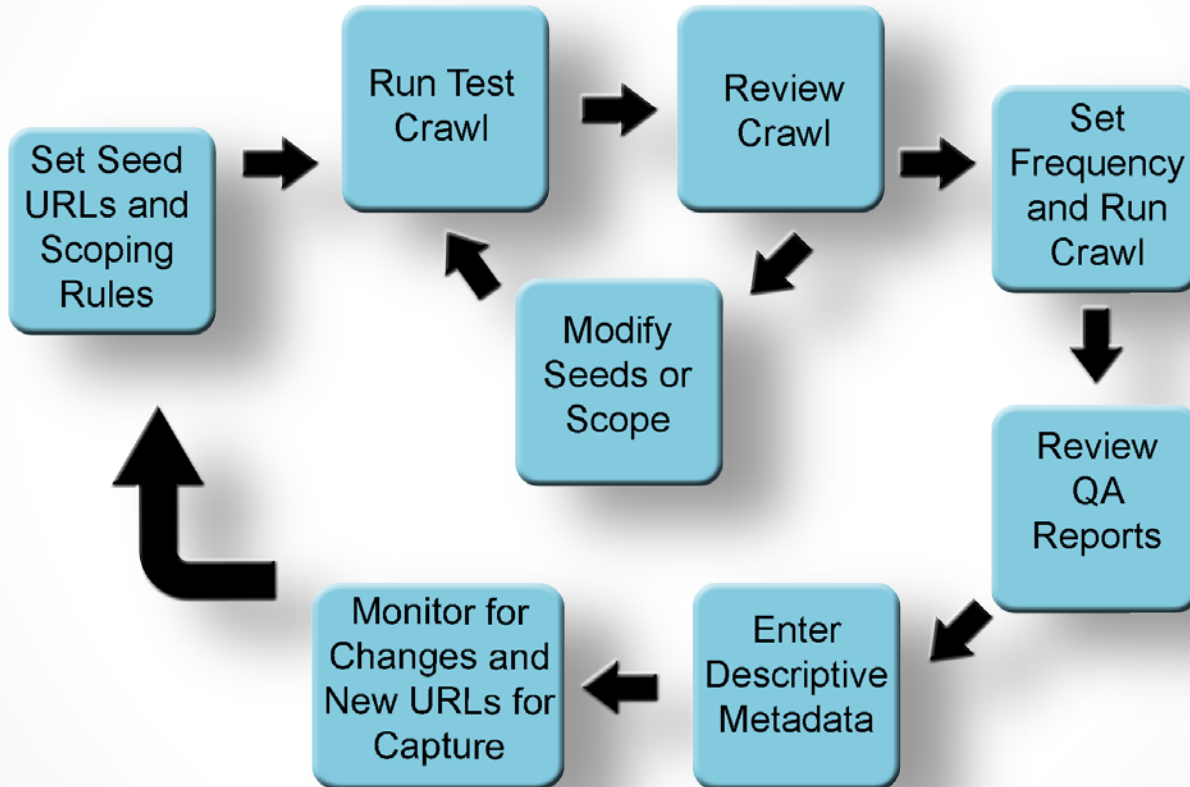
Collector Add

Language Add

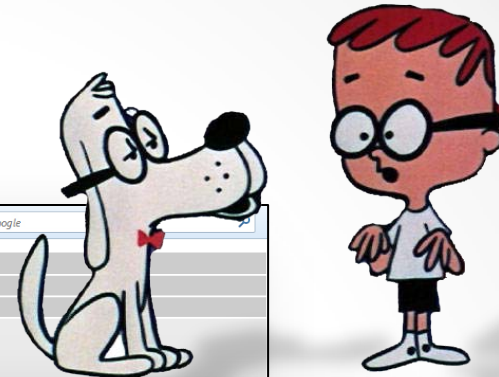


Rebecca Goldman. 2009. "Core Values." *Derangement and Description*.
<<http://derangementanddescription.wordpress.com/2009/07/13/core-values/>>

Establishing a Workflow



The Wayback Machine



wayback.archive-it.org/3139*/http://www.ketr.org/

Found 56 Captures between Sep 7, 2012 - Feb 8, 2013

2012	2013
Sep 7, 2012 * 43 pages	Jan 4, 2013 * 13 pages

wayback.archive-it.org/3139/20120327184713/http://web.tamuc-commerce.edu/aboutUs/administrativeOffices/president/default.aspx

You are viewing an archived web page, collected at the request of Texas A&M University - Commerce using Archive-It. This page was captured on 18:47:13 Mar 27, 2012, and is part of the University websites and social media collection. The information on this web page may be out of date. See All versions of this archived page.

TEXAS A&M UNIVERSITY COMMERCE


CURRENT STUDENTS FACULTY/STAFF VETERANS

ABOUT US ACADEMICS ADMISSIONS

Office of the President

Office of the President Home

- President's Biography
- President & Staff
- Director of Athletics
- Chief Information Officer
- Faculty Senate
- Budget Review and Development Council
- Strategic Enrollment Management Committee
- Addresses & Presentations
- Community Updates
- The Pride Alumni Magazine
- Sam Rayburn Series on Rural Life
- Strategic Plan



Home > About Us > Administration > Office of the President

Office of the President

[Guiding Principles](#)

[Texas Budget Request](#)

[Budget 2012 Suggestions](#)

Thank you for your interest in the 120-year legacy of providing high quality education and bringing together exceptional people.

ARCHIVE-IT

HOME EXPLORE LEARN MORE CONTACT US

A web archiving service to harvest and preserve digital collections a service of the Internet Archive

Explore >> Texas A&M University - Commerce

Texas A&M University - Commerce

Archive-It Partner Since: Mar, 2012
Organization Type: Colleges & Universities
Organization URL: <http://www.tamuc.edu/>

Description: Texas A&M University - Commerce is a public four year teaching and research institution located in Northeast Texas. Founded in 1889 by William Mayo as East Texas Normal College, the university has over 11,000 students and is the second largest in the A&M system. The James G. Gee Library at Texas A&M University - Commerce provides access to more than 1 million cataloged items to support the teaching and research needs of its students and faculty. The Special Collections Department is dedicated to preserving documents relating to the history of the university and the Northeast Texas region.

Narrow Your Results

Subject: Universities & Libraries (1)

Enter search terms here

Collections Sites Search Page Text

Page 1 of 1 (1 Total Results)

Sort By: Collection Name (A-Z) | Collection Name (Z-A)

University websites and social media

Archived since: Mar, 2012

Facebook

Capture

- Ignore robots.txt for host and stylesheets
- Set Document limit ~2000
- Expand scope SURT rule +http://(net,fbcdn,
- Language page duplicates
- Public Personal Pages

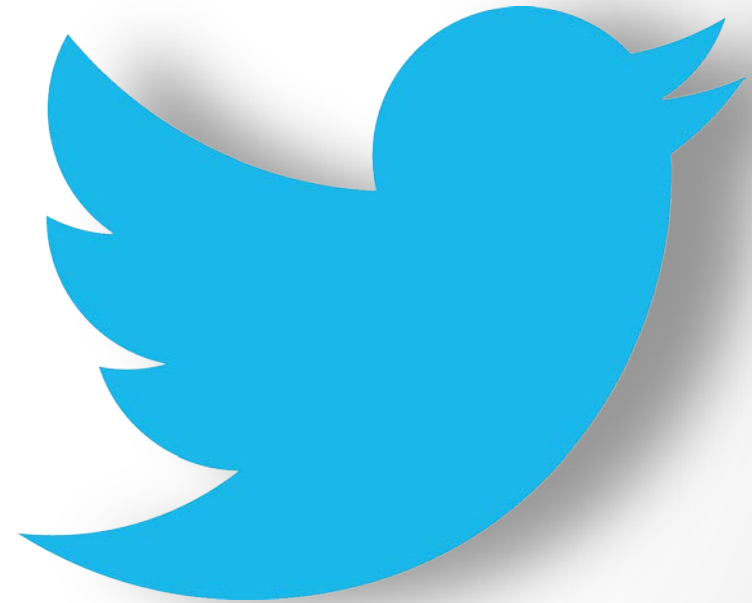


Display

- Infinite reloading loop, try disabling Javascript
- More comments, Older Posts
- Timeline View
- Uploaded Media

Twitter

- Must be a public twitter feed (default)
- Expand scope SURT rule:
+http://(com,twitter,)/i/profiles/show/
- Add rule "contains the following text"?lang= to avoid duplicates
- Expand Scope rule for urls that contain: http://t.co/ - captures links which utilize url shorteners



YouTube

Capture

- Channel as News/RSS Feed
- Crawling a Channel using a Document Limit
- Crawling Channel Videos Using an Uploaded Document RSS Feed



Display

- Video from Wayback page
- Video from specific crawl in Archive-it crawl report
- Video from the linked Video page

Capturing password protected content

- Added to Archive-it 4.8
(May 2013)
- Only in Alpha release

The screenshot shows the 'Settings' tab of the Archive-it interface. At the top, there are three tabs: 'Metadata', 'Settings', and 'Comments'. Below the tabs, there are navigation links: 'Back to Seed Management', 'Previous', and 'Next'. The main content area is divided into several sections, each with a label and a corresponding control:

- Frequency:** A dropdown menu set to 'One-Time'.
- Activation Status:** The status is 'Active' (in green), with a 'Deactivate' button.
- Status On Live Web:** The status is 'Unchecked', with a 'Verify Now' button.
- Public Site:** A checkbox labeled 'Show on public site' is checked.
- Seed Type:** A dropdown menu set to 'Default'.
- Group:** A dropdown menu set to 'None', with a 'Create New Group' button.
- Login Username:** A text input field.
- Login Password:** A text input field.

The 'Login Username' and 'Login Password' fields are highlighted with a red rectangular border.

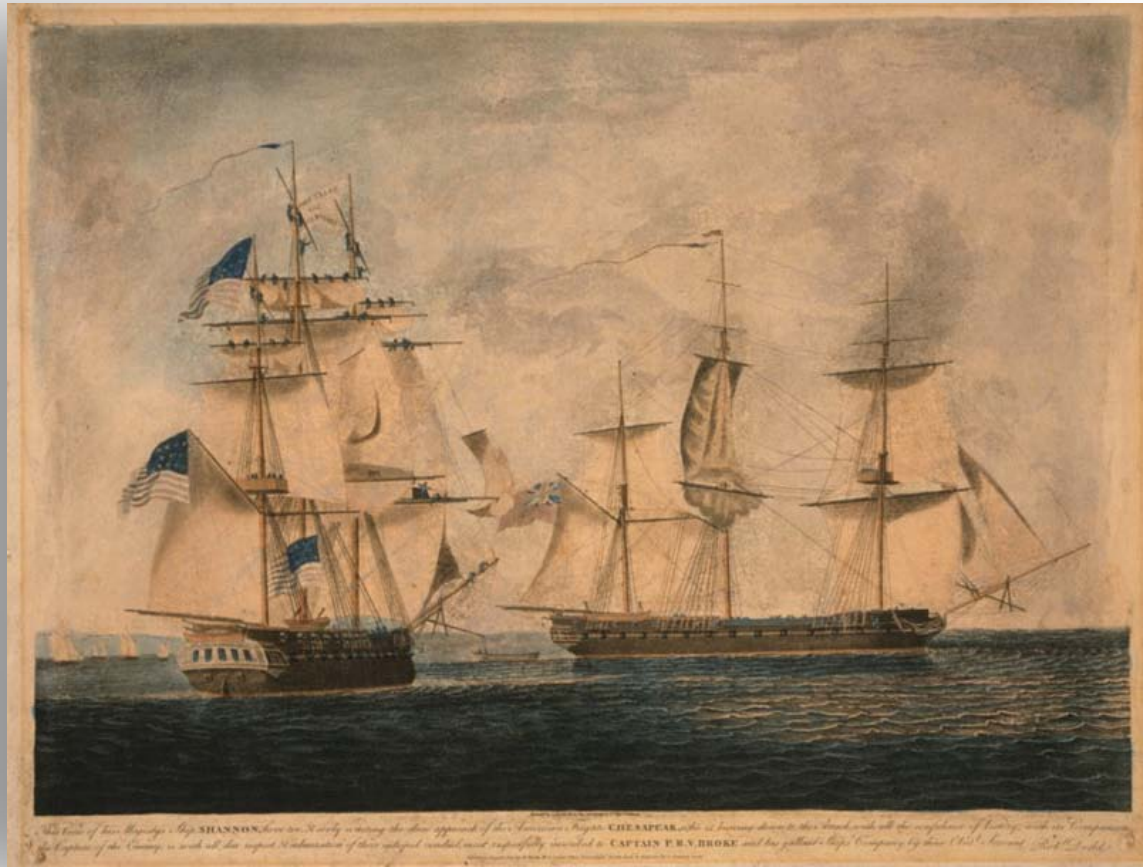
Zombie URI's

Captured Javascript reaching out to the live web = zombies in the web archive

The screenshot shows the Internet Archive interface for the URL <http://www.cnn.com/>. The page is dated September 3, 2008, and shows 15,098 captures. The main content area features a large article titled "Palin's path from city hall to governor's mansion" with a sub-headline "Republican National Convention 2008". The article text is dated "updated 4:24 p.m. EDT, Wed September 3, 2008". To the right, there is a "Latest News" section with a list of headlines, including "Poll measures race in three key states" and "Dems blast Lieberman, say he lied to delegates". Below this is a "Republican National Convention" section with a list of items like "Bernstein: Democrats better take note" and "Bush, McCain still an uneasy alliance". On the right side, there is a "Video" section with thumbnails for "Gustav recovery briefing", "Dogs die during lunch break", and "The RNC in St. Paul". At the bottom right, there is a "CNN TV" section with a thumbnail for "Palin in prime-time". The page layout, including the navigation bar and search box, is consistent with the live CNN website from that time, illustrating how the archived page still contains references to live content.

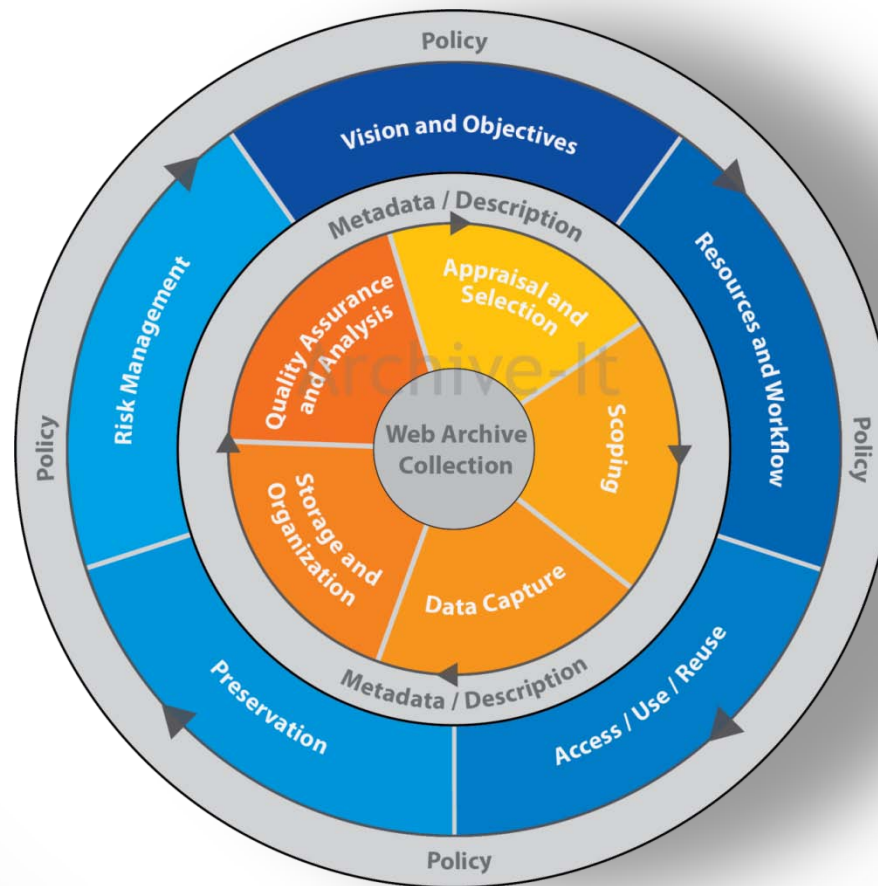
Example from Justin F. Brunelle, Web Science and Digital Libraries Research Group at Old Dominion University <<http://ws-dl.blogspot.com/2012/10/2012-10-10-zombies-in-archives.html>>

Web Archiving and the Ship of Theseus Paradox



Dodd, Robert. 1813. "This view of his majesty's ship Shannon" aquatint with engraving, hand-colored from Library of Congress Print and Photographs Online Catalog <<http://www.loc.gov/pictures/item/98519184/>>

Archive-It Web Archiving Life Cycle Model



Bragg, Molly, Kristine Hanna, et al. 2013 *THE WEB ARCHIVING LIFE CYCLE MODEL*. <http://archive-it.org/static/files/archiveit_life_cycle_model.pdf>

Curating the Web: Using web archiving to supplement traditional collections

Andrea Weddle, CA
Head of Special Collections and Archives
Texas A&M University-Commerce
Andrea.Weddle@tamuc.edu
@aeweddle

Topical web archives

- Web archiving gives us the opportunity to capture web-based content that fits within our collection development policy
- Gives us the capability to capture websites related to spontaneous events that effect our campus or region
- Proves that web archiving is not only for colleges, universities, and governmental offices

Archive-It Partners

- Colleges and universities
- Museums
- Local/state governments
- K-12 schools
- Organizations (governmental and non-governmental)
- Public libraries
- Politicians
- Individuals

@ A&M-Commerce

- Our main CDP includes a broad statement related to materials pertaining to Northeast Texas
- Added a web component in April 2012 to collect regional news websites
- Many longstanding newspapers in our region are moving to web-only
- New web-based news websites are increasing in popularity

A look at the data...

Weekly Northeast Texas News crawl

<u>Northeast Texas News</u>	
Weekly (ID #20130510141239362)	
<< Back to Reports	
Summary	Hosts
Statistics	
Started	May 10, 2013 9:12:39 AM
Completed	May 13, 2013 10:44:25 AM
Status	Finished (time limit)
Average Doc Rate	1.00 urls/sec
Average KB Rate	34.0 KB/s
Total Documents Archived*	264,851
New/Changed Documents	126,650
Unchanged Documents	138,201
Total Data Crawled	8.6 GB
New Data Archived*	4.9 GB
* This number applies to your Archive-It account budget.	

Quarterly University website crawl

<u>University websites and social media</u>	
Quarterly (ID #20130503142151743)	
<< Back to Reports	
Summary	Hosts
Statistics	
Started	May 3, 2013 9:21:51 AM
Completed	May 4, 2013 8:16:44 AM
Status	Finished
Average Doc Rate	1.00 urls/sec
Average KB Rate	108.0 KB/s
Total Documents Archived*	82,436
New/Changed Documents	54,514
Unchanged Documents	27,922
Total Data Crawled	8.6 GB
New Data Archived*	3.1 GB
* This number applies to your Archive-It account budget.	



Northeast Texas News

Collected by: [Texas A&M University - Commerce](#)

Archived since: May, 2013

Description: A collection of newspaper websites in the Northeast Texas region.

Subject: [Spontaneous Events](#), [Electronic newspapers](#)

Creator: [Texas A&M University - Commerce](#), [University Archives and Special Collections](#)

Type: [Interactive resource](#)

Collector: [Texas A&M University-Commerce](#)

Rights: Texas A&M University-Commerce does not claim copyright to any of the materials within this web collection.

Narrow Your Results

Creator

Sort By: **Count** | [\(A-Z\)](#)

[Mount Vernon Optic-Herald](#) (1)
[North Texas e-News](#) (1)
[The Herald Banner](#) (1)
[The Sulphur Springs News-Telegram](#) (1)
[eParis Extra!](#) (1)

Publisher

Sort By: **Count** | [\(A-Z\)](#)

[Mount Vernon Optic-Herald](#) (1)
[North Texas e-News](#) (1)
[The Herald Banner](#) (1)
[The Sulphur Springs News-Telegram](#) (1)
[eParis Extra!](#) (1)

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Page 1 of 1 (6 Total Results)

Sort By: [Title \(A-Z\)](#) | [Title \(Z-A\)](#) | [URL \(A-Z\)](#) | [URL \(Z-A\)](#)

Title: [The Herald Banner, Greenville, TX](#)

URL: <http://heraldbanner.com/>

Captured 2 times between May 3, 2013 and May 10, 2013

Creator: [The Herald Banner](#)

Publisher: [The Herald Banner](#)

Rights: © Community Newspaper Holdings, Inc

URL: <http://www.clarksvillenews.net/>

No Captures were found for this URL

Title: [eParis Extra!](#)

URL: <http://www.eParisextra.com/>

Captured 2 times between May 3, 2013 and May 10, 2013

Creator: [eParis Extra!](#)

Publisher: [eParis Extra!](#)

Rights: (c) Hometown Web Media, LLC

Looking to the future

- Develop a web archiving collection development policy
- Create a web archiving FAQ
- Develop a means for campus entities and local news agencies to either opt-in or opt-out of the web collection
- Increase the number of seeds

What others are doing

- University of Texas San Antonio
 - 19 total collections, three university collections
 - Topical web archive collections include bilingual education, Texas-related LGBTQ websites, renewable energy in Texas, San Antonio HIV/AIDS-related websites, San Antonio restaurants and organizations
- Primary Special Collections website includes an extensive FAQ and guidelines
- <http://www.archive-it.org/organizations/318>

UTSA LIBRARIES

[HOME](#)[ABOUT](#)[FIND INFORMATION](#)[SERVICES](#)[HELP](#)[Library Home](#) > [Find Information](#) > [Special Collections](#) > Collections

COLLECTIONS

[Search Special Collections](#)[University Archives](#)[Manuscripts](#)[Rare Books](#)[Photograph Collections](#)[Web Collections](#)[Digital Collections](#)

ABOUT SPECIAL COLLECTIONS

[Hours and Locations](#)[Contact Information](#)[History and Purpose](#)[Gifts and Donations to Special Collections](#)

PLANNING A VISIT

Web Collection Development Guidelines



Purpose:

The University of Texas at San Antonio Special Collections Department began partnering with the Internet Archive's **Archive-It Program** in September of 2009 to preserve web content that is of enduring value to both South Texas and the University of Texas at San Antonio. The Archive-It Program allows UTSA's Special Collection Department to capture relevant web content and ensure its long-term access through the Internet Archive's website. The Archive-It Program selectively crawls either web domains or individual web pages, taking a snapshot of the page, and storing a copy in the Internet Archive. The web page is then made publicly accessible on both **the University of Texas at San Antonio Special Collections' homepage** and at the **Archive-It partner page**.

Method:

The University of Texas at San Antonio Special Collections Department identifies either a web domain (e.g. <http://www.utsa.edu>) or a particular website (e.g. <http://www.utsa.edu>) to archive, and then periodically makes a copy of the web site to ensure long-term access to the web site's content. Depending upon the frequency in which the web content changes over time, the web crawler may be set to crawl a particular web site either twice daily, daily, weekly, monthly, quarterly, semi-annually, annually, or a single time. Since most web sites either self-archive a majority their content, or only change their content periodically, UTSA only crawls most web sites on either a semi-annual or annual basis. After the web site or web domain is crawled by Archive-IT, the web content is described and indexed by Special Collections staff and made available on the Archive-IT web site.

<http://lib.utsa.edu/collections/web-collection-development-guidelines/>

UTSA (cont.)

[Ask Special Collections](#)

[Request Access to a Collection](#)

[Photocopy and Reproduction Services](#)

[University Archives Forms and Handouts](#)

[Reference Services](#)

[Request Instruction Session](#)

Types of Web Content Collected:

- Official University of Texas at San Antonio web sites
- University-affiliated web sites (student organization web sites, Facebook, Twitter, Flickr pages created by organizations at the University of Texas at San Antonio)
- Web sites relevant to Texas and the Southwest (Border Studies, Gender Studies, South Texas and San Antonio History)
- Web pages (HTML), photographs (jpegs and tiffs), embedded video (mpegs), embedded audio (wav., mp3), and PDFs
- Publicly available web content that is not password protected

Types of Web Content Not Collected:

- Web sites created by individual students
- Password protected sites
- Databases
- Calendars
- JavaScripting
- Streaming audio/video
- YouTube videos
- Web sites that have robots.txt exclusion requests

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)

JOHN PEACE LIBRARY

Located on the Main Campus

One UTSA Circle
San Antonio, TX 78249-0671

Office of the Dean:
(210) 458-7506

DOWNTOWN LIBRARY

Located on the 2nd floor of the Buena Vista Building, Downtown Campus

501 W. César E. Chávez Boulevard
San Antonio, Texas 78207-4415

Phone: (210) 458-2440

APPLIED ENGINEERING & TECHNOLOGY LIBRARY

Located on the 2nd floor of the AET building on the Main Campus

One UTSA Circle
San Antonio, TX 78249-0671

SPECIAL COLLECTIONS

Reading Rooms located on the 4th floor of the John Peace Library and HemisFair Park Campus

One UTSA Circle
San Antonio, TX 78249-0671

<http://lib.utsa.edu/collections/web-collection-development-guidelines/>



Contemporary Women Artists on the Web

Collected by: [National Museum of Women in the Arts](#)

Archived since: Dec, 2011

Description: This is a collection of contemporary women artists on the web. There are three components to this collection: 1. Individual websites created by artists working in conceptual or new media art 2. Artist profiles of women artists represented by contemporary art galleries 3. Women artist organizations

Subject: [Arts & Humanities](#)

Narrow Your Results

Subject

Sort By: **Count** | [\(A-Z\)](#)

[Abrahams, Annie, 1954-](#) (2)
[Amer, Ghada, 1963-](#) (2)
[Auerbach, Tauba, 1981-](#) (2)
[Bourgeois, Louise, 1911-2010.](#) (2)
[Harkness, Hilary, 1971-](#) (2)

[More ▾](#)

Creator

Sort By: **Count** | [\(A-Z\)](#)

[Galerie Hauser & Wirth](#) (11)
[Cheim & Read](#) (8)
[Deitch Projects](#) (7)
[Mary Boone Gallery \(New York, N.Y.\)](#) (5)
[Sean Kelly Gallery](#) (2)

[More ▾](#)

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Sites

Search Page Text

Page 1 of 1 (90 Total Results)

Sort By: [Title \(A-Z\)](#) | [Title \(Z-A\)](#) | [URL \(A-Z\)](#) | [URL \(Z-A\)](#)

Title: net art, video, performance | [Annie Abrahams](#)

URL: <http://aabrahams.wordpress.com/>

Description: Artist Website

Captured **19** times between **Jan 11, 2012** and **May 5, 2013**

Videos: 63 Videos Captured

Subject: [Abrahams, Annie, 1954-](#)

Title: [Amy Alexander](#)

URL: <http://amy-alexander.com/>

Description: Artist Website

Captured **11** times between **Jan 11, 2012** and **May 5, 2013**

Subject: [Alexander, Amy](#)



April 16 Archive

Collected by: [Virginia Tech: Crisis, Tragedy, and Recovery Network](#)

Archived since: [May, 2007](#)

Description: The April 16 Web Archive captures a wide variety of content related to the April 16, 2007 tragedy at Virginia Tech. It includes memorial and tribute sites, commercial and non-commercial media, and other relevant web-based materials. This collection was specifically developed as a complement to the April 16 Archive (www.april16archive.org), which is dedicated to collecting and preserving individual stories, images, and files related to the events of April 16.

Subject: [Spontaneous Events](#), [Shooting](#), [virginia tech](#), [campus shooting](#), [violence](#), [United States](#)

Creator: [Brent Jesiek](#), [Center for Digital Discourse and Culture \(CDDC\)](#), [Virginia Tech](#)

Date: [May 21, 2007](#)

Collector: [Virginia Tech](#)

Narrow Your Results

There are no further ways to narrow your results.

Sites for this collection are listed below. Narrow your results at left, or enter a search query below to find a site, specific URL or to search the text of archived webpages.

Sites

Search Page Text

Page 1 of 1 (88 Total Results)

Sort By: [Title \(A-Z\)](#) | [Title \(Z-A\)](#) | [URL \(A-Z\)](#) | [URL \(Z-A\)](#)

Title: [ABC News: Massacre at Virginia Tech](#)

URL: <http://abcnews.go.com/US/VATech/>

Captured 7 times between [Jun 13, 2007](#) and [Apr 22, 2008](#)

Videos: 25 Videos Captured

Title: [The Associated Press: People at Virginia Tech remember mass shooting of 1 year ago](#)

URL: <http://ap.google.com/article>



Occupy Web Archive

UCLA Library

- Home
- About
- Site List
- Search
- Help
- Contact Us

Refine site list

lookup by site name

Go

Clear

Site list by topic:

- Arizona
- Bay Area
- Berkeley
- Boise
- Brazil
- California
- Canada
- D.F.
- Eugene
- Fresno
- Idaho
- Las Vegas
- Long Beach
- Los Angeles
- Mexico
- Nevada
- Oakland
- Olympia
- Oregon
- Pasadena
- Phoenix
- Portland
- Porto Alegre
- Reno
- Rio de Janeiro
- Sacramento
- Salvador
- San Francisco
- San Jose
- Santa Cruz
- Santa Rosa
- Sao Paulo
- Seattle
- South America
- Vancouver

Occupy Berkeley Hide Info

Latest Starting URL

<http://occupyberkeley.org/> [\(live link\)](#)

Description

Occupy Berkeley

Captured

- 04/29/12 08:16 AM
- 05/02/12 11:32 AM
- 05/07/12 02:33 PM
- 05/13/12 06:59 PM
- 05/20/12 08:09 AM
- 06/04/12 06:04 AM
- 07/02/12 06:42 PM
- 08/02/12 06:41 PM
- 10/04/12 04:58 PM
- 11/02/12 06:32 PM

Occupy Boise Show Info

Occupy Eugene Show Info

Occupy Fresno Show Info

Occupy Las Vegas Show Info

Occupy Long Beach Show Info

Occupy Los Angeles Show Info

Occupy Oakland Show Info

Occupy Olympia Show Info

Occupy Pasadena Show Info

Occupy Web Archive

UCLA Library

- Home
- About
- Site List
- Search
- Help
- Contact Us

Description

The UCLA Library Occupy Web Archive is a collection of websites documenting local Occupy movements and events on the west coast of the United States, Mexico, and Brazil. The geographic scope complements the UCLA Library's traditional print collection strengths from these regions. Extant websites are captured once a month.

<http://webarchives.cdlib.org/a/occupy>



Prisoners Rights and Political Prisoners Web Archive

New York University Libraries / Tamiment Library (Labor & the Left)

[Home](#) [About](#) [Site List](#) [Search](#) [Help](#) [Contact Us](#)

Refine site list

[American Gulag](#) [Hide Info](#)

Latest Starting URL

http://www.infoshop.org/prisons/public_html/ [\(live link\)](#)

Description

American Gulag - Prison News was an electronic discussion group, last active in 2009, for activists and others interested in defending the rights of prisoners in the US, and presented a radical critique of US prisons and the prison industrial complex.

Captured

03/15/11 06:30 PM

[Attica is All of Us](#) [Hide Info](#)

Latest Starting URL

<http://atticaisallofus.org/> [\(live link\)](#)

Description

Attica is All of Us is a group of Attica Correctional Facility survivors from the 1971 prison riot, their attorneys, friends, family and allies that raise awareness of prison conditions, call for the reform of the prison system and the end of mass incarceration, and racism in the prison system. The website contains a petition to close Attica, resources, an events calendar, and information on their Attica is Now program.

Captured

05/02/12 12:33 PM

08/02/12 04:20 PM

11/02/12 04:20 PM

[BreaktheChains.info](#) [Show Info](#)

[Business Of Detention](#) [Show Info](#)

[Campaign to End the Death Penalty](#) [Show Info](#)

Building Support

- The first step is to research, research, research!
 - Investigate the most common service providers
 - Understand the technology:
 - What web archiving does well
 - What web archiving does poorly
 - What web archiving cannot do
- When requesting funds for a web archiving project, you want to make sure stakeholders understand the limitations of web archiving

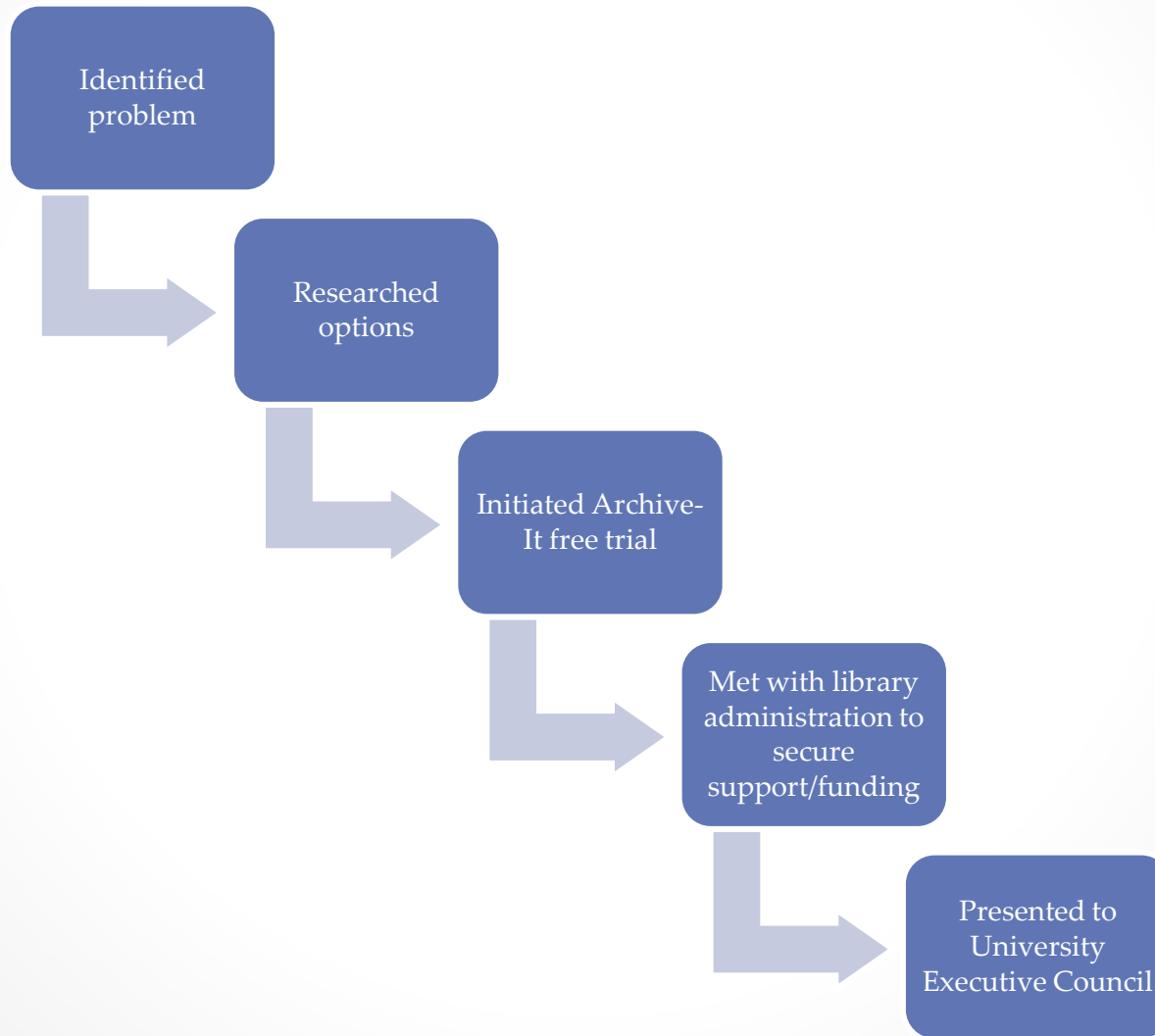
Building Support (cont.)

- Analyze your website
 - What vital documents are published exclusively online?
 - Can paper copies be derived?
 - Are paper copies derived and delivered to archives?
 - Are the original electronic records transferred to archives?
 - Compare what is exclusively published online to your retention schedule
 - How pressured are you to collect or seek out these records?
 - Level of participation from external departments
 - Don't forget about electronic ephemera
- Consider participating in a free trial from Archive-It

Meet with Administration

- Emphasize that web archiving is in addition to, not instead of traditional practices
- Collect a list of institutions in your region who have adopted web archiving
- Consider gathering feedback from current users or members of the SAA Web Archiving Roundtable

Our process



Where we went wrong

- Don't forget your IT Department, Webmaster, and/or CIO



Where we went wrong (cont.)

- Involving IT
 - In selecting Archive-It, we did not require any technical assistance
 - Viewed web archiving as solely the responsibility of the archives
- The problem
 - Heritrix crawls the IT servers
 - It could be perceived as a threat
- The courteous thing to do would be to involve the necessary IT individuals in early discussions

Questions?

Andrea Weddle
Andrea.Weddle@tamuc.edu
@aeweddle

Adam Northam
Adam.Northam@tamuc.edu
@atomic_red77

Jeremy Floyd
jeremyjfloyd@gmail.com
@jjamesfloyd

<http://www.archive-it.org/organizations/615>