

Project to Production: The Web Archiving Service DLF Fall Forum, 2010 Tracy Seneca







LIBRARY OF CONGRESS	ASK A LIBRARIAN DIGITAL COLLECTIONS LIBRARY CATALOGS GO Optio
The Library of Congress > Digital F	Preservation > Partners > Web-At-Risk Project
DIGITAL PRESERVATION	Partners
	Partners Home Partners by State Partners by Activity Alphabetical List
Search this site GO	Tools and Services Publications
 Digital Preservation Home News & Events Partners Tools & Services Publications Features Video Presentations 	PRESERVING web Sites
O Podcasts	Web-At-Risk Project
 Personal Archiving About the Program Contact Us 	Lead Partner: California Digital Library at the University of California Additional Partners: New York University and the University of North Texas. Additional support from Stanford University, the San Diego Supercomputing Center, the Arizona State Library, the Library of Congress, and the University of California libraries.
Related Resources	Web-at-Risk enables librarians and archivists to capture, curate, preserve, and provide access to Web-based government and political information. The primary focus of the
> <u>Digital Formats</u> Sustainability	collection is state and local government information, but it may grow to include Web documents from federal and international government, and from non-profit sources.

Web-at-Risk: 2005-2009

Today:

- Whirlwind tour
- Project to production
- Current / future work
- The DLF community

Build unique archives for local research communities

San Diego Local Governments Web UC San Diego Libraries) Archive		
Home About Site List Search Help Contact Us	3		
Description	Search		Quick Facts
The San Diego Local Governments Web Archive includes materials from government agencies around San Diego County (e.g. Citty of San Diego, Citty of Chula Vista, County of San Diego, etc.) and semi-official (e.g. League of Women Voters of San Diego) information related to topics of local interest.		Search	Sites: 40 Oldest tis: 09/19/07 Most recent site: 12/06/09
Photos used under Creative Commons licensing: harbor seal trolley more			

Geographically focused archives support local research

- Los Angeles
- Monterey Bay
- Orange County
- San Diego
- Santa Barbara

Topical archives support special research collections

• Guantanamo Bay / Tamiment Library

• California Water Districts / Water Resources Center Archives

THE TAMIMENT LIBRARY & ROBERT F. WAGNER LABOR ARCHIVES					
Guantanamo Bay Detention Camp Web Archive New York University Libraries / Tamiment Library (Labor & the Left)					
Home About Site List Search Help Contact Us					
 Previous 1 2 Next > 100 Days to Close Guantánamo and End Torture A campaign initiated by Witness Against Torture Act Against Torture - Stop Torture and Indefinite Detention - Home American Civil Liberties Union : Safe and Free American Civil Liberties Union , Accountability for Torture, Restore the Rule of Law, Enhanced Interrogation Techniques, Waterboard American Civil Liberties Union, Accountability for Torture, Restore the Rule of Law, Enhanced Interrogation Techniques, Waterboard 	lookup by site name Go Clear Click the "information" icon to see further information about the site, including the dates on which each site was captured. The site lookup box will find words anywhere in the site name.				

Archive Ephemeral Content



Analyze Site Change



Are there new documents on this site?

Are there documents in my archive that have been removed from the live site?

Easy to Use : Simple Workflow



Tailor Capture Settings

Edit Site: Deepwater Horizon Response on Facebook	
Capture Settings Scheduling Descriptive Data	Eacus on sita directory
* Required field *Site Name: Deepwater Horizon Respon	Focus on site, directory or page
*Seed URLs: Ex.: http://www.example.com	Set appropriate frequency
Scope: Page Edit Site: Deepwater Horizon Response	on Facebook
Max. Time: Brief Capture (1 hot	
Robots.txt overridden: No (may only be Capture Frequency: Off Image: Capture Frequency: Off Image:	✔ 26 ✔ 2010 ✔
Custom	Cancel Save (all tabs)

http://was.cdlib.org



web archiving service

build an archive log in

Capture todays web ° Build tomorrow's archives

			Home	Information for:	Partners	WAS News	Contact WAS
ublications are	e fragile, an ed. The We	ed our access to info d ready access to V b Archiving Service enge.	Veb resources	cannot be	• Res • WA	tion for ential Partners earchers S Curators omasters	
access to web cont	e been capturing ent. Public arch Access to pub	g and preserving critical web ves can be browsed or searc lished archives is open to all ls.	ched and provide persis	stent links to	2003 Califo 2007 South	r over an archive name ornia Recall Electio hern California Wild	n Ifires Web Archive
published or kept fo	nt you can captu r private study. 1	rre, analyze and archive web The WAS curator tools are ea _ provides training and guida	asy to use, fully hosted	d, and allow	AFL-CIO - C African Po	Influenza A (Swine Change to Win: the litics Web Archive Mass Media / New	open web archive
	ountains Wildfi	re Web Archive	2			n Web Archive State Government:	.ca.gov Web
Santa Cruz Mountains Wildfire Web Archive	Sites: 35	Iniversity Library, UC Santa (Oldest site: 05/23/08 The Santa Cruz Mountain Wi	Most recent site:			Water Districts We sm, Socialism, Trot	
	images, news Fire of May 20	aper articles, and personal b	ologs related to the Sa	nta Cruz Summit	I	no Bay Detention C	amp Web Archive

Archive Home Page



California State Government: .ca.gov Web Archive

UC Government Information Librarians

Home About Site List Search Help Contact Us

Description

The ca.gov domain is the 3rd largest domain in the U.S. government web. A great deal of that content is published by the State of California, and represents a rich source of material for researchers, students, the general public and state agencies themselves. This archive preserves access to hundreds of state agency sites that are captured on a regular basis.

State agencies utilize their websites to publish everything from press releases, agendas, minutes, events, reports and statistics. This materia...

more . . .

-	e 2	FC	•	
•				

Search

Quick Facts

Sites: 299 Oldest site: 10/24/08 Most recent site: 01/08/10

Automatic List of Sites



Los Angeles Local Government Web Archive

UCLA Library

Home About Site List Search Help Contact Us

 « Previous 1 2 3 Next »
 07 HomeWalk LA - Home Show Information
 1st District Gloria Molina Show Information
 2005 Greater Los Angeles Homeless Count 2 Show Information
 2007 Greater Los Angeles Homeless Count Show Information
 2007 Greater Los Angeles Homeless Count Hide Information
 2nd District Mark Riddley Thomas Hide Information
 Starting URL http://ridley-thomas.lacounty.gov/ (live link)
 Captured

12/11/08 02:20 AM 12/23/08 12:51 AM 12/27/08 04:13 PM lookup by site name

Go Clear

Click the "information" icon to see further information about the site, including the dates on which each site was captured.

The site lookup box will find words anywhere in the site name.

Search Across All Sites or Target Just One



San Diego Local Governments Web Archive

UC San Diego Libraries

Home About Site List Search Help Contact Us

Search		Your search terms will be found anywhere in the full text of the web
Type words or a URL:	Filter by:	pages and documents in this
border	All sites Search	archive. You can search for key words or for particular URLs.
« Previous 1 2 3 4 5 6 7 8 9 10 Next »	All file types	You can narrow your search to find the words in particular web sites, or to find only particular file types, such as PDF files.
Border Affairs - SignOnSanDiego.com Captured: Sun Dec 06 17:31:50 UTC 2009 www.signonsandiego.com/news/border-affairs/	36.8 KB text/html	You are not required to type a key word in your search; you can select a site to review all of the documents from that site.
	KB text/html Corporate Responsibility Thousands of site The CFO (<mark>Border</mark>	from that site.
The Border Captured: Sat May 17 12:04:20 UTC 2008 www.pbs.org/kpbs/theborder/history/hdrhistory.html	527 Bytes text/html	

University of California Curation Center, California Digital Library

WAS Snapshot: October 2010



Stats: Since January 2007
17 organizations
4,509 sites captured
33,619 captures run
21.7 terabytes
29 archives published

Project to Production

• Scale

 Infrastructure: ensure continuity of service through server upgrades

• Policy

– Setting up service agreements

Engage the users

(Needs assessment doesn't stop)

Assessment Path Final Report December 27, 2007	Site Summary: congressional budget office Ste created by: Tracy Seneca Capture Settings Scheduling Descriptive Data Capture History 1-1 of 1 CAPTURE DATE STATUS FILES DURATION ACTIONS 04/17/09 03:12 PM Settings: Host site only. 1h Captured by: Tracy Seneca 1.1 of 1	Site Tags berkeley (*) blog (*) books (*) conference (*) los angeles (*) open source (*) sacramento (*) (*) c add enters for disease control critical mass cyclist
Prepared by : Kathlean R. Murra kmurray@library. Univerity of Net	Limits: topic = Law School display 25 50 100 all 1-25 of 38 next > brief records site name only seed URL only select all soft by: site name URL	gress captured
	American Civil Liberties Union Student Chapter (University of Michigan) American Constitution Society (University of Michigan) Asia Law Society (University of Michigan) Tag:	since:
	Asian Pacific American Law Students Association (APALSA) (University of Michigan) Law Sc Black Law Students Alliance (BLSA) (University of Michigan) Schedule All Business Law Association (University of Michigan) All	
	Entertainment, Media, and Arts Law Student Association (University of Michigan) Keyword Environmental Law Society (University of Michigan) Image: Stepsilon of Michigan) Federalist Society (University of Michigan) Image: Stepsilon of Michigan)	name

"Production" is not just for us!

- How does WAS fit into existing workflows?
- Is there a 'web archivist'? Student workers?

Friday November 5th 10-11:30 am

Curators from UC San Diego, NYU, University of Michigan and UC Berkeley share their approaches to web archiving.

http://was.cdlib.org WAS News



Archiving the State of California

Learning from the archives





Archiving the Gulf oil spill

Improving support for collaboration



527 sites 9288 + captures 1.6 TB Began May 5

LSU tags relevant sites in Delicious CDL imports Delicious JSON feed into WAS

del	licious Home Bookmarks - People	▼ Tags ▼		Join Now! W
• Taqs	Recent oil and spill Bookmarks Recent Popular oil ×) spill × Type another tag	RESOURCES ABOUT US DEPAR	TMENTS HELP SERVICES A TO A	LOUISIANA STATE UNIVERSIT
4 JUL 10	Recovery effort falls vastly short of BP's promises s	INDEXES & DATABASES EJOURNALS GO	OVERNMENT INFORMATION SUBJECT GUIDES	LIBRARIANS OTHER RESOURCES
	The Road to Deepwater Horizon - Newsweek save The backstory on BP resultiv Carl Safina: The oil spill's unseen culprits, victims ' resultive bp business drilling en	Enth See Leb 11 May 2010 Coastal Spelles Institute Coastal Spelles Institute	A States of the second	
	Deepwater Horizon / BP Oil Spill Response Recer Ocean Service Office of Response and Restoration r timpagel	LSU Librar	ies Oil Spill Informa	ation Service
	Ocean Service Office of Response and Restoration	LSU Librar	ies Oil Spill Informa Search the Oil Spill Subject Guid	
	Ocean Service Office of Response and Restoration timpagel Visualizing the BP Oil Spill SAVE	LSU Librar	Search the Oil Spill Subject Guid	

Interoperability of CDL Services



UC 3 M	erritt	Logout Profile Contact UC3 Help Logged in as Tracy Seneca	
Collection home Ad	dd object Change collection 🔻		
Collection: CDL Quali Merritt > Collection: CDL Quality Ass	-	Object Lookup	
Primary Id	Object info 1 to 10 of 22 < prev next >	Search object author, title, date or identifier.	
ark:/99999/fk4p55nbx	General Drapery		
ark:/99999/fk4sx6d2q	Tracy Seneca. Tumble Bug, Luna Park. Sydney		
ark:/99999/fk4xp74sd	Tracy Seneca. Golden Dragon Restaurant Sign (2009-09-01T00:00:00)	Overview	
ark:/99999/fk42f7n8m	UC3 Merritt	Logout Profile Contact UC3 Help Logged in as Tracy Seneca	
ark:/99999/fk49w0f9v			
ark:/99999/fk4fn1610	Collection home Add object Change collection 🔻		
ark:/99999/fk4kd1xr8			
ark:/99999/fk4q52pgd	Object: ark:/99999/fk42f7n8m	Object Lookup	
ark:/99999/fk4tx3f7p	Merritt > Collection: CDL Quality Assurance > Object: ark:/99999/fk42f7n8m	Go	
	Orange Public Library. Crate label for the Orange Fruit Company (1945)	Search object author, title, date or identifier.	
ark:/99999/fk4zp45zc	primary identifier: ark:/9999/fk42f7n8m title: Crate label for the Orange Fruit Company creator: Orange Public Library date: 1945 local id: last modified: 2010-10-20 01:39 pm created: 2010-10-20 01:39 pm total size (all versions): 1.9 MB total versions: 1 Download object Merritt is a service of the University of California Curation Center	Current Version Version 1: 2010-10-20 01:39 pm • master.tif • mrt-erc.txt • thumbnail.gif • use.jpg	-
	Copyright © 2009-2010 The Regents of the University of Control of the C		

DLF Community

Tools

- WAS
 - UC libraries, NYU, Stanford
- Archive-It
 - Columbia, Rice, Indiana...
- Other
 - Library of Congress, North Texas, Los Alamos

Common issues/developments

- Memento
- Policy
 - Can / should Facebook block government pages from capture?
- Collection
- Collaboration
- Imagination

In-depth demos: Nov 8th and 16th Workflows discussion: Nov 5th

- <u>http://www.facebook.com/webarchiving</u>
- <u>http://was.cdlib.org</u> > WAS News

	Web Archiving Service Wall Info Photos Discussions +			
	What's on your mind?		b	uild an archive log
web archiving service	Attach: 🗐 🔳 🗭	Everyone Share		_
No Day Lond Sectorem Re Autors	Web Archiving Service + Others Web Archiving Service			
An	Web Archiving Service Today we meet with partners from Ithaka s + r to discuss a new rou needs research. Thanks to everyone who's con http://www.cdlib.org/uc3/partners/webatrisk.h	und of Web archiving user ntributed their time!	WAS News	Contact WAS
	University of California California Digital Library Find out more about the pa behind the Web Archiving	ast and current work of the NDIIPP grant		
capture, analyze, preserve	Yesterday at 1:20pm · Comment · Like · Share · Pro			
websites publish, browse, search web archives was.cdlib.org	Web Archiving Service On Monday May 10, hosting an on-site Personal Archiving Day. A re provides further information personal archiving http://wp.me/pCH4C-1gV May 7 at 11:27am · Comment · Like · Promote	ecent CDL blog post		
California Digital Library	Lynn Jones likes this.			