

WAS User Guide

Updated December 2010

Contents

- **WAS: Quick Overview**
- SITES
 - **Selecting Sites to Capture**
 - **Creating / Editing Sites**
 - **WAS Default Capture Settings**
 - **Managing Sites**
 - **Tagging Sites**
 - **WAS Site Registry**
- CAPTURES
 - **Starting and Stopping Captures**
 - **View Captures**
 - **Capture Overview Report**
 - **Search Tips**
 - **Display Captures**
 - **Compare Captures**
- PUBLIC ACCESS
 - **How Public Access Works**
 - **Steps to Publish a Project**
 - **Project Configuration Tips**
 - **Rights Management**
- ADMINISTRATION
 - **Your Account**
- OTHER TOOLS
 - **WAS Browser Button**
- OTHER GUIDES
 - **Collection Plan Guidelines [PDF]**
 - **Collection Plan Form [DOC]**
 - **WAS Rights Management Practices [PDF]**
 - **Project Administrator Guide [PDF]**
 - **Institution Administrator Guide [PDF]**
 - **Glossary**

WAS Quick Overview

Step 1: Define a site

Sites are the starting points for creating your Web captures. You will typically provide a name and at least one seed URL to define a site. You can also provide descriptive information, set up capture scheduling and change default capture settings.

Step 2: Capture the site

Once you create a site entry, you will have the option to capture the site. When you click the **Capture** button, a *capture* will be created and scheduled. Your capture may be queued behind others, so it may not run immediately. When the crawler has finished, you will be notified by email that the capture is complete.

Step 3: Review results

After a capture is complete, you can browse, search and display the archived content. Content from each capture is saved separately. Each capture has an overview report and detailed reports. These reports let you check for errors and assess the effectiveness of your capture. You can also compare different captures of the same site on different dates.

Step 4: Public Access

You can publish WAS projects to create publicly accessible web archives. End-users can search the entire archive, or limit searches to particular sites. End-users can also browse the web archive by site name. You can configure your archive as needed to provide descriptive text, navigation to related sites, and your organization's branding.

Selecting Sites to Capture

When investigating sites to capture, consider the following:

How much of the site do you want to capture?

Explore the site to determine if you only want pages from that site or if you also want to capture linked pages from other sites. In some cases, you may only want a specific page, or pages within a specific directory.

How many servers or hosts does the site consist of?

In some cases, particularly sites that provide extensive multimedia or PDF files, a web site is delivered from more than one server. When you create a WAS site entry, you'll be able to enter as many seed URLs as needed to effectively capture the site.

Is the site protected by a robot exclusion file?

In some cases, the site owner or administrator may have placed a file on the site prohibiting crawlers from capturing some or all of the site's content. To see if the site you want prohibits crawlers, try viewing the site's robots.txt file (<http://www.example.com/robots.txt>).

General Web Crawler Considerations

- Web crawlers have difficulty capturing sites with content generated from databases instead of static HTML pages.
- The Heritrix crawler can only capture URLs beginning with "http" or "https". You cannot capture FTP sites with the crawler.

Creating/Editing Sites

You will need to add an entry for each site you plan to capture. WAS "Sites" usually correspond to web sites, but may not always. In some cases you may choose to capture only specific files or a specific directory of a site. In other cases, you might define a site broadly. For example, you might consider the National Library of Medicine to be a site, or you might consider it to be part of the National Institute of Health site.

Create Site

Capture Settings | Scheduling | Descriptive Data

* Required field

* Site Name:

* Seed URLs:
Ex: <http://www.example.com>

Scope: Host site

Capture Linked Pages: No Yes

Max. Time: Brief Capture (1 hour)

Cancel Save (all tabs)

Capture Settings

Site Name

What you name a site is up to you. A site name does not need to match the HTML <title> text of its pages. During a capture, the HTML <title> information is gathered for each page along with the rest of the content.

The site name will display on the **Site List** of a public archive. Consider that a broad range of users might someday see this site name, so make sure it is adequately descriptive.

Seed URL(s)

The URL(s) where the crawler should go to begin capturing files.

You may enter more than one URL if the site has multiple subdomains or servers. Example:

`http://www.nasa.gov/`
`http:// science.nasa.gov/`
`http://history.nasa.gov/`
`http://visibleearth.nasa.gov/`

If you have already entered `http://www.nasa.gov/`, you DO NOT also need to enter URLs beneath that, such as: `http://www.nasa.gov/directory/file.html`

Seed URLs must begin with **http** or **https**. The crawler cannot capture FTP sites.

Scope

How much of the site you plan to capture. Specifies how much of the target site and whether information will be included from related sites.

Page

This setting will only capture the specific files listed in the Seed URL box. If you have HTML files listed here, any embedded images or multimedia files will also be captured.

Directory

To capture the site of an organization that is part of a larger organization, you may need to limit your capture to a specific directory. Example: `http://www.ucop.edu/budget/` With this setting, the crawler will look for files whose URLs begin with the text in the seed URL.

Host site (default)

This is the basic site capture option. It will capture the website you specify as completely as possible.

Capture linked pages (yes or no)

This option can be combined with any of the three scopes above. If you select **yes**, the crawler will also capture any immediately linked pages and their components (images etc.), no matter what site the page is served from. The crawler will not continue past immediately linked pages from other hosts. The default **no** setting will not collect links from other sites.

Max Time

The maximum amount of time the crawler will gather files.

There are two time-limit settings: brief (1 hour) and 36 hour. The capture will stop before the time limit if the site has been completely captured in less time.

Scheduling Options

You can schedule sites be captured on a daily, weekly, monthly or less than monthly (custom) basis.

You can still use the capture button to run a scheduled capture at any time; this will not impact existing capture scheduling.

Edit Site

The screenshot shows the 'Edit Site' interface with three tabs: 'Capture Settings', 'Scheduling', and 'Descriptive Data'. The 'Scheduling' tab is active. Under 'Capture Frequency', the 'Off' radio button is selected. Other options are 'Daily', 'Weekly', 'Monthly', and 'Custom'. To the right, the 'End Date' is set to April 15, 2010. At the bottom, there are 'Cancel' and 'Save (all tabs)' buttons.

- You must select an end date for daily captures. If the max time setting for the site is full (36 hours) it will be updated to brief (1 hour) to prevent our crawlers from stressing the server.
- When you set a capture frequency, the confirmation screen will tell you when the site will be queued for capture. Captures may not run exactly at the time scheduled if there is a long line of sites in the queue.
- Project administrators will receive an email notification every time a scheduled capture has started.
- The Manage Sites screen will indicate when a site has a capture frequency. It will also indicate who scheduled the site to be captured.
- The home page will provide a link to a list of captures that are scheduled to run.
- The scheduling tab on the site summary screen will tell you the date and time the capture will run next.

Descriptive Data

You can enter descriptive text about this site and the following selected Dublin Core fields: Creator, Publisher, Subjects, Coverage (geographic). The **Description** will display in a public archive on the **Site List** page. The remaining descriptive elements will display when the user chooses the **Show Details** tab when viewing archived content.

WAS Default Capture Settings

Here is information about some of the important default crawler settings for all captures:

Scope (depth) settings

All jobs will default to "max-link-hops = 25" scope setting. This means that the crawler will follow links from the URL(s) you enter for the site, and continue gathering files until it gets 25 links away from the starting point. This should provide a thorough capture of most sites.

Politeness settings

WAS crawlers are configured with the highest "politeness" settings. This means that the crawler will force a delay between requests to the same server, so that the content owner's site performance is not adversely affected by crawler activity.

Robot Exclusions

WAS honors robots exclusion files posted by content owners. Contact washelp@ucop.edu to review options if a robots.txt file is preventing a site from being captured effectively.

Order.xml

One of the reports available to you for each capture is the "order.xml" file. This file contains a record of every setting the crawler used including all default settings.

Managing Sites

The Manage Sites screen lets you edit site entries, view site summaries, view capture histories, start captures, view capture progress and stop captures. You can also look up sites and either start or reschedule sites in batches.

The Site List

Web Archiving Service

Home | Help | Contact WAS

Logged in as: tseneca | Log Out

Project: California Government Sites Test | Change Role

Sites Captures Administration Public Access

Manage Sites

display: 25 | 50 | 100 | all
brief records | site name | seed URL

1-25 of 295 < Prev 1 2 3 4 ... 12 Next >

Select all sort by: site name | URL

Acupuncture Board (3)
Seed URL(s): <http://www.acupuncture.ca.gov/>
Tags: Health
Status: Preserved
Current settings: Host site only, 1h
Last captured 2 days ago
EDIT CAPTURE DEACT DELETE

Area VI Developmental Disabilities Board (3)
Seed URL(s): <http://www.areaboard6.ca.gov/>
Tags: Health
Status: Canceled
Current settings: Host site only, 1h
EDIT CAPTURE DEACT DELETE

Arts Council (2)
Seed URL(s): <http://www.cac.ca.gov/>
Tags: Culture, Education
Status: Preserved
Current settings: Host site only, 1h

Limit site list by:

All active sites
 In progress
 Never captured
 Inactive sites

Captured since:
month day year

Tag:
Everything

Scheduled:
Any

Keyword:

site name
 seed URL

limit clear

The default view of your site list shows brief records for the first 25 sites. This includes the site name, seed URL(s), tags, current status, settings, date last captured and any future scheduling information. You can change the format of this screen to show just the site names or just the seed URLs. You can also adjust the number of sites that show in the list.

- Clicking on the site name will bring you to the detailed summary screen for that site.
- Click on the number of captures to see details for each capture history.
- Click on a seed URL to open a second window displaying the live web site. This can help you test the seed URL to see if it is still valid.
- Clicking on a tag will list all other sites with the same tag.
- Move the mouse over a status to see what that status means.
- The **Edit** icon will let you change the site name, description and capture settings.
- The **Summary** icon will show you the site details, including the full description, any metadata, and the detailed capture history
- The **Capture** button will start a capture.
- The **Deact** icon will remove this site from the (default) active sites list. Use this when you don't expect to capture the site again in the future. This will not affect any content you have already captured from this site.
- The **Delete** icon will delete a site entry entirely. This is only accessible if there are no captures associated with a site.

Navigating the List

Limit site list by:

All active sites
 In progress
 Never captured
 Inactive sites

Captured since:
month day year

Tag:
Everything

Scheduled:
Any

Keyword:

site name
 seed URL

You may accumulate hundreds of sites in your site list. The **Manage Sites** screen offers some tools to help you navigate that list. The limiting options on the right-hand side of the screen can be combined to help you zero in on the sites you need to work with.

- You can show all active sites, those currently being captured, those never captured and sites you have de-activated.
- Flexible date-limiting options are available.
- You can limit to only sites with a particular tag. You can also list only sites that have not been tagged yet.
- You can list all of the sites that are scheduled on a daily, weekly, monthly or other basis.
- You can look up a site by a keyword in the name, or in the seed URL.
- Limiting options can be combined; for example, you can list all sites scheduled to run weekly with "blog" in the site name.

Batch Actions

Each site in the list has a checkbox. You can use the **Capture Selected** or **Reschedule selected** buttons at the bottom of the list to capture or reschedule all of the sites you have checked.

Tagging Sites

When an project consists of a few hundred or several hundred websites, you may need additional tools to help you organize and navigate your site list. You can create tags to organize your sites both on the **Manage Sites** screen of the curator tool, and on the **Site List** screen of a public archive. You can add multiple tags to any site. Tags are tools to help you organize your sites, and to help researchers understand what your archive contains.

Creating Tags

If you're creating a tag for a site that already exists, go to the **Descriptive Data** tab for that site. The tagging controls will appear on the right-hand side of the screen.

Site Summary: Area VI Developmental Disabilities Board

Description: Area Board VI works to protect the rights of people with developmental disabilities and their families, who live in Amador, Calaveras, San Joaquin, Stanislaus, and Tuolumne counties.

Creator:

Publisher:

Subjects:

Geographic coverage:

Site Tags

H

Health

User tags to organize the site list on your [manage sites](#) page and the public [site list](#) page. For more information see [Tagging Sites](#)

Tips: Descriptive Data

The descriptive data you provide here should apply to the entire site. The description will display on the [site list](#) page of a public archive when the user clicks on [show info](#).

You can type an entirely new tag into the tag entry box, or type an existing tag. The tag box will list any existing tags that begin with the text you type.

When you click the add button, the tag will be added to the site entry. You will be able to remove a tag by clicking on the red X next to it.

When you delete a tag from a site, it will remove the tag from that site, but won't affect any other sites using that same tag.

Site Summary: Area VI Developmental Disabilities Board

[Capture Settings](#) | [Scheduling](#) | [Descriptive Data](#) | [Capture History](#)

Description: Area Board VI works to protect the rights of people with developmental disabilities and their families, who live in Amador, Calaveras, San Joaquin, Stanislaus, and Tuolumne counties.

Creator:

Publisher:

Subjects:

Geographic coverage:

Site Tags

Health ●

Use tags to organize the site list on your [manage sites](#) page and the public [site list](#) page. For more information see [Tagging Sites](#)

Tips: Descriptive Data

The descriptive data you provide here should apply to the entire site. The description will display on the [site list](#) page of a public archive when the user clicks on [show info](#).

If you are creating a new site, you'll have the option to add a tag after you save the site entry.

Using Tags for Navigation

There is a drop-down selection list for tags on the Manage Sites screen, and all tags will display in the Site List of a public archive:

Manage Sites

1-25 of 25 display: 25 | 50 | 100 | all
brief records | site name | seed URL

Select all sort by: site name | URL

- Arts Council
- Bureau for Private Postsecondary and Vocational Education
- California Community Colleges System Office
- California Department of Education
- California Education Audit Appeals Panel
- California High School Proficiency Examination
- California Science Center
- California State Summer School for Mathematics and Science (COSMOS)
- California State Summer School for the Arts
- California State Teachers' Retirement System (CalSTRS)
- Center for California Studies, CSU Sacramento
- Commission on Teacher Credentialing
- CSU Board of Trustees

Limit site list by:

All active sites
 In progress
 Never captured
 Inactive sites

Captured since:

Tag:

- Education
- Everything
- Untagged
- All tagged
- Culture
- Economy
- Education
- Elections & Politics
- Environment
- Government
- Health
- Labor
- Law & Judicial
- Science
- Social Services
- State Infrastructure
- Transportation

California Government Sites Test
California Digital Library Quality Assurance

Home About Site List Search Help Contact Us

Refine site list

lookup by site name
Go Clear

Site list by topic:

- Culture
- Economy
- Education**
- Elections & Politics
- Environment
- Governor
- Health
- Labor
- Law & Judicial
- Science
- Social Services
- State Infrastructure
- Show All Topics

Showing sites with the topic Education

- Arts Council [Show Info](#)
- Bureau for Private Postsecondary and Vocational Education [Show Info](#)
- California Community Colleges System Office [Show Info](#)
- California Department of Education [Show Info](#)
- California Education Audit Appeals Panel [Show Info](#)
- California High School Proficiency Examination [Show Info](#)
- California Science Center [Show Info](#)
- California State Teachers' Retirement System (CalSTRS) [Show Info](#)

Tags vs. Subjects

- Tags provide archive navigation; subjects are strictly descriptive.
- Tags will display on the site list screen. Subjects will display when you look at the detailed record for a rendered site.
- Subjects are provided as part of the basic Dublin Core elements to describe each site. These may eventually be used to generate OAI records that can be harvested into a catalog.
- Tags should be simple and brief; subjects may be complex LCSH headings, if you so chose.

Tagging Recommendations

- Limit tag length to less than 28 characters. This is the width of the tag list on the public Site List page.
- Don't use a tag unless it applies to at least 5 sites. Following a link that only leads to one or two items can be disappointing, so make it worth the researcher's while to click on a tag in the site list.
- You may not need tags if there are less than 50 sites in your archive.
- You may not need to tag every site in your site list; you can just use them to highlight major themes for the content in the archive.
- If you have already made your archive publicly accessible, the tags you apply to sites will show in the Site List right away. When you are setting up tags for a public archive, you should determine what the tags will be in advance, and set aside time to add them to sites in one session.

Managing Tags

When you delete a tag from a site on the Descriptive Data screen, it will remove the tag from that site, but won't remove it from your list of tags. You can edit or delete tags in your tag list by choosing the **Administration** menu, then **Manage Tags**.

Manage Tags for California Government Sites Test

Tags in Project

<input type="checkbox"/> California (0)	<input type="checkbox"/> Elections & Politics (3)	<input type="checkbox"/> Labor (8)	<input type="checkbox"/> State Infrastructure (10)
<input type="checkbox"/> Culture (16)	<input type="checkbox"/> Environment (59)	<input type="checkbox"/> Law & Judicial (17)	<input type="checkbox"/> Transportation (5)
<input type="checkbox"/> Economy (32)	<input type="checkbox"/> Governor (5)	<input type="checkbox"/> Science (8)	
<input type="checkbox"/> Education (25)	<input type="checkbox"/> Health (41)	<input type="checkbox"/> Social Services (27)	

Remove Selected Change Selected

From the Manage Tags screen you can see all tags being used in the project and how many sites are associated with each tag. If you select one or more tags then choose **remove selected**, it will delete the tags from all sites and remove the tag from the tag list. If you select one or more tags and click the **change selected** button, you will be able to edit the tag text. The new text will appear for all sites associated with that tag.

Old Value	to	New Value
Environment	→	<input type="text" value="Environment"/>
Governor	→	<input type="text" value="Governor"/>

Site Registry

The Site Registry will tell you which websites have been archived by other WAS curators. You can search the registry by keyword in the site name or by seed URL.

Curators have access to the WAS Site Registry from the Sites menu in the top navigation bar. Institution Administrators can access it via the Site Registry link on the Admin home page.

This registry is a report only; it does not search captured content and will not allow access to captured content. The registry is available only to WAS curators; it is not an access system to captured content. The registry will report information about sites included in both public and dark archives.

Searching the Registry

The words you type will be found anywhere in the site name or seed URL. Multiple terms will be connected with AND. Sites with identical seed URLs but with different names will be found together by a URL search. Search results will include a number indicating how many separate captures of that site exist.

WAS Site Registry: Results

<p>Search</p> <p><input checked="" type="radio"/> Site name <input style="width: 150px;" type="text" value="water"/></p> <p><input type="radio"/> Seed URL <input style="margin-left: 100px;" type="button" value="look up"/></p>	<p>Browse Site Name</p> <p># A B C D E F G H I J K L M N O P Q R S T U V W X Y Z</p>
--	---

141 results

Click a site name for details

- [ABAG Bay Area Water \(1\)](#)
- [Alameda County Flood Control & Water Conservation District \(5\)](#)
- [Alameda County Water District \(3\)](#)

If multiple organizations are capturing the same site, the results will be listed separately – they will not be merged into a single entry. In the example below, two organizations are capturing the Southern California Association of Governments. Click on the site name to see details about who is capturing it and the number of times captured:

WAS Site Registry: Results

Search

Site name

Seed URL

Browse Site Name

A B C D E F G H I J K L M
N O P Q R S T U V W X Y Z

2 results

Click a site name for details

[Southern California Association of Governments \(8\)](#)

[Southern California Association of Governments \(6\)](#)

Created by: UC Riverside Libraries
 In project: Riverside California Inland Empire Web Archive
 Publicly viewable: No
 Seed URLs: <http://www.scag.ca.gov>

CAPTURE DATE / SETTINGS	STATUS	FILES	SIZE	DURATION
12/22/08 04:31 PM <small>Settings: Host site only, 30h</small>	Preserved	11539	7.2 GB	36h
11/16/08 08:18 AM <small>Settings: Host site only, 30h</small>	Preserved	3952	7.2 GB	19h 36m 18s

The details will show context about the site being captured; the organization, which project, whether it is publicly available, the seed URL and information about each capture.

IMPORTANT

If you find that another organization is capturing a site of interest to you, it is still possible that:

- They may delete their captures
- They may decide not to provide public access to the content
- They may have used different capture settings than you would choose

Starting and Stopping Captures

Starting

You can schedule captures from the **Edit Site** screen, or you can capture sites on demand or in batches from the **Manage Sites** screen.

After capture starts

You will get a confirmation screen showing capture status and your capture will be queued by the crawler. As soon as the capture starts, you will see the status change to **running** and you'll see updates of the number of files captured. This status information is also available on the "Manage Sites" and "View Captures" screens.

When a capture is in progress, you can continue to use the Web Archiving Service, or logout without affecting captures in progress. You will receive an email when each capture has finished.

Stopping

Once the capture progress information is visible, you will also see a link allowing you to stop the capture. You will be asked to confirm that you want to stop the capture. It may take a moment for this action to take effect.

Note that you will not be able to view reports or content for captures you stop. When you stop a capture, it will show with a status of "Canceled". We recommend that you delete these captures.

View Captures

The **View Captures** screen lists all of the sites that you have begun to capture. The number next to each site name indicates how many times a capture has been started. This number includes captures currently in progress as well as captures you have canceled. When there are two or more completed captures for a site, you will also see a **compare** link.

Click on a site name to see a list of captures associated with that site. Basic information about each capture is provided, including the date and time the capture started, who started it, settings selected, status, number of files captured, and the capture duration.

View Captures

Click  to view the captures for a site.

1-25 of 124 ◀ Prev 1 2 3 4 5 Next ▶ display: 25 | 50 | 100

SITE NAME / CAPTURE DATE	STATUS	FILES	DURATION	ACTIONS
 Acupuncture Board (3)				Compare
12/03/10 02:09 PM Settings: Host site only, 1h Captured by: Scott Fisher	Preserved	71	39s	View Results 
11/30/10 11:02 AM Settings: Host site only, 1h Captured by: Scott Fisher	Preserved	396	14m 15s	View Results 
11/22/10 04:24 PM Settings: Host site only, 1h Captured by: Tracy Senese	Preserved	396	14m 3s	View Results 
 Area VI Developmental Disabilities Board (3)				Compare
 Arts Council (2)				Compare
 Assembly Democratic Caucus (1)				
 Assembly Republican Caucus (1)				

Limit capture list by:

Captured since:

Tag:

Scheduled:

Keyword:

site name
 seed URL

For each capture, you can **View Results** or click the  delete icon. **The DELETE ICON will remove content from the archive.**

When you click on **View Results** for a capture, you will go to the overview report screen, which will give you links to search or browse the results.

Capture Overview Report

When you choose to display your captured content, the first screen you see is the overview report. This page offers a summary of the capture result, provides tabs to **search** the captured content or browse all captured **files**, view a map of where content was provided from, and provides links to more detailed reports.

Link to Home Page

The first link in this report lets you render the home page for this capture. This link looks for an exact match to the first seed URL you enter for this site. It may not always find the right page; if the site contained a forwarding mechanism or alias, the real home page URL may be different.

General Info

This area conveys basic capture status, such as size and number of files.

Capture Settings

This section repeats the settings you chose to capture the site. If the capture reaches a time limit, you will be able to see the limit you originally selected. A link to edit these settings is available if needed. Edited settings will be applied to future captures.

Other Statistics

This section includes four reports that you can expand on this screen. Click the  **plus** icon to expand these reports.

Robot Exclusions

A robot exclusion is a method that site administrators use to convey instructions to web crawlers. This can be a separate text file or a META tag placed within an individual page. A robots exclusion rule may prohibit all content from being captured, specific pages or specific directories.

A crawler can be configured to ignore robot exclusions; the WAS will honor robot exclusions by default.

The overview report will list only robots.txt files encountered on the seed URL servers; these will have the strongest impact on your results. Results may sometimes be impacted when there are robots.txt files restriction the capture of material from related hosts.

For help interpreting rules in robots.txt files, see **Web Server Administrator's Guide to the Robots Exclusion Protocol**

Hosts

This section of the report can tell you when a site may require more than one seed URL to capture, particularly if you selected a host site + linked pages setting.

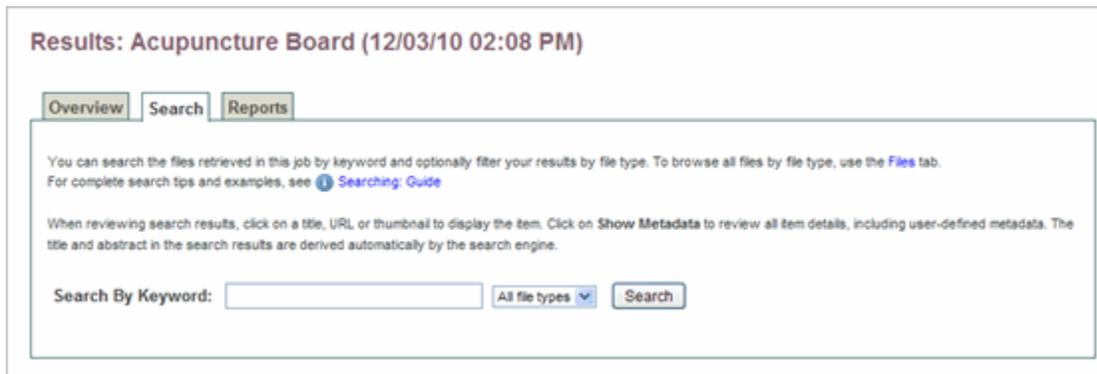
Response Codes

This area includes the most frequently returned response codes. This gives you a sense of how many broken links and other problems the crawler encountered.

Mimetypes

This area includes a list of the different file types or mimetypes found in this capture.

Search Tips



When you search in the curator view, you are searching a single capture, not the entire project. Curator search tools are intended to support QA, and will help you understand the scope and effectiveness of a capture.

A keyword search will find the terms you type anywhere in the full text or URL of each file. **Example:** *index* will retrieve any file with the term *index* anywhere either in the URL or the full text. Full text terms are found in HTML, PDF, Word, Excel, Flash, and text files.

Multiple search terms will return results containing both terms anywhere in the text. **Example:** *fishery plan* will find records referencing the *fishery protection plan*.

Use quotes to require the search terms immediately next to one another. Example: "*financial information*" will find only records with the exact string *financial information*.

Do not use the boolean operators AND, OR, NOT in your search; the search tool will ignore these, and will always combine your terms with AND.

You can limit a search to selected types of files: HTML, PDFs, MS Word, Images, Audio or Video.

You do not need to type any keywords to use the file type limit dropdown. For example, if you select PDFs from the filetype dropdown without including a keyword, you will get a list of all PDFs in the capture.

Display Capture Results

Once you have selected the Home Page link, a search result or a file from the files list, the file you've selected will display in a new browser window. There may be a delay when you display older captures; older captures may need to be retrieved from the repository and are being uncompressed for you to view.

Top Navigation Bar

The Society of St. Vincent de Paul :: Council of Los Angeles
http://www.svdpla.org/
Tue Aug 14 17:34:14 -0700 2007
CAPTURE [show Detailed Record](#) Close

QUICK LINKS
Home
About SVDPLA
Donate
Volunteer
Conferences
Programs
News & Events
Directory

SOCIETY OF ST. VINCENT DE PAUL COUNCIL OF LOS ANGELES

Donate Volunteer Get Help

About SVDPLA

The displayed file will open underneath a page heading that indicates the material is part of a web archive. The heading provides the page title and URL, and provides a link to a detailed record for this file.

Detailed Record

The Society of St. Vincent de Paul :: Council of Los Angeles
http://www.svdpla.org/
Tue Aug 14 17:34:14 -0700 2007
CAPTURE Show File Close

System-Defined Metadata

Title: The Society of St. Vincent de Paul :: Council of Los Angeles
Archive Date: Tue Aug 14 17:34:14 -0700 2007
URL: http://www.svdpla.org/
Abstract: ... The Society of St. Vincent de Paul :: Council of Los Angeles homelink QUICK LINKS . History . Board of Directors . Employment Opportunities . Promo ...
Size: 10.6 KB
Mimetype: text/html

User-Defined Metadata

Site Name: Society of St. Vincent DePaul, Council of Los Angeles.
Site Description: Contains information of interest to research on local homelessness issues.
Creator: Society of St. Vincent DePaul
Subjects: Vincentians Homeless Services
Coverage (Geographic): Los Angeles, CA

Comment(s) for this file:

Add Comment

Collections that include this item

None

Add to: -- Select a Collection -- Add

The detailed record shows all of the automatically derived metadata about the item, as well as any metadata that you supplied for the web site the page is from. You also have the ability to add comments about this particular page.

Compare Captures

Whenever a site has been successfully captured more than once, you will have the opportunity to compare the results for any two dates the site was captured. This feature is available on the **View Captures** screen, or from the **Captures** menu > **Compare Captures**.

When you choose the **Compare** link, you will need to specify which two captures you want to compare. If there is a long list, you should examine the list details carefully. The original scope setting, the number of files found and the duration have been provided to help you choose captures to compare.

Compare Captures for Governor's Office of Planning and Research

Captures available for comparison are listed below. Check two captures to compare and click "Submit."

	CAPTURE DATE	STATUS	SCOPE	FILES	DURATION
<input type="checkbox"/>	02/01/09 09:36 PM	Available	Host site + linked pages	2332	7h 15m 20s
<input type="checkbox"/>	10/25/08 09:23 PM	Available	Host site + linked pages	2595	7h 15m 24s

What the Comparison Will Tell You

- **If the captures were run with different settings:** the comparison screen will show you the impact of changing your settings. This can help you decide if "host only" scope is more effective than a "host + linked pages" scope, if you've tried both.
- **If both captures were run with exactly the same settings:** the comparison screen will show how the site has changed between the two times it was captured.

The Comparison Screen

Compare Captures for Governor's Office of Planning and Research

Earlier capture date: **10/25/08 09:23 PM**

Scope: Host site + linked pages, Files: 2595, Duration: 7h 15m 24s

Later capture date: **02/01/09 09:36 PM**

Scope: Host site + linked pages, Files: 2332, Duration: 7h 15m 20s

Limit: PDF

CHANGED (2) Documents in both captures, content not identical

NEW (24) Documents in later capture, not in earlier

- <http://www.opr.ca.gov/ceqa/pdfs/june08-ceqa.pdf>
- http://www.opr.ca.gov/military/docs/Military_Mailing_Addresses_SB1462.pdf
- <http://www.opr.ca.gov/planning/publications/2009bol.pdf>
- http://www.opr.ca.gov/planning/publications/AB_642_Memo_12_31_2008.pdf
- http://www.opr.ca.gov/planning/publications/Memo_re_CDFG_fees_2009.pdf
- http://www.opr.ca.gov/sch/docs/G-August_16-31-2007.pdf
- http://www.opr.ca.gov/sch/docs/G-December_1-15-2008.pdf
- http://www.opr.ca.gov/sch/docs/G-December_16-31-2008.pdf
- http://www.opr.ca.gov/sch/docs/G-November_1-15-2008.pdf
- http://www.opr.ca.gov/sch/docs/G-November_16-30-2008.pdf

< Prev 1-10 of 24 Next >

MISSING (5) Documents in earlier capture, not in later

UNCHANGED (387) Documents in both captures, content identical

The comparison screen is divided into four sections:

- **Changed:** lists URLs that were present in both captures, but that have changed in size since the first time they were captured. Note that for HTML files, this may not always indicate meaningful change in the content, as it will include changes in dynamic ads, page visit counters, etc.
- **New:** lists files that appeared in the later capture, but not in the earlier one. If the two captures used the same settings, this may serve as a list of new publications.
- **Missing:** lists the files that were in the older capture, but are not in the more recent one. Note that these files are still stored in the Web Archiving Service, so even if they have been removed from the live web, they are still part of your archive.
- **Unchanged:** Lists files that appear in both captures and have not changed at all.

Expand each section and browse the list by clicking on the plus-sign icon (circled in red on image). Files will be listed ten at a time by URL; you will have a **Next** link to take you through the list.

Limiting Comparison by Filetype

There is a dropdown menu at the top of this screen to limit your comparison to specific filetypes. This will allow you, for example, to see only the new PDF files since the previous capture.

How Public Access Works

The projects you create can be published to become publicly accessible archives. Public archives will be accessible to anyone, can be searched by keyword or URL. Searches can be limited by site and by file type. End-users can also browse a site list and see information about how often each site has been captured. Statistics about the archive, such as the number of sites included, the oldest date of capture and the most recent date of capture will be automatically generated to help end-users understand the scope of the archive. Each displayed document will show the archival URL for that document, enabling researchers and librarians to create citations or catalog records for stable, archived files. End users will have a contact form to submit questions about the archive; all messages will be sent to the project administrator and cc'd to the CDL Web Archiving Service support staff.

The project administrator can customize project settings to control public access and tailor the appearance of the archive. This includes adding descriptive text, customized banner images, adding links to related resources and suppressing particular URLs or hosts from public view if needed.

The project administrator can preview any description, banner or other settings prior to publishing the archive. Project settings can be changed and previewed by clicking the Public Access icon on the project home page, by selecting the **Public Access** menu, or from the **Administration** menu.

In accordance with the **Section 108 Study Group recommendations** concerning web archiving, an embargo period will be observed before captures can be made publicly available. A project can be published as soon as the oldest capture is six months old. Any captures in your project more than six months old will automatically be indexed and made searchable together, so that you can preview what search results will return for end-users searching the published archive. Once a project has been published, captures will automatically be added to the archive as they emerge from the embargo period. This gives curators 6 months to conduct quality assurance on any captures run in a public archive before they become publicly accessible.

Project administrators have the option to leave a project unpublished and to maintain it as a dark archive. There are also features available to the curator if there is content included in the project that would not be appropriate for public display; this content can be suppressed from public view. A published project can be unpublished if necessary, but this is not recommended; it would cause any links to archived content to stop working.

Steps to Publish a Project

There are three required steps a project administrator must take to create a public archive:

1. Create a project description to provide end-users with information about the archive.
2. Choose a directory name for the public archive. This will create the URL for the public archive home page. Example: <http://webarchives.cdlib.org/lagovdocs>
3. Click the **Publish changes** button.



There are also a number of optional steps the project administrator can take to tailor the appearance of the public archive:

- Upload a tailored banner image to display on the archive home page, and a smaller icon image that will display at the top of each rendered page.
- Add links to other important resources that are topically related to the archive.
- Choose which project administrator will get email via the **contact** form, if there is more than one administrator for the project.

Changes you make to project information will not be reflected in the public view until you click the **Publish changes** button.

Project Configuration Tips

Project Overview Tips

Changes you make to project information will not be reflected in the public view until you click the **Publish changes** button.

We strongly recommend that you include the words **web archive** in the project name. This name is the end-user's strongest cue that they are interacting with an archive.

The overview tab will tell you the visibility status of the project.

If a project has been published, there will be a **disable public visibility** button. We strongly recommend that you not remove archives from public view, as any links to archived documents will stop working.

You must choose a URL to publish a project. Consider the directory name carefully, as it will be part of the path to these archived materials. This cannot be changed once it is set.

The number of public sites listed includes only those captures older than the six-month embargo period. It may not reflect all of the captures you have run.

Project Description Tips

The information you include here will provide context for users who find your web archive pages. Use the description to tell users the scope and purpose of the archive.

The text you provide here will appear on the web archive home page (in part), on the "About" page for the archive (in full), and on the CDL gateway to web archives (in part).

Click on the text to edit.

Changes you make to project information will not be reflected in the public view until you click the **Publish changes** button.

Related Links Tips

The links you create here will appear on the web archive **About** page.

You can create links to other web archives, your library or unit home page, your digital collections or any other relevant resource.

Changes you make to project information will not be reflected in the public view until you click the **Publish changes** button.

Project image tips

You can use images to customize each of your web archives.

The **banner image** will appear on archive home page, the site list, search pages and help screens.

The **icon image** will appear in the heading for displayed content. It will also be used on the CDL page that lists all WAS web archives.

The name of your project does not need to appear in the image; the project name will appear prominently in the public interface.

Changes you make to project information will not be reflected in the public view until you click the **Publish changes** button.

Contact Info Tips

End-users will have a **contact us** form to submit questions and comments about the public archive.

If there is only one project administrator, end-user questions will be sent to the project administrator.

If there is more than one project administrator, you can choose who will get end-user questions.

All email will automatically be cc'd to washelp@ucop.edu. WAS technical support will respond to any technical questions submitted by end-users.

You can edit your email address if needed by selecting the **Administration > Your account** menu.

Changes you make to project information will not be reflected in the public view until you click the **Publish changes** button.

Content Suppression Tips

There may be cases where you need to suppress content from public view without deleting an entire capture from your archive.

You can enter specific URLs, directories, or host names to suppress content from public view.

Researchers accessing your archive may find records for the content when they search, but will receive a message that it is not viewable.

Curators will still be able to see the content when they are logged in to this project in the curator interface.

Changes you make to project information will not be reflected in the public view until you click the **Publish changes** button.

Rights Management Tools

The Web Archiving Service provides the following rights management features:

- An embargo period of six months is automatically applied to all captured content before it can be made publicly accessible.
- Content owners have the right to "opt out" of the archive. While the Section 108 Study Group recommendations state that U.S. government agencies should not be able to opt out of web archives, these are still only recommendations. CDL will respect any content owner's request to opt out of the archive, but will strongly encourage them to allow their materials to be archived.
- The project administrator has the ability to suppress specific URLs, directories or hosts from public view. This option will suppress content from view only within a specific project; it will not impact content in other archives. This option may be used if a content owner has contacted a curator to opt out of the archive. If this occurs, the curator should communicate

this to the WAS support staff as well. This option may also be used if there is information in the archive that violates privacy issues, or objectionable material has been captured in the archive.

- The WAS administrators at CDL have the ability to stop any capture in progress if it is harming or impacting the content owner's server. You will be notified if a capture in progress has been stopped for this reason. WAS support staff will attempt to alter crawler settings if possible, or will inform you if the site cannot be captured.
- The WAS administrators have the ability to block certain hosts from being captured. Databases licensed to the University of California are blocked from capture, in accordance with UC's licensing agreements.
- The WAS administrators at CDL have the ability to suppress content from public view in all archives. This will be done in response to a content owner's request to opt out of the archive. The WAS support staff will notify any curators who are using the URLs in question as seed URLs for captures.
- Curators should contact washelp@ucop.edu if a robots.txt file is preventing the effective capture of a site, including for government agencies. While Section 108 Study Group Recommendations state that government agency sites should not prohibit libraries from capturing web content, we are still not able to override robots.txt files without permission from the content owner. The WAS support staff can provide you with a form letter and guidance for requesting that content owners update their robots.txt files to allow access to our crawlers (preferred), or allow CDL to override.
- The Web Archiving Service will automatically leave CDL contact information on each content owner's server. Content owners will be able to follow a link that describes the service, the value of the archives, and details for contacting the WAS support staff.

Administration: Your Account

Every user can review and edit basic information about his/her account. From the **Administration** menu, select **Your Account** to keep your email and other contact information up to date.

You can also set your default time zone. This setting determines how capture times are reported. When you change this setting, the start and end times will be adjusted to reflect your preferred time zone. You will need to log out and log back in for this change to take effect.

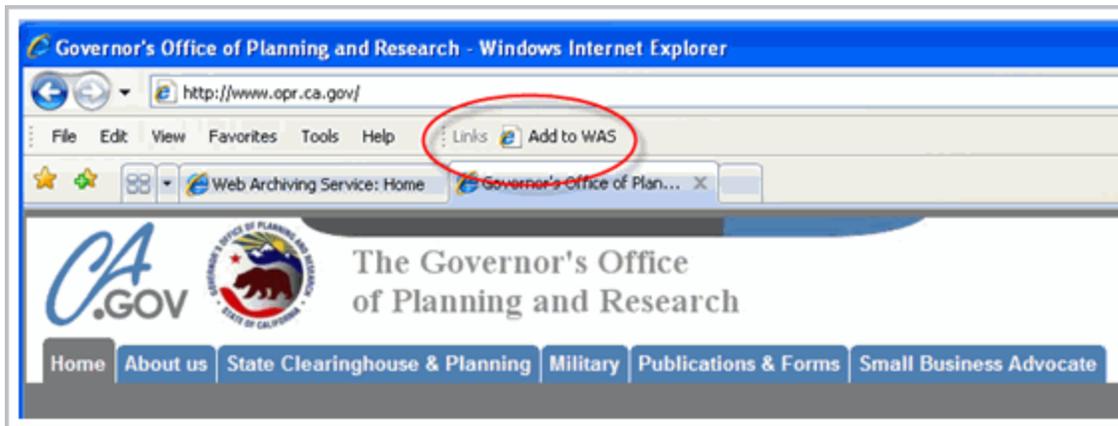
The **List Projects** screen lists all projects you are a member of and your permissions level for each project. Only Project Administrators can add or remove users from projects. If you need to be removed from a project, Project administrator contact information is provided on this screen.

The **Change Preferences** page allows you to control email notifications. If you are a project administrator, or if you are the person who created a site entry, then by default you will get an email whenever a capture of that site is started. This screen lets you opt out of email notifications for scheduled or manually started captures.

The WAS Browser Button

The WAS Browser Button is a tool that allows you to quickly add a site or URL to the Web Archiving Service to be captured later. This type of button uses your browser's bookmark or favorites toolbar to create a browser "button" action (known in online terminology as a bookmarklet). This enables you to quickly add sites to WAS while you browse.

When you install the browser button, you will place it on your browser's toolbar for easy access. Any time you wish to add a page to WAS, simply click the **Add to WAS** button while you're browsing the Internet.



Once you've clicked on **Add to WAS**, you'll be connected to the Web Archiving Service. You may be prompted to log in. WAS will confirm the site information, using the default page title and URL. You will have the opportunity to edit this later if the page title is not accurate. You also have the option of adding the selected URL to an existing site entry if you've found an important subdomain URL for a site.

After you **Add** the site, you have the option to add descriptive metadata, or you can continue browsing and adding more sites.

How to Install the WAS Browser Button in common browsers:

- Internet Explorer 8
- Internet Explorer 7
- Internet Explorer 6
- Firefox 2.x and 3.x
- Google Chrome
- Safari 4.x
- Other Browsers

Installing for Internet Explorer 8

1. Be sure the Favorites Bar is enabled. Select **Tools, Toolbars** and be sure the **Favorites Bar** is checked.

2. Right-click on this **Add to WAS** link and choose **Add to Favorites**.
3. If you receive a warning that the link may not be safe to use, choose **Yes** to continue. This warning occurs for any link that contains Javascript (which is necessary for the browser button to function correctly).
4. You will see the **Add a Favorite** dialog box. Next to **Create in**, choose the **Favorites** folder and choose **Add**.

Installing for Internet Explorer 7

1. *Be sure the links toolbar is enabled.* Select **Tools, Toolbars** and be sure the **Links** toolbar is checked.
2. Right-click on this **Add to WAS** link and choose **Add to Favorites**.
3. If you receive a warning that the link may not be safe to use, choose **Yes** to continue. This warning occurs for any link that contains Javascript (which is necessary for the browser button to function correctly).
4. You will see the **Add a Favorite** dialog box. Next to **Create in**, choose the **Links** folder and choose **Add**.

Installing for Internet Explorer 6

1. *Be sure the links toolbar is enabled.* Select **View, Toolbars** and be sure the **Links** toolbar is checked.
2. Right-click on this **Add to WAS** link and choose **Add to Favorites**.
3. If you receive a warning that the link may not be safe to use, click **Yes** to continue. This warning occurs for any link that contains Javascript (which is necessary for the browser button to function correctly).
4. You will see the **Add a Favorite** dialog box. Next to **Create in**, choose the **Links** folder.

Installing for Firefox 2.x and 3.x

1. *Be sure the bookmarks toolbar is enabled.* Select **View**, go to **Toolbars** and make sure the **Bookmarks** toolbar is checked
2. Drag the **Add to WAS** link in this step and drop it on your bookmarks toolbar at the top of your browser.

Installing for Google Chrome

1. *Be sure the bookmarks toolbar is enabled.* Click the wrench icon in the top right corner of the window and make sure **Always show bookmarks bar** is checked.
2. Drag the **Add to WAS** link in this step and drop it on your bookmarks toolbar at the top of your browser.

Installing for Safari

1. *Be sure the Bookmark Bar is enabled.* Select **View** and choose the option for **Show Bookmark Bar** if that option is available.
2. Drag the **Add to WAS** link in this step and drop it on your Bookmark Bar at the top of your browser.

Installing for Other Browsers

The WAS Browser Button is likely to work in most browsers that have bookmark or favorites functionality and that support Javascript.

To install the browser button, simply add this **Add to WAS** link to your browser bookmarks or favorites in the way indicated by your browser's maker. Many browsers have an easily accessible toolbar for the bookmarks you specify. For quickest access you may wish to place it on this toolbar, but it will work the same way whether it's in your regular bookmarks list or on a bookmarks toolbar.

When you wish to use the browser button, simply click it from your bookmarks or favorites area for any page on which you'd like to us it.