

# Web Archiving Service Collection Guidelines

February 2009

Web Archiving Service Collection Guidelines .....	1
Overview .....	1
Terminology .....	2
The Collection Plan .....	2
1. General Project Information .....	2
2. Project Scope and Purpose.....	3
3. Site List: Evaluating sites .....	3
4. Descriptive Metadata.....	5
Rights .....	6
In Brief.....	6

## Overview

A Web Archiving Service account allows an institution to create **Projects** for developing different web archives. Each project can be seen as a 'holding bin' for captured web sites that are related topically. Each project is shaped by at least one project administrator, and is guided by a Web Archiving Service Collection Plan, which will be written prior to the creation of the project.

The purpose of a Web Archiving Service Collection Plan is two-fold. The first is to help curators to define the scope and purpose of the planned web archive. This will help each institution to plan and prioritize web archiving activity. The second is to alert the California Digital Library to forthcoming web archiving activity. This will help to insure that CDL can support web archiving activity across all institutions. The collection plan is guided by the features available in the Web Archiving Service and the decisions that curators will need to make when they use this service.

The web is a quickly changing and dynamic source of content. While the collection plan you create should specify the sites intended for capture as much as possible, it is not expected that the collection you build will be strictly limited to the sites you have specified in advance. However it is advisable to review and update your collection plan on a yearly basis where appropriate.

The following document will outline what each WAS collection plan should cover, and the steps that curators should take in evaluating the content that will be collected. When finished, the collection plan will be reviewed by the WAS administrator for your institution and sent to [washelp@ucop.edu](mailto:washelp@ucop.edu) if the project is approved. The information requested for each plan is shaped by the capture options available in the Web Archiving Service.

If there is an important collection constraint or feature that is not mentioned in this document, you are encouraged to mention it in your plan, so that the CDL web archiving team can learn about issues that are critical to your evaluation process.

Note that some of the issues to consider here focus on forthcoming features in the Web Archiving Service, such as public access. Though these features will not be immediately enabled, they are covered here because they will shape your collection development strategies.

## ***Terminology***

There are some important terms to clarify before you begin the collection planning process.

**Site:** A site is whatever you define as the target for the crawlers to capture. It may correspond to an entire web site or it may only specify part of a web site. WAS provides tools to help you describe and organize your sites.

**Capture:** A capture is a single archived instance of a web site's content. A single site can have many captures, all run on different dates.

Curator:

**Project:** A Web Archiving Service project is a 'holding bin' for topically related sites and captures. This allows curators to evaluate the success and relevance of the results before deciding if the content belongs in a collection. Different curators or teams of curators have permission to work with different projects. A project is also distinguished by the metadata that will be used to describe the content. Forthcoming WAS features will allow project administrators to define what metadata fields can be used to describe sites in each project.

**Web Archive:** A publicly accessible project. Anyone may search or browse the web archive you've created, and you may continue to add new captured sites on an ongoing basis.

## ***The Collection Plan***

A WAS collection plan form has been provided to help curators envision and describe their web archiving projects in advance: [http://was.cdlib.org/docs/was\\_collection\\_plan\\_form.doc](http://was.cdlib.org/docs/was_collection_plan_form.doc). The following information is requested in the plan:

### **1. General Project Information**

#### **PROJECT NAME**

This will be the name of the web archive as displayed to the public. Be sure to choose a name that is descriptive of the content. "My Library Web Archive" might not be sufficient for users who aren't already familiar with your library's collection.

#### **INSTITUTION**

The organization that has the Web Archiving Service account. Example: UC Berkeley Libraries.

#### **CURATOR(S)**

Who is responsible for this collection? Specify both the project administrator(s) and the submitters. A project administrator is responsible for shaping the scope of the collection, inviting other curatorial users and controlling settings for public access. A submitter is able to create site entries, capture sites, search and display content and build collections from captured content. Do you anticipate collaborating with curators outside of your library (either faculty members or others who you plan to invite to contribute to the project)?

## 2. Project Scope and Purpose

### PROJECT CONTENT

What subject area or theme unites the web sites in the web collection? What makes the content under consideration a good candidate for web archiving? (consider both the value of the content and the relative risk or stability of the content).

### PROJECT SCOPE

Will material be collected for the project on an ongoing basis? Is it an event-based project with a presumed end-date for collecting?

### PROJECT PURPOSE

What research areas and programs will the collection support? How might researchers use the resulting web archive?

### PUBLIC ACCESS

Do you plan to make the project publicly accessible? Do you plan to create catalog records for some of the documents you capture?

### USER GROUPS

Who are the user groups for the web collection? In many cases there will be more than one user group that will use a collection, for example faculty, students, and the general public. How might the user groups impact the decisions you make when building the collection? Consider that while your project may be targeted to a particular user group, such as University of California, Berkeley researchers, anyone will be able to access the public web archive you build.

## 3. Site List: Evaluating sites

The most important part of the collection plan will be the site list with the names and URLs of the sites you plan to capture.

Prior to creating a site list for your collection plan, we recommend taking the following steps to evaluate the sites you plan to capture. These criteria will help to shape the capture settings and descriptive data for each site.

### SITE NAME

While a web site's home page may have a title tag, this is not consistently used or accurate. You will be asked to supply your own name for each web site, which is often based on the most prominent heading on the site's home page or on your own knowledge of the site. Please be as specific as possible and consider that ultimately end-users may find this archived content in some other context. For example, "Mayor's Office" is not as ideal a site name as "Mayor of the City of San Francisco".

Note that other factors can influence what you choose as the site name. In the previous example, if you intend to always have the S.F. Mayor's office collected under the same title, even as administrations change, "Mayor of the City of San Francisco" is appropriate. If you intend to provide access to the site based on the person occupying that office, "Gavin Newsom, Mayor of San Francisco" would be a more appropriate site name. You would then create a new site entry when the administration changes.

Assume that site names will eventually become a means by which end-users can explore your web archive collections when public access is available.

## **SEED URL(S) FOR THE SITE**

In order to capture a site or part of a site, you will need to tell the crawler what page to start on. This is usually straightforward, but can sometimes be complicated, depending on how the site is designed. This decision is closely tied to deciding the capture scope settings for the site.

If you're capturing the entire site, the seed URL is usually simply the home page URL, or just the domain section of the home page

Examples:

California Department of Forestry and Fire Protection:  
**<http://www.calfire.ca.gov/>**

Cal-Fed Bay Delta Program  
**<http://calwater.ca.gov/index.aspx>**  
In this case, the URL that shows when you load the home page is  
<http://calwater.ca.gov/index.aspx>. All you need to enter is  
<http://calwater.ca.gov/>

In some cases, the site you wish to capture may be a directory within a larger site. City government offices (such as a mayor's office) are often delivered as part of the city's web-presence as a whole (such as the city of Los Angeles). It is up to you to decide whether you want to capture the entire city site or if you want to capture parts of it separately. For directories, you need to make sure to specify the specific directory for the seed URL.

**[http://www.nasa.gov/mission\\_pages/GLAST/](http://www.nasa.gov/mission_pages/GLAST/)**

If you don't want the entire website, you can also specify a single page or list of pages. Be sure to select "page" for your scope setting if you choose to do this.

Some web sites just are not designed optimally for web capture, particularly if they are dynamically generated or rely heavily on content management systems. The San Francisco Mayor's Office is one such example; it is quite difficult to identify a single URL or directory that will fully capture the Mayor's office and its departments without also capturing the rest of the city. If you encounter a site that is structurally difficult to understand, you can always contact the WAS-SUPPORT mailing list for help in identifying a capture strategy.

## **SITE SCOPE**

You don't need to know the scope settings in advance, but it may be helpful to consider them in advance as you assess the sites in your collection plan.

As you explore the site, determine how much of the site you wish to capture and archive. There are two settings that will influence this; the first setting defines the scope of the site you plan to capture:

- Page
- Directory

- Host site

The second defines whether *related* information will be captured – that is, whether the crawler will capture linked pages from hosts that do not appear in your seed list. This “capture linked pages” (or “+1”) setting can be any of the three scope settings above. Use this setting when you are capturing a site, directory or page that is particularly good at linking out to content from a number of sources on the web. The crawler will not capture the entirety of those other sites, just the pages that are relevant to your site and the embedded content necessary to fully render those pages.

Example: one strong value of blog pages is that they often link out different resources on the web. During major events, most of the relevant discussion occurs on the main page of the blog. In this case you would not want the entire blog, and you would want the relevant content from other sources, so “page + 1” would be a good scope option.

It can sometimes be difficult to tell just how much external material a site is linking to until after you capture it. You can always adjust your settings after analyzing capture results, try again, and delete any older captures that did not work to your satisfaction.

### **FREQUENCY**

WAS provides the option to schedule captures to run automatically. Volatile sites of critical importance can be run on a monthly, weekly or daily basis. As you evaluate sites, consider how frequently they are likely to change.

Note duplicate reduction is still a forthcoming feature; WAS does not yet have an option to capture \*only new or changed\* content. This means that you want to be careful with setting up capture frequency. In some cases you may decide to capture a site once in its entirety, then set up scheduled captures with a narrower scope to capture changes close to the site's home page.

Evaluating the capture frequency will also give you and CDL a sense of how large your planned collection will be.

### **MAX TIME**

This is not included on the collection plan form, but it will be important to consider when you create event-based projects or when you plan to schedule a capture frequency. The Web Archiving Service allows brief (1 hour) or full (up to 36 hour) captures of a site. Because we capture sites individually, the 36 hour option is usually sufficient; only 2.8% of our full captures hit a time limit before fully capturing the site.

During events, or when capturing the same site frequently, meaningful changes in content tend to occur on or close to the home page. If you begin with a full capture, then follow up with brief captures, you can get important changes to content without getting a great deal of duplication.

## **4. Descriptive Metadata**

The Web Archiving Service currently provides four default Dublin Core fields for describing sites:

- dc.description
- dc.subject
- dc.coverage.geographic
- dc.creator

While it is not yet possible to tailor the metadata fields for a project, it will be extremely helpful for CDL to know your descriptive metadata needs, and will be a useful tool to help you distinguish when you need different projects for different content.

As you plan each project, determine what metadata you will need to describe the sites you capture. Consider that this metadata will become a means of organizing long lists of sites, and will ultimately a means of browsing sites when your web archive becomes publicly available. Example: planned site metadata: California Wildfires

- Fire name:
- City:
- County:
- Site type: [state agency, city, news, blog]

## ***Rights***

The Web Archiving Service will follow the Section 108 Study Group Recommendations for the Preservation of Publicly Available Online Content, released in March of 2008. Those recommendations are covered in detail in the “Web Archiving Service Rights Management Practices” document: [http://was.cdlib.org/docs/was\\_rights\\_management.pdf](http://was.cdlib.org/docs/was_rights_management.pdf)

In brief, you do not have to request advance permission to capture either government or commercial websites, but commercial websites do have the right to block Web Archiving Service crawlers, or to request that their materials be suppressed from an archive.

For collection planning purposes, it is helpful to understand the role of a “robots.txt” file that content owners can use to prevent the capture of their sites. These are files that contain instructions for web crawlers, and the Web Archiving Service obeys them by default. If a federal, state or local government agency, a political party or candidate or a political action committee is using a robots.txt file to prevent capture, you can contact [washelp@ucop.edu](mailto:washelp@ucop.edu) to evaluate how to capture it. When evaluating sites, you can tell if the capture will be impacted by a content owner’s instructions by entering *just* the host section of the URL followed by /robots.txt

Example:

<http://www.huffingtonpost.com/robots.txt>

If you do not get a result, then there are no rules prohibiting capture. If you do get a result, rules for interpreting robots.txt files can be found at: <http://www.robotstxt.org/>

## ***In Brief***

Each WAS Collection Plan should contain at least the following information:

- Project Scope and Purpose - with an indication of the collection subject area, priority for your institution and the curator names.
- Site List with site names and URLs.

It will also be helpful for CDL to see what metadata fields you would like to use for your project.