Capturing Sites: 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: Build a collection**

Once you complete one or more captures and confirm that you've captured the materials you want for your collection, use the Collections menu to create a new collection and add content to that collection. You can either add entire captures, or select as little or much of the captured content as you want. You can add individual files as you search or browse capture results.

**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).

**Useful Resource**

The Alexa site (http://www.alexa.com) can be an excellent source of background information about a web site. It can tell you how busy a site is, and provide any subdomains associated with a site.

Type the address of the site can click Go. The Traffic Details report will contain a list of all the site's subdomains at the bottom of the page.

**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://". 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 documents 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.

Capture Settings

Site Name
What you name a site is up to you. A site names 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.

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://. 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 Captures
You can now set sites up to be captured on a daily, weekly or monthly basis.

- Daily: you will need to 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.
- Weekly: weekly captures will be queued to run at midnight on Friday to allow longer captures to run through the weekend.
- Monthly: monthly captures will be queued to run on the 15th of the month.

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

- 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.
- You 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, Subjects, Coverage (geographic).

**Rights Data**
For WAS-6 (August 2008) this screen will allow you to add rights information pertaining to a site. This will include entry fields for any copyright statements you found on the site, and contact information. It will also include a rights designation to help you keep track of rights management decisions and activity.

**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 documents until it gets 25 links away from the starting point. This should provide a thorough capture of most sites.

**Politeness settings**

During pilot testing, the WAS crawlers will be 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**
Until rights management features are in place, WAS will honor robots exclusion files prohibiting crawlers from capturing web sites.

**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 to issue the job, 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.

**The Site List**

From left to right, each row in the site list provides a checkbox for starting groups of captures, the site name and description, the current status, the number of captures that have been started for that site, and the actions that can be taken for that site.

- Check the checkbox to select sites to be captured in a group. The "Capture checked sites" button is at the bottom of the list.
- Click on the number of captures to see details for each capture history.
- 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.

**Navigating the List**
You are likely to accumulate hundreds of sites in your site list. The Manage Sites screen offers some tools to help you navigate that list. You can limit the list to sites based on when they were last captured. You can page through your sites, or jump to a specific page in your site list. You can also look at the list of sites you have deactivated; that list will allow you to reactivate sites if needed.

Starting and Stopping Captures

Starting

Running batch captures

For pilot test 5, you can check multiple sites for capture, but they do need to be listed on the same page. You can control the number of sites listed on the page to increase the number of sites in the batch. The sites you've selected will be queued for capture at once, but they will still be crawled separately, just as if you started them one at a time.

After capture starts

You will get a confirmation screen showing capture status and your capture will be queued by the crawler. The amount of time it takes to begin depends on the number of other captures (for all WAS users) waiting in the queue. 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 "Cancelled".

View Captures

The View Captures screen is designed to help you find the content you need even after you've run hundreds of captures. This 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 cancelled. When there are two or more completed captures for a site, you will also see a compare link.

Click on the name of the site whose content you're interested in seeing. A list of captures associated with that site will open up beneath the site name. Basic information about each capture is provided to help you select the correct one from the list. This includes the date and time the capture started, who started it, settings selected, status, number of files captured, and the capture
duration.

For each capture, you can View Content or click the delete icon. The DELETE ICON will remove content from the archive. We strongly recommend that you only delete captures that were unsuccessful. Sometimes you don't know which captures are going to contain documents that later disappear from the web, so keeping it in the archive is a safeguard.

When any content from a capture has been added to a collection (either the entire capture or individual files) the used icon appears. Clicking on this icon will tell you which collections contain files from this capture. In order to delete this capture, you will have to go to the Manage Collections screen and remove the content from the collection first.

When you click on View Content 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: Guide**

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. If the crawler reached a limit set by you, the capture status will be highlighted in red. You will also see if this entire capture is part of a collection, and you have the option of adding this entire capture to a collection.

**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
The following search tips apply both to searching individual captures and to searching collections.

Keyword
Keyword is the default search option. This 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*.

**Limit by File Type**

The search screen now includes a limit option for common file types. This will allow you to find a particular keyword only in PDF files or other specific file types. Note that your capture or collection likely has many more file types than you see here. See below for information about searching for file types not included in this list.

**Search or limit by URL**

Include **url:** in front of your search term to find that term only in the URL text. Use this when you wish to narrow a collection search to a particular site. Do not include a space after the colon. **Example:** *url:calwater drought* finds the term *drought* on any page that includes *calwater* in the URL.

**Search by file type**

Include **type:** in front of your search term to find only files of a specific type. Do not include a space after the colon. **Example:** *type:pdf*

Some common type searches:

- type:html
- type:pdf
- type:msword
- type:excel
- type:powerpoint
- type:image
- type:audio
- type:video
- type:application
  (will find PDF, Word, Excel and other complex files.)

**Important:** the file types are automatically derived from the record’s HTTP headers, and are not always accurate. PDF files will often appear to be MSWord files, since they were often originally created with Word.

**Search by title**

This search is not currently recommended, particularly for finding PDF files. When you include **title:** in front of your search term, the search engine will
look in the document's META tags for title information. These tags are rarely used for PDF and Word files, and are often inaccurate for HTML files.

**Other Notes**

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 cannot combine two field-specific searches. The search `type:pdf url:budget` will not produce useful results.

**Browse Files**

The following tips apply to browsing file lists for both individual captures and collections.

The **Files** list is an inventory of every file in the current capture or collection, including individual image files, javascript files, html files and PDFs. You can limit this list to particular filetypes in order to get (for example) a quick list of every PDF file found for a capture. Note that the **types** drop-down only shows a few commonly sought file types; there may be many more file types in your capture or collection. See the **Search Tips** guide for instructions on searching for any file type.

If you are browsing files for a capture, you can use this page to add batches of files to a collection. Select the check box next to the URL you wish to add, then choose a collection at the bottom of the screen. You will get a confirmation when the files are added, or a warning if some files were already in the collection.

**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 the first time you display a file from a capture; the capture results are being retrieved from the repository and are being uncompressed for you to view. Subsequent files you view from the collection will load much more quickly.

**Top Navigation Bar**
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 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 have the ability to add comments about this particular page, and the option to add this page to a collection.

**Geographic Results Mapping**
The **map** tab on the Capture Results screen shows the geographic locations of all host servers that provided content while capturing this site. This will include all related servers for captures that were run with a **host + linked pages** scope.

Click on a pointer to see the list of hosts encountered at that point on the map.

- 👑 Fewer documents hosted
- 🌟 More documents hosted

Use the + plus icon to zoom in to a location on the map. Click and drag inside the map to scroll.

This map is generated from the "hosts report" generated by the Heritrix crawler, a database of geographic locations for IP addresses, and Google Maps.

### 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.
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

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, you can still add them to a WAS collection.
- **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.

**Add files to Collection**

You can add files to a collection as you analyze the change results on this screen. Use the check boxes next to each file and the drop-down collection list at the bottom of the screen. You do have to add checked files to a collection before paging ahead in the file list.

**Build Collections**

**Collection**

A group of resources related by common ownership or a common theme or subject matter. A web collection consists of one or more jobs that capture a group of related websites (e.g., candidate websites for state election campaigns). A WAS collection brings together the content you've captured into a single, searchable web archive. You can add entire captures to collections, or choose specific pages from a capture. A WAS collection can include any number of sites, and can include content from the same site captured on many different dates.

**Create Collection**

Supply a name and a description for the collection. Once you have created the collection, it will be available in the add to collection drop-down menus when you're reviewing search results or browsing file lists.

After you create a collection, the confirmation screen will prompt you to add content to the collection.

**Manage Collections**

The Manage Collections screen lists all of your existing collections. It allows you to view collection contents, edit collection title and description or delete the collection. You can also add content to the collection or remove content from the collection.
Selecting the Add icon will take you to a list of sites with finished captures available for adding to a collection. Click a site name to see the captures available, then either select the checkbox to add the entire capture, or use the search files or list files options to add selected files from that capture.

Selecting the Remove icon will let you remove content from the collection. This will take you to a screen that lists the collection contents, and provides a trash can icon for removing either entire captures or individual files from the collection. At this time there is not yet a way to remove only a single file if what you added was the entire capture; you can only remove files individually when you added them individually.

Selecting the trash icon will remove data from the collection, but will not delete it from the archive. If you delete an entire collection, you will not impact or remove the archived data. The files and capture jobs will all remain in the archive.

View, Display and Navigate Collections
View Collection Contents

To review and display a collection's contents, go to the Manage Collections screen and click on the View icon.

This will take you to the Collection screen, which provides an overview of how many files are in the collection, and a table of contents of included sites. This table of contents expands when you click on it to show the details of each site. Over the course of time, your collections are likely to include a very long list of sites so this screen is intended to simplify a potentially complex list.

The contents list is updated immediately whenever you add or remove data from a collection.

Search & Browse Collection Contents

The Search and Files tabs for collections work much like they do for captures. The Search feature will find your search terms only in files included in the collection. See the Search Tips guide for a detailed searching guide. The Files screen looks slightly different than for captures because it no longer includes collection-building features (since you are already in a collection). The Search and Files screens take about an hour to update after you have added or removed content from your collection, so you may see a mismatch between the contents list and your search results.

Display Collection Contents

When you've searched or browsed and then select a file to display, the page will display much like it does for captures. You may encounter a delay the first time you attempt to display content, while the associated files are retrieved from the repository and uncompressed. You will see a top information bar, with the rendered page underneath. The top bar will indicate that you are in a collection. Because you can add different captures of the same site over time, the page history on top bar will indicate
when there are multiple versions of the same item in the collection. Use the drop-down to navigate to the other dates for the same file.

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 View 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 WAS Bookmarklet

The WAS Bookmarklet is a tool that allows you to quickly add a site or URL to the Web Archiving Service to be captured later. Now you can add sites to WAS while you browse.

When you install the bookmarklet, you will place it on your links toolbar for easy access. Any time you wish to add a page to WAS, simply click Add to WAS while you’re browsing.
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 Bookmarklet**

**Installing for Firefox**

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 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 bookmarklet to function correctly).
4. You will see the Add a Favorite dialog box. Next to Create in, choose the Links folder.

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 bookmarklet to function correctly).
4. You will see the Add a Favorite dialog box. Next to Create in, choose the Links folder.

Installing for Other Browsers

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

To install the bookmarklet, 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 whether it's in your regular bookmarks list or on a bookmarks toolbar.

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

RSS Feeds for Monitoring Capture Status

WAS now makes RSS feeds available for monitoring the status of your WAS captures. This allows you to check whether your captures have completed successfully without logging into WAS.

The RSS feeds list the names, statuses, and start times of your 10 most recent WAS captures. Click the link below to view your feed:

Capture Status

There are also links on the Manage Sites and View Captures pages.

You can set up an RSS reader to automatically check your feed. There are many free
readers available, including the most recent versions of the major Web browsers.

The Wikipedia has more information on RSS and feed readers.

Setting Up RSS Monitoring in Internet Explorer 7

1. Click the **Capture Status** link. Internet Explorer formats and displays the RSS feed in the browser window.
2. Click the **Subscribe to this feed** link. Subscribing allows you to monitor your feed from in Internet Explorer.
3. The Subscribe popup window appears. Select **Feeds** from the Create In menu.
4. Click **Subscribe**. Internet Explorer creates the subscription.
5. Click **Tools**, **Toolbars**, and then **Feeds** to view your feed subscriptions.
6. Click the WAS feed to open it in the browser window.

![The Feeds toolbar and a formatted feed in Internet Explorer 7.](image)

Setting Up RSS Monitoring in Firefox 2

1. Click the **Capture Status** link. Firefox formats and displays the RSS feed in the browser window.
2. Select **Live Bookmarks** from the menu at the top of the screen. Live Bookmarks allows you to monitor your feed in Firefox.
3. Click the **Subscribe Now** button.
4. The Add Live Bookmark popup window appears. Select **Bookmarks Toolbar Folder** from the Create In menu. Click **OK**.
5. Make sure the Bookmarks Toolbar is visible in your Firefox browser. You can click **View**, **Toolbars**, and then **Bookmarks Toolbar** to view it.
6. A Live Bookmark for your captures appears in the Bookmarks Toolbar. Click it to view the status of your captures.

A formatted RSS feed and open Live Bookmark in Firefox 2.