

Curators' Evaluation of WAS Release 2/3

March 16, 2007

Prepared by:

Kathleen R. Murray
krmurray@unt.edu

University of North Texas

Contents

1 Introduction 2

2 Site Definition, Capture, and Management 3

3 Reports and Analysis 5

4 Search & Display 7

5 Overall Reactions 10

1 Introduction

The Web Archiving Service, WAS Release 2/3, was available to the project's curators from January 30, 2007 – February 13, 2007. During this time 19 curators trialed the service with varying frequencies, as measured by the number of capture jobs they submitted. Subsequent to their trials, 17 curators submitted a web-based survey evaluation of their experience. The survey consisted of 27 questions, divided into four sections: Site Definition, Capture, and Management; Reports and Analysis; Search and Display; and Overall Release.

Overall, curators thought the workflow reflected in the redesigned interface was a great improvement from the first release and they noted that the service was intuitive and easy to use. This release effectively addressed the three key functions that posed the most confusion in WAS release 1. These were: (1) Starting a capture after defining it; (2) Determining when a capture was completed; and (3) Determining how to view captured results.

While opinions about the new and enhanced features in this release were not unanimous, each major enhancement was welcomed by some. The features many curators received well were searching and browsing, reports, filtering and sorting of files, and longer capture times. Several curators noted that the overall performance of this release was very slow and some were frustrated that many or all of their captures were not successful. Additionally, many were unable to display files and some found the browsing interface a bit of a barrier to accessing and viewing a captured site in it entirety. The major functions that were either confusing or problematic in this release were:

1. Estimating max time capture settings for sites

2. Inability to display files and receiving either an error message or a "Not in Archive" message for files listed in the captured files list

3. Unexpected search results and lack of understanding about what content was being searched

4. Discovery and display of discrete files of publications or documents, especially PDF files

5. Discovery and display of the "home page" for seed URLs (versus individual files)

To assess the differential influence of frequency of use on some survey results, the 19 curators were assigned to one of the three groups in Table 1 according to the number of capture jobs they submitted during the trial. Due to the small number of curators, it is not possible to infer any significance or relationship between frequency of use and other measures; group effects can at best suggest the possible influence of frequency of use.

Group	Range of Capture Jobs	#
High	10 – 15	4
Moderate	4 – 9	9
Low	1 - 3	4

Table 1. Curators by Frequency of WAS Release 2/3 Use

The remainder of this report summarizes the feedback received from the curators and provides details to illustrate the areas that were either confusing or problematic to them.

Their recommendations as well some considerations for future development are also included.

2 Site Definition, Capture, and Management

2.1 Site Capture Settings

Determining site capture settings is a significant area of uncertainty for curators. In particular determining the optimal time for a capture to run is problematic for most. Overall, 35% of curators had difficulty determining optimal scope settings for their captures and conversely, 65% of curators reported difficulty with determining the optimal max time settings. As Figure 1 shows, moderate users account for the largest variance, with both low and high users split 50-50 in regards to having difficulty with determining settings.

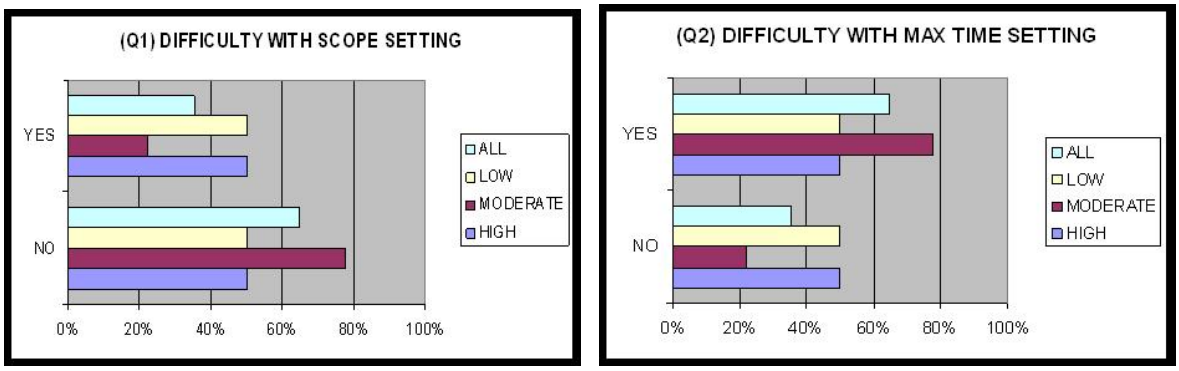

Figure 1. Difficulty with Site Definition Settings by Use Group (*N*=17)

Curators (*n*=12) commented that they were largely unsure how to determine optimal scope settings, especially on the initial capture of a site. Five curators indicated that in the absence of being able to predict optimal settings, they basically used a trial-and-error method but had varying success with this method. Many experimented with time settings, some setting the initial limits high to ensure their content was captured and others starting with a minimal time setting of 10 minutes and then increasing it. A few found that no time setting successfully captured the intended sites or content.

Comments from Curators:

- I had no idea how many files were linked or how much time it would take to do a capture but in trying a 10 minute crawl it helped to get some idea.
- It's simply hard to tell how extensive the site is and how long it will take the crawler to capture it.
- It is too hard to determine the best combination of scope and time for 'one' best crawl at the beginning.

Some commented that the definitions of scope settings were either unclear to them or did not coincide with the actual capture results. A few found the term "seed files" unfamiliar.

Comments from Curators:

- The 'directory only' didn't work (though the documentation warned about that) and seemed to just go to everything, only stopping when it ran out of time. However, the 'host site + linked pages' appeared to do the same thing.

- For the 'Host site only' scope setting, I was unclear how many links away from the seed urls the crawler would capture. Is there any limit?

2.2 Capture Status

Indications of capture status were well-received by curators. Over half (53%) of the curators (*N*=17) were both very satisfied with the in-progress status indication of their capture jobs and definitely found email notification of completed captures beneficial. An additional 35% were somewhat satisfied with the in-progress status indication and an additional 24% found the email notification somewhat beneficial.

Some curators encountered problems in this area. A few curators never received email notifications of completed capture jobs. One commented that the email messages were inaccurate, reporting 'successful' completion for jobs that had been terminated. Another encountered an inaccurate in-progress status display of 'running' although the job had stopped.

Problems Reported by Curators:

- I received an email each time that my captures were 'successful' when in fact all were terminated because the maximum time was reached.
- My City of Irvine search displayed 'running' when it had encounter an error and had stopped.

2.3 Manage Sites: Filtering and Sorting Sites

Six of the curators, four in the moderate use group and two in the low use group, did not try the filtering and sorting features available through "Manage Sites". Sixty-four percent of the remaining 11 curators reported that these features were very beneficial to them (Table 2). Some curators could anticipate that these features would be of more benefit when the service was generally available and they were running many more capture jobs.

N=11		**Filter Sites**		**Sort Sites**	
		#	**%**	**#**	**%**
	Very Beneficial	7	64%	7	64%
	Somewhat Beneficial	3	27%	4	36%
	Not Beneficial	1	9%	0	0%

Table 2. Benefit of Filtering and Sorting Sites Feature

2.4 Recommendations and Considerations

Recommendations from Curators:

- Create tools for estimating the settings for a capture, especially the time needed
- Improve definitions for the scope settings
- In the documentation, "provide a sample URL with a diagram of the parts that would be crawled for each scope setting."
- "The filtering options and the sorting options provided in the Manage Sites feature would be very helpful in the View Results feature as well."
- Include additional capture settings:
  - Limit to a specific file type
  - Exclude certain file types
  - Exclude certain sub-directories

- "Make the email notification optional. That way I'll know about sites I'm really tracking, but won't have my in box filled with the others."
- Add an option to receive capture completion notices via text message

Considerations:

- Using experiences from the trials, is there some automated tool that can be created to estimate the initial scope and time settings for captures?
- What additional guidance or training can be created to assist curators with estimating settings?

3 Reports and Analysis

3.1 Overview Report

All curators indicated the Overview Report was helpful in evaluating their completed capture jobs, with 53% (*n*=9) rating it as very helpful and 47% (*n*=8) rating it somewhat helpful. While curators in the low use group rated it highest there is no meaningful difference between the ratings of the three groups (Figure 2).

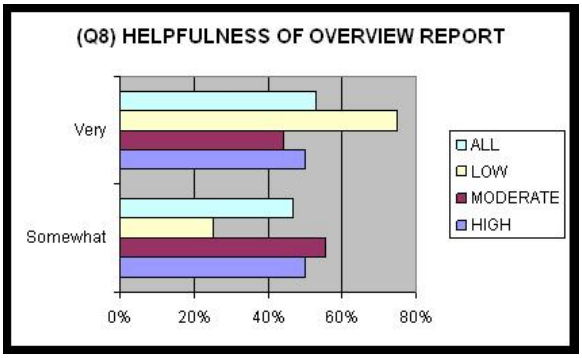

Figure 2. Helpfulness of Overview Report by Use Group

One curator found the indication of "finished" for a capture inaccurate because the capture had actually failed. Another curator encountered a problem with the captured file type data reported on the Overview Report. Specifically, files types listed as captured in the Overview Report (for Goleta, CA) were not in the archive. The curator observed there was a robots.txt file listed for the site but it was not clear to the curator if this accounted for the files not being captured. Even if that were the case, it still would not explain why the Overview Report indicated the file types had been captured. This same curator noted that the text of the report stated that Response Code Help was supposed to be available but could not find a link to this help.

3.2 Other Reports

The majority of curators (76%; *N*=17) did not encounter any problems accessing specific reports about completed capture jobs. One curator did encounter an "intermittent lack of response" when trying to access reports. Another curator downloaded a URI Errors report but it contained no data. The curator was not certain how to interpret this: Were there no URI errors or was the report data missing? Another curator discovered that a "report showed data for the wrong capture in some areas". [*Note*: This curator is likely referring to individual reports accessed from the Overview Report but does not specifically state this.

This conflation of data from two different capture jobs in reports for a single capture also occurred during the qualitative evaluation.]

3.3 Files: Filtering and Sorting

The Files feature provided file filtering and sorting options for the results of capture jobs. Curators were asked to indicate how beneficial they found each option when evaluating the success of a completed capture job. Table 3 lists the file filtering and sorting options in rank order by curators' ratings of "Very Beneficial". Curators rated "filtering captured files by mime-type" as the most beneficial and "sorting files by size" as the least beneficial. In fact, four curators indicated "sorting files by size" was not beneficial in evaluating capture results.

	n	Very Beneficial		Somewhat Beneficial		Not Beneficial		Not Tried
		#	%	#	%	#	%	#
Filter captured files by mime-type	10	8	80%	2	20%	0	0%	6
Sort captured files by URL	10	7	70%	2	20%	1	10%	7
Sort captured files by mime-type	12	8	67%	4	33%	0	0%	5
Filter captured files by seed URL	10	6	60%	4	40%	0	0%	7
Sort captured files by size	11	3	27%	4	36%	4	36%	6

Table 3. Benefit of Filtering and Sorting Files

Not all curators tried all options and a few reported that they were either not sure of the difference between sorting and filtering or that they were not sure how to apply them in their evaluation of capture results. In contrast to this, one curator found the use of system generated metadata to enable sorting and filtering of files to be "GREAT".

Comments from Curators:

- I am still not entirely sure I understand the difference between filtering and sorting?
- Seeing lists of thousands of files can be a bit overwhelming.
- I really really like that I can see all of the URLs by file type; and then be able to sort them by the URL - this lets me see if there are dups; sort things like agendas and minutes in chron order - provided the site used a good directory structure and naming conventions.

3.4 Subsequent Captures

Only about one third (35%) of curators improved the success of subsequent site captures by changing site scope and/or max time settings based on their review of reports about capture jobs (Table 4). Six curators did not attempt to do this. However, of the 10 curators who attempted subsequent site captures, six curators usually improved their success and three sometimes did so. The remaining two curators were unable to improve subsequent captures after reviewing reports.

N=17		#	%
	Usually	6	35%
	Sometimes	3	18%
	Never	2	12%
	Not Tried	6	35%

Table 4. Improvement in Subsequent Captures after Evaluating Reports

Comments from Curators:

- Our one attempt to experiment with improving the scope (switching to 'host site + linked pages' in order to capture some important external links) turned out to be worse that the original attempt.

3.5 Recommendations and Considerations

Recommendations from Curators:

- From the Overview Report add a link to the seed URL(s)
- Set the default for the Files page to list the seed URL(s) at the top
- Add response code definitions to the Response Code Report

Consideration:

- Provide more instruction on how to interpret and apply reports. Perhaps expanding "Help" pages or offering brief, topical, online training modules might be useful.

4 Search & Display

4.1 Searching

Most curators (71%, *N*=17) did not encounter any problems doing keyword searches. A few curators were unclear about what content had been indexed (e.g., only certain file types) and about what exactly was being searched, (e.g., just text in URL's). While one curator was not certain if phrase or proximity searching was possible, a different curator commented that they "liked the phrase searching ability", and a third curator suggested that both Boolean searches and truncated searches seemed possible. One curator only used searching to locate the home page of captured sites and indicated for a second time that they would like to see a direct link to the home page as most of their "users want to mimic the experience of the original [web site], which means starting at the home page". Another curator echoed this in an expression of frustration at not being able to easily find the home pages of captured sites.

Three curators reported problems with the search function. These are related in the comments that follow.

Problems Reported by Curators:

- I was trying to find the publications from the CA Postsecondary education site and searched on 'publication' and 'pub' and didn't find any. I knew that pub appeared in the URL's. Did the keyword search search the URL as well?
- I attempted to search for 'NASA' and 'shuttle' at separate times within my completed 'Return to Flight' archive, but only came up with one hit for each, which seems incorrect.

4.2 Search Results

Eighty-two percent (82%, *N*=17) of curators indicated that the brief item records in search results provided enough information for them to identify specific items of interest. Curators reported the following problems.

Problems Reported by Curators:

- The display of results seems to have a problem displaying diacritics. While the results set may extend to multiple pages, one can only see the first page, that is, no navigation to subsequent pages. The 'return to top' button didn't seem to function.
- It states 'Click on Show Detailed Record to review all item details.' but there isn't a Show Detailed Record link and I couldn't figure out how to display a detailed view.
- A couple of links gave me machine mark-up or was not archived.

4.3 Browsing

Three curators did not attempt any of the file sorting and filtering options and other curators selectively tried certain options. This resulted in ratings of usefulness from 11 curators for each option. Table 5 ranks the browsing options by the degree of their usefulness to curators.

n=11	Very Useful		Somewhat Useful		Not Useful	
	#	%	#	%	#	%
Filter captured files by mime-type	8	73%	2	18%	1	9%
Sort captured files by URL	7	64%	4	36%	0	0%
Sort captured files by mime-type	7	64%	4	36%	0	0%
Filter captured files by seed URL	6	55%	5	45%	0	0%
Sort captured files by size	4	36%	3	18%	4	36%

Table 5. Usefulness of Browsing Options

The results are similar to the results reported in Table 3 regarding the benefit of sorting and filtering files to evaluate capture results. This is not surprising as the function of browsing to identify content logically precedes the function of evaluating the content. Curators found "filtering captured files by mime-type" to be the most useful browsing option and found "sorting captured files by size" to be the least useful, with four curators indicating that the latter option was not useful.

Problem Reported from Curator:

- When I attempted to browse my content for 'Return to Flight,' the filtering URL I was given was for a different capture, the 'Medicaid Commission.' This is similar to the problems I encountered in a few of the reports, one or two of which mistakenly pointed to the medicaid instead of the NASA data.

4.4 Display & Navigation

Overall, 70% of curators (*N*=17) were either very satisfied (*n*=5) or somewhat satisfied (*n*=7) with the display of their captured content (Table 6). One curator did not attempt to display content.

N=17		#	%
	Very Satisfied	5	29%
	Somewhat Satisfied	7	41%
	Not Satisfied	4	24%
	Not Tried	1	6%

Table 6. Satisfaction with Display of Captured Content

Four curators were not satisfied with the display of captured content (Figure 3). This dissatisfaction was seemingly not impacted by frequency of use of the WAS.

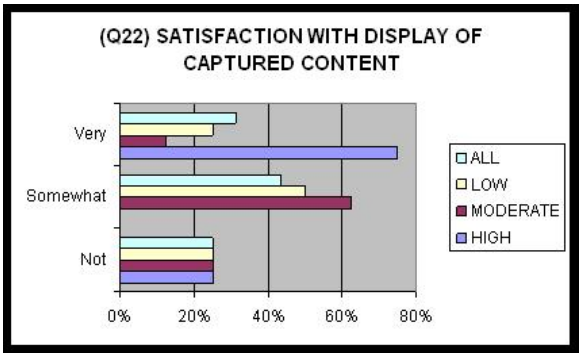

Figure 3. Satisfaction with Display by Use Group (*n*=16)

When viewing their captured content, 10 curators tested hyperlink functionality within captured sites. The same four curators who indicated they were not satisfied with the display of captured content reported that the hyperlinks in captured web sites accessed web content outside of the WAS repository. Another curator commented that they were not certain they wanted hyperlinks to content outside the archive to be blocked.

Many curators commented that they were unable to access any files and encountered either error messages or a message that files were not in the archive. This latter message was quite confusing to several curators who wondered how the files could be listed but the content not be captured in the archive. Curators who indicated they were particularly interested in collecting at the document level were frustrated at their inability to view specific files, especially PDF files. The following curator comment exemplifies the common display issues.

Comment from Curator:

- I'm not sure of the success of the crawls--the results indicate that the files were successfully captured but they're not in the Archive. I couldn't get any to display from the browsing feature. I was able to view some captured files after having done a keyword search, 'goleta.' But then, when I tried to view the document itself (a pdf file off the City Council meeting page, for example), these either took me outside the archive and onto the live web, without notifying me, or it came up blank. It's the documents themselves that I'm interested in capturing--not the webpages that list them.

Problems Reported from Curators:

- Some of my captured content did not display correctly - some embedded images and video content were not present.

- There seems to be a problem for me in the files tab. The URL for what I consider the top level html page actually brings up a lower level html page that, in the online site, has a different URL. This is one instance, but didn't analyze in detail.

4.5 Recommendations and Considerations

Recommendations from Curators:

- Additional search criteria or capabilities:
  - File type
  - URL
  - Date or date range
  - Keyword search limited to mime type or host
- Changes to the brief item records resulting from searches:
  - Referring URL for links external to host URL
  - Indication if file is actually in the archive and if not, why not
  - Content from <meta> tags in files
  - On "table of contents" pages, include total file size or time spent on crawl
- Additional browsing and searching capabilities:
  - Provide easy access to the "home page" of a seed URL (i.e., the initial web page presented when a seed URL is accessed via a browser)
  - Provide site-level searching across the archive that results in discovery of all capture instances for a given site of interest
- Additional display and navigation features:
  - Add the capture date in the top frame when displaying files
  - Present files in a manner that allows easy access to the "home page" of a seed URL; alternately, if possible, display files hierarchically

Considerations:

- There are opportunities to improve understanding of the search and display features as well as increase confidence in the completeness of captures.
  - Create documentation that describes both what content is indexed and what searching capabilities are available.
  - Correct the disjunction between files presented in a browse list and the message "Not in Archive" returned when curator attempts to open file.
- Some curators want to collect specific publications, frequently published in PDF format. Are there optimal strategies for using the WAS that can be identified at this stage to assist in these curators?

5 Overall Reactions

5.1 WAS Help

Most curators thought the documentation provided was either very helpful or somewhat helpful in answering questions about features and functions of the WAS (Table 7).

	N	**Very Helpful**		**Somewhat Helpful**		**Not Helpful**		**Did Not Use**	
		#	**%**	**#**	**%**	**#**	**%**	**#**	**%**
Side Bar Tips and Icon Keys	16	9	56%	7	44%	0	0%	0	0%
Help Information (Detailed Guides)	17	7	41%	8	47%	2	12%	0	0%
Contextual Help	17	7	41%	5	29%	0	0%	5	29%

Table 7. Helpfulness of Documentation

Problems Reported by Curator:

- Possibly due to display settings (800 x 600) and/or the larger text size setting in the browser, Sidebar Tips and Icon Keys were displayed at the end of pages and not as expected on the right hand side of the display. Since the side bars were not initially visible, the useful information they provide was not readily discovered.
- Response Code Help text did not hyperlink to any help. Also, the Help available by selecting Help in upper right of reports page did not include Response Code Help in its contents list.

Recommendation from Curator:

- Suggested that contextual help be added for the Action icons on the Manage Sites page.

5.2 What Curators Liked the Most

1. Interface and Navigation
   - Great improvement over release 1
   - Clean, clear, intuitive, easy to use
   - Visual clues on initial screen
   - Ease of changing capture settings
   - Inclusion of URL on display screen
   - Ability to switch between file display and metadata display
2. Searching and Browsing
3. Reports
4. Filtering and Sorting Results
5. Longer Capture Times

5.3 Areas for Improvement

1. Speed
   - "Painful" response lag between pages in the interface
   - Browsing very slow
   - Reports extremely slow to appear
   - Files loaded very slowly
2. Capture Settings
   - Guidance for specifying max time
   - Add 'file type' setting: specific file types or all types
   - Add 'schedule capture job' setting so curator can specify optimal time
   - Add 'frequency' setting: daily, weekly, monthly, etc.

3. Access to Site Home Page
   - Ability to replicate site's original look and feel
   - Add a direct way to open site in browser to check links
   - Add easy access to site home pages from View Results

4. Desirable Feature Enhancements
   - Ability to search and display multiple capture results for any site in the archive, regardless of which curator captured site
   - Ability to view entire archive or a portion of archive for a specified group of curators, for example, all curators at one campus
   - Easier method(s) of dealing with the volume of files in the files list
   - Include some Overview Report data in the Manage Site table
   - Identify screen resolution and browser settings for interface
   - Ability to simultaneously browse capture results from different points in time for the same site
   - When viewing PDF file with active links, display the URL for the link along with the linked file
   - Explanation of why files were not captured
   - In addition to the page numbers and (< >), add Next Page and Previous Page text for navigation through results pages
   - Ability to print the files list in its entirety
   - For multiple captures of same site:
     - Indicate in capture results if site changed since last capture
     - Retain only non-redundant data