Curators' Evaluation of WAS Release 1

October 20, 2006

Prepared by:

Kathleen R. Murray
krmurray@unt.edu

Inga K. Hsieh
ikh0003@unt.edu

University of North Texas

Contents

Introduction

The Web Archiving Service, Release 1, was available to the project's curators from September 11, 2006 - September 26, 2006. During this time the curators had an opportunity to trial the service. Subsequent to their trials, curators submitted a brief web-based evaluation of their experience.

Overall, curators were pleased with the service and found it easy to use. Curators found the help screens and user guide useful and the interface clean. They are generally optimistic about future releases. From a usability perspective, the three functions that posed the most confusion were:

1. Starting a capture after defining it
2. Determining when a capture was completed
3. Determining how to view captured results

To address these areas, curators provided the following recommendations:

- Make the RUN CAPTURE icon more obvious; add a RUN button on the capture definition screen to initiate a newly defined capture. A RUN button to initiate a capture
- Provide service-initiated "positive feedback": after captures complete, send an email message to signal capture completion and to notify curators with a report that includes *what* was captured

Many problems with capture results were associated with incomplete captures, likely due to the 10 minute capture constraint for this trial. Some other problems seemed related to the Firefox browser or to the Wayback Machine. When evaluating their capture results, a few curators reported problems that might be of concern:

1. File listed in crawl log but not found in the archive
   - A few curators compared their crawl logs with their capture results. One curator searched the archive using file names on the crawl log and was unable to locate one PDF file listed on the log.

2. Damaged file error message for file that appeared to be intact on the site
   - "Capture of a pdf link: http://ca.water.usgs.gov/issues/ retrieved a 'file is damaged' note"

3. Displayed images turning gray
   - In one case, images initially appeared properly on displayed web pages and subsequently turned gray in color. (The curator wondered if this might be a Wayback Machine problem.)

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. Additionally, many of the observations and suggestions in the *Quick Heuristic Evaluation of the WAS* echo the feedback received from the curators. These commonalities are noted in the document.

1 Problems Encountered

Overall, the curators reported encountering few functional problems with the WAS as indicated by the percentages of 'No' responses listed in Table 1 (average 82%; range 67% - 94%). Logging into the service and viewing captured results or reports were each problematic for only one person. Four curators (roughly 1 in 4) encountered some problems with both defining sites and defining captures. One in 3 curators encountered some problems with the execution of their captures.

		No		Yes	
Function	**# Responses**	**#**	**%**	**#**	**%**
Login	18	17	94%	1	6%
Site Definition	18	14	78%	4	22%
Capture Definition	18	14	78%	4	22%
Capture Execution	18	12	67%	6	33%
View Results & Reports	18	17	94%	1	6%
Overall	90	74	82%	16	18%

Table 1. Problems Encountered

1.1 Site Definition

- Capture timeout yielded insufficient results to determine if modifications to the site and capture definitions were needed
- Site definition field length was too short and some text was lost
- Identical site names were created by making one uppercase and one lowercase
- URL format was rejected:
  - http://goleta.govoffice.com/index.asp?Type=B_BASIC&SEC={0968C707-44CF-4897-A189-30CBD22AB1CD}

1.2 Capture Definition

- Dropdown menu options and scope options did not always display when using Firefox browser
- Capture definition field length was too short and some text was lost
- Max time setting for capture was confusing

1.3 Capture Execution

- Not intuitive how to execute a capture;
- Difficult to determine what was captured;
- Max time of 10 minutes was insufficient to capture desired content
- Length of capture exceeded 3 hours; inference is this seemed excessively long

Recommendations from Curators:

- A RUN button to initiate a capture
  (*Comment*: The suggestion in the heuristic evaluation (Manage Captures, page 5) echoes this recommendation.)
- Service-initiated "positive feedback": after captures complete, send an email message to signal capture completion and to notify curators with a report that includes *what* was captured

(*Comment*: Two suggestions in the heuristic evaluation (Manage Captures, page 6, and View Results, page 6) echo this recommendation.)

Consideration:

- What type of message or report would be helpful for the WAS to send curators upon completion of a capture? What should the content include?

2 Interface & Terminology

Curators reported some confusion or lack of clarity in regard to some functions and terms. These are listed below along with their recommendations for improvement.

- Initiating or starting to run a capture: Where is this functionality? "Save Capture" icon on Create New Capture page does not start the capture (as some expected). Looking for a more obvious RUN CAPTURE icon on Manage Captures screen. (*Comment*: The confusion or misinterpretation of the Save Capture button is also reported in the heuristic evaluation (Manage Captures, page 5).)
- Finding capture results: Why is Viewing Results a separate function from Managing Captures? Why is there a View History function on the Manage Capture screen but not a View Results on that screen?
- Editing or viewing capture settings: Functionality of the Edit icon on Manage Captures screen is not intuitive; it is not obvious that capture settings can be viewed and accessed from the Edit icon.
- Determining if a capture completed
- Adding new sites after running captures: Where is this function located? Manage Sites does not imply Add New Site for some.
- Max time setting for capture: How should this be determined? What does it mean? What was the default setting? (*Comment*: Most curators seemed well aware of the default capture time setting.)
- Description of a site versus description of a capture: How are they different? (*Comment*: Perhaps this arises when only one site with one seed URL is included in a capture, which might have happened in the trial?
  (*Comment*: This is somewhat similar to an item in the heuristic evaluation (Create New Capture, page 4), which suggests that the functionality to create a simple capture of a site with one URL be combined. In a similar vein, the heuristic evaluation (Overall Suggestions, third suggestion, page 8) states: "For a simple site with only one seed the work of defining a site and defining a capture separately causes more steps to be taken than is necessary.")

Recommendations from Curators:

- Include a Getting Started tab on the interface to the left of the Captures tab. (*Comment*: This could include the information in the Overview of Capture Process section of the Release 1 WAS User Guide. "Getting Started" is a fairly intuitive starting place.)
  (Comment: The heuristic evaluation (Overall Suggestions, first suggestion, page 8) suggests a wizard interface for the capture process, which might address some of the confusion curators reported.)
- Make the RUN CAPTURE icon more obvious
- Associate a VIEW CAPTURE SETTINGS feature or option with viewing capture results

Considerations:

- Is the max time setting in the capture definition meaningful to curators or important for them to specify? If yes, how can they do so accurately or is "hit-and-miss" a good approach?
- What are the circumstances in which a curator would only specify one site per capture?

3 Help Screens

Fourteen curators used the help screens. All 14 rated the screens as either somewhat helpful or very helpful (Table 2). Overall, the help screens appear quite useful.

Rating	#	%
Not Helpful	0	0%
Somewhat Helpful	6	33%
Very Helpful	8	44%
Not Used	4	22%

Table 2. Usefulness of Help Screens (N=18)

4 Capture Results

A few curators were unable to determine the effectiveness of their capture results and a few rated the capture results as not effective. Likewise two curators did not respond to this question, one who had not viewed their results and the other for an unknown reason. Nine of the remaining 12 curators rated the capture results as moderately effective while three curators rated their capture results as very effective. (See Table 3.)

Rating	#	%
Not Effective	2	13%
Moderately Effective	9	56%
Very Effective	3	19%
Don't Know	2	13%

Table 3. Effectiveness of Captures (N=16)

Many curators reported that their captures were incomplete: pages missing and images missing. Most suggested that this was likely due to the 10 minute time constraint on captures. A few people thought it may have been due to their max data setting (1 MB or 100 MB) and another to a robots.txt barrier.

A few curators reported increasing the max data setting for an incomplete capture or executing smaller captures. In these cases the capture success improved, sometimes reaching 100%.

In one case, images initially appeared properly on displayed web pages and subsequently turned gray in color. The curator wondered if this might be a Wayback Machine problem.

One curator compared their crawl log with their capture results. The curator searched the archive using file names on the crawl log and was able to locate all but one PDF file listed in the log.

Some curators encountered error messages for pages that were not captured or were damaged but did not fully understand the reasons.

- "APP_DISPLAY_FILESYSTEM.ASP: Error:The following file was not found: 'D:datawwwSANDAG/services/sandag_general/hometext.asp'"
- "Capture of a pdf link: http://ca.water.usgs.gov/issues/ retrieved a 'file is damaged' note" (*Comment*: Inference was that the file was not damaged on the site.)
- "A sample search for a keyword in one captured site showed a missing page (404 error) that apparently was not related to a robots.txt file (at least, I didn't see it for that site listed in the crawl log)."

Some curators seemed somewhat puzzled by what was not captured:

- "There was never a case where I could capture more than 2 links away from the main site."
- "The City of Goleta capture at first reported encountering an 'internal error' but then later, the single homepage was captured but no pages below that."
- "It seemed that items that weren't there when I first viewed the results ended up being there the next day."
- "The raw data size in bytes for timed out captures ranged from 5mb to 69mb. Why the big difference?"

5 Overall Reactions

Curators also identified what they most liked and disliked about the service and had an opportunity to add any additional comments they wished. Some of the areas that presented problems for curators were known issues or the results of constraints established for the trial, for example, the 10 minute capture time limit.

5.1 What Curators Liked the Most

1. Documentation & Help Information
   - Thoughtful documentation
   - Useful user guide
   - Contextual (mouse-over) help for Capture Status (x2)
   - Helpful alt-tag content for icons
   - Help information clearly written
   - Information box containing tips, page-specific information, etc.

2. Uncluttered, Clear & Intuitive
   - Easy to read
   - Clean design for forms
   - Clean design for viewing crawl size and duration
   - Intuitive GUI; easy to understand and navigate
   - Clear, predictable, easy-to-learn interface

3. Ease of Use & Speed
   - Fairly easy to use
   - Easy to start captures
   - Fast
   - Quick and easy to define a capture
   - Quick and easy to add a new site

4. Miscellaneous
   - Captured all significant files and not extraneous files
   - Detailed crawl reports
   - "View Results" pick list

5.2 Areas for Improvement

1. Operating Confusion: Defining v. Starting Captures
   - Figuring out how "to drive" the service
   - Difference between defining a capture and starting a capture
   - Pull-down menu to execute a search rather than a button
   - "I can't tell what it did." (need for confirmation or status)
   - "Was it really still running? Was it indexing? Didn't know."

Comment:
Many curators thought that after they defined a capture, they should be able to execute it; they didn't get that they had to go to a different place to start the crawl. They were uncertain what happened when they hit that "Save Capture" button: Did the crawl execute? How would you know?

2. Indication of Capture Status: What's happening?
   - Determining when a capture was completed
   - Change in status from "Ready → Running → Ready" with no status to indicate "Completed"
   - Had to check on capture status versus being notified when a capture completed
   - One capture was initiated 9/21. Status until 9/25 was "Running", then the status changed to "Ready". The status message in a report was not available until the status was "Ready". The message provided insight into the length of time the capture status was "Running": 'capture terminated due to server unavailability'. (*Note*: The inference here seems to be that it would have been helpful to know this information at an earlier time.)
   - Indication of status as capture is being done

3. Viewing Captures
   - Hard to determine how/where to locate a captured site
   - On Manage Captures page, expected to link to the captures possibly through the "Ready" status indication
   - More helpful if viewing results was integrated with the list of captures
   - Wayback Machine interface & presentation
     - Search not as good as Google
     - The list of all files captured is not needed, "unless there is a problem" with the capture; better to present the archived copy of the site as it was presented on its original server
       (*Comment*: The potentially long list of files presented by the Wayback Machine provides excessive detail to curators interested in simply viewing a captured site. The heuristic evaluation identifies a 'length of list' issue in another functional area of the WAS (Manage Sites, page 2) and suggests creating "some kind of information hiding" to address the issue.)
     - Arabic text not displayed
     - Pages listed returned "Not in Archive" message when link was clicked
     - Need to know dates to find content versus just browsing content
   - Firefox seemed to ignore the <base href> resulting in links not functioning properly. IE did not have this problem.

4. Evaluating Capture Results & Reports
   - Detailed analysis of reports is needed to evaluate the success and completeness of captures
   - Difficult to read

- Jumbled crawl log; seemingly no delimiters in the .txt files
- Key to the response codes not accessible from or included with the Response Code Report, therefore the report was not meaningful/decipherable
- "I wasn't sure what to do with the Processor's report."
- Not the information wanted by one curator
  - Wanted report of captured files by file type
  - Wanted for PDF files: filename, size, date created, first 250 characters, indication if file had been previously captured

Comment: The heuristic evaluation reports a related issue (Detailed Crawl Results, page 7 and 8) and suggests that a consolidated report be designed.

5. Miscellaneous
   - No way to determine if someone else had already added a site; database of captured sites desirable
   - 10-minute capture constraint
   - Delayed timing of the release; need more advance notice to manage department
   - Unclear which "types" of web sites are good candidates for successful captures (need to know what types of files will encounter problems - asp, flash, etc.)
   - Inconsistent capture results among captures
   - Failure to capture anything from one site
   - Determining how to set up a daily capture (*Comment*: This feature wasn't implemented in this release.)

(*Comment*: The heuristic evaluation (Overall Suggestions, second suggestion, page 8) suggests adding a scheduling feature for captures in a wizard interface for the capture process.)

5.3 Other Feedback

Many curators repeated their praise for the overall ease of use of the service and could appreciate the work that had gone into it. Many were also eager to see future rollouts and enhancements. There was a general sentiment of "job well done" and "this is going to be great". A few provided the feedback below, which might be helpful for future development.

More options needed for configuring the crawler to capture content:
- This site
- Not this site
- Sufficient time to capture large files beyond the root URL

Unclear about how to delete:
- A capture
- A site

Refresher information will be needed in future regarding:
- Determining the number of servers for a site
- How to calculate the "level" of links -- to determine the link hop setting

Question
- Where will the captured files ultimately be preserved?

Nice Touch
- Converting crawl sizes to MB