Web Archive Quality Assessment

A Beginning for the Beginners
by the Beginners

By: Kim Schroeder,
Wayne State University
Background

- Why go to this level?
Background

- No formal academic articles to support this research.
- There are a few case study assessments on a local sense.
- Nothing to assist someone in choosing a tool or service.
- Rather than fall into the best user interface, cheapest or well supported tools, I wanted to evaluate the end results.
Background

- That being said, there is some difference in results based on user skill, technology available and the overall complexity of the sites being captured.

- The following results were based on our learning curve and do not reflect on the tools and services. This preliminary research. (Note added for public posting 7/25/2014)
Methodology

- Five of my graduate students volunteered to review the quality of captures via five different web capture processes.
- Existing web archives had their captures evaluated.
- HTTrack and Blue Crab were also used to capture sites anew.
- Archive Team was a bit of a hybrid as existing captures were evaluated as were new captures initiated in their Archive Bot.
Methodology

- A spreadsheet was created of 95 sites.
- Current sites were chosen to represent a broad-spectrum based on cultural popularity, national, regional, and minority news, Library Science Schools, and pop culture sites.
- Defunct sites captured by services such as Archive Team and California Digital Library Archives were also evaluated.
Methodology

- Each student was assigned a tool or existing website archive to evaluate the captures.
  - California Digital Library Project (WAS)
  - Archive Team
  - Internet Archive (tried to separate Archive Team feeds from IA captures)
  - HTTrack – PC Based
  - Blue Crab – Mac based
Methodology

- A criteria for evaluation was created:
  - For those existing captures housed in openly available archives:
    - Percentage closest to original – Rendering accuracy
    - Storage Format – WARC?
    - Metadata and Harvesting Data
    - History of Captures and Recency
Methodology

The Recency scale:
0 - No existing captures on this site
1- One existing capture more than 3 years old
2 - One existing capture within the last 3 years
3- Several existing captures spread over time but nothing more recent than the last 3 years
4- Several existing captures spread over time and the last within the last 3 years
5- Multiple captures spread over time including recent captures within the last 6 months.
Methodology

- A criteria for evaluation was created:
  - For those new captures:
    - Percentage closest to original – Rendering accuracy
    - Storage Format – WARC?
    - Number and Type of Ingest Errors
    - Intervention Usability
    - Robots.txt (How handled)
    - Metadata and Harvesting Capability
    - History of Captures and Recency
    - Ease of Capture for Interface
Execution

- Some students searched existing archives to determine if each site had been captured.
- Some input the sites on our list into Archive Team (if not already captured)
Execution

- Another team input the site list into Blue Crab ($24) and HTTrack (Free) to assess results. These are considered web freezers which allow for offline browsing of websites.

- Only website homepages were evaluated at this time. Though several were captured in whole this deeper evaluation is Phase 2.
Time Analysis – New Captures

- Majority of time in captures was spent nursing sites through the process:
  - 1) Internet Locks Up
  - 2) Software Freezes
Time Analysis - New Captures

- 3) Evaluating capture failures
  - technology problems or
  - robots.txt limitations

NOTE: Error Logs were critical in evaluation. Those with hard to access logs or unavailable logs, utterly limit the user’s ability to problem solve.
Time Analysis - Existing Captures

- Majority of time searching existing sites to find urls was slow with no searchability. All but the Internet Archive was only folder searchable by topic only so searching was slow.
Results by Tool – Archive Team

- **Archive Bot**
  - Designed by Jason Scott and his organization. Is defined a crowd-sourced crawler.
  - [https://archive.org/details/archivebot](https://archive.org/details/archivebot)
  - Volunteers and zealots saving at risk sites.
  - Some content is on their site and some they have uploaded to the Wayback Machine.
  - Allowed us to use Archive Bot for our additional captures.
Results by Tool - Archive Team

- Overall the 95 sites that were reviewed resulted in an assessment completion of 77 percent which gathered some data on the site.
- The number is drawn down by sites that simply did not copy at all. This is partially due to robots.txt restrictions and other errors that would have to be nursed through the process.
Results by Tool – Archive Team

- Recency was strong coming in at 4.9 out of 5.0 but partially due to our own ingest actions.
Top Stories

Fatal shooting investigated near downtown
Police are on the scene of a fatal shooting on Braden Street.
More

Clouds gather Sunday night, showers loom
Showers, storms possible Monday and Tuesday
More
Results by Tool – HTTrack (PC Based)

- Self-described “offline browser utility tool”
  - [https://www.httrack.com/](https://www.httrack.com/)
- Of the 95 sites examined with this tool:
  - 42% were captured at 80 percent or higher assessment.
  - 32% of sites had no content captured.
  - Tended to be either no capture or full. With two exceptions, the zero captures were robots.txt issues. For instance, Disney.com copied faster than other sites and delivered no visual content only this:
Site mirroring finished - [disney.wmt]

Minoring operation complete.
Click Exit to quit WinHTTrack.
See log file(s) if necessary to ensure that everything is OK.

Thanks for using WinHTTrack!
Tip: Click [View log file] to see warning or error messages.

HTTrack3.48-17+hlts.gif+hlts.java launched on Tue, 15 Jul 2014 10:33:05 at disney.com by "httrack" (HTTrack 3.48; Windows 7) as "Mozilla/4.5 (compatible; HTTrack 3.48; Windows 7)"

Warning/Errors log file

[Logs and warnings related to the mirroring process are shown here.]

Information, Warnings, and Errors reported for the mirror:
- The mirror log file, and its cache folder, may contain sensitive information, such as username/password authentication for websites mirrored in this project.
- Do not share these files/folders if you want these information to remain private.

10:33:05 Warning: Moved permanently to disney.com/robots.txt.

HTTrack Website Copier/3.48-17 mirror complete in 1 seconds: 2 links scanned, 1 files written (14,538 bytes overall), [33811 bytes]
(no errors, 3 warnings, 0 messages)
Screenshot of captured website [www.archives.gov](http://www.archives.gov) (many of the captured websites were not allowed to run scripts or Active X controls)
Screenshot of www.archives.gov home page
HTTrack 3.48-13+htsswf+htsjava launched on Sat, 12 Jul 2014 19:29:00 at
(winhttrack -qwr0%e1C2%Ps2u1%s%uN0%l0p3DaK0H0%kf2A25000%f#f -F "Mozilla/4.5 (compatible; HTTrack 3.0x; Windows 98)" -%F "<!-- Mirrored from %s%s by HTTrack Website Copier/3.x [XR&CO'2014], %s -->

Information, Warnings and Errors reported for this mirror:
note: the hts-log.txt file, and hts-cache folder, may contain sensitive information,
such as username/password authentication for websites mirrored in this project
do not share these files/folders if you want these information to remain private
19:29:00 Warning:  Found for occupyphx.org/robots.txt
19:29:00 Warning:  Redirected link is identical because of 'URL Hack' option:
occupyphx.org/robots.txt and occupyphx.org/robots.txt
19:29:00 Warning:  Can not bear crazy server (Found) for
occupyphx.org/robots.txt

HTTrack Website Copier/3.48-13 mirror complete in 1 seconds : 2 links scanned, 1 files written (282 bytes overall) [771 bytes received at 771 bytes/sec], 282 bytes transferred using HTTP compression in 1 files, ratio 118%
(No errors, 4 warnings, 0 messages)
Results by Tool – HTTrack (PC Based)

- Testing was done to adjust settings to go 0, 1, 2, and 3 levels deep.
- Sometimes this fixed problems such as missing social media boxes, videos, links, etc.
- Sometimes it did not.
- When tested on the Michigan Chronicle site all four tests turned out exactly the same which was a complete capture of the Homepage with interactivity to the links.
Results by Tool – Blue Crab (Mac Based)

- Also marketed as an offline browser.
- Overall the tool was strong at capture and the brunt of the moderate captures were robots.txt based and not the fault of the tool.
Publisher's Description

From Limit Point Software:

Blue Crab is a versatile and thorough program that you use to copy the contents of a web site to your computer, in whole or in part, and then search or browse it offline.

+ Conduct fast offline browsing and searching without an internet connection.
+ Perform batch downloads of URL's
+ Create a snapshot of a website for historical archiving.
+ Batch download web archives.
+ Collect specific types of resources such as images or email addresses.
+ Search current content more thoroughly than a search engine right on your own computer.
+ Check a site for broken links, or generate an HTML sitemap
+ Create full page images of URL's (JPEG, BMP, TIFF, PNG, PSD, etc.)
+ Perform "Google Image Search" batch downloads: download images found using an image query on Google.

With Blue Crab you can download all the content including HTML, PDF, graphics, video, file archives, etc., or use selective filtering to restrict downloads to specific kinds of files. For example, you can choose to save only the JPEG images Blue Crab finds, or just the PDF's.

Blue Crab has a special feature called the "Media Grabber" which you can use to easily download just the images (or movies) on a web site. Moreover, you can view a slide show of the images as they are downloaded. You also have the option of "flattening" the images, i.e., putting them all into one folder, or preserving the folder structure on the server (just as when downloading a complete web site for offline use).
Results by Tool – Blue Crab (Mac Based)

- Of the 55 sites evaluated in Blue Crab, an average of 77 percent of sites examined captured some data.
- This number was generated via the “Download One Page” feature to assess the homepages only. This meant many links, ads and videos were not working. To fix this a custom profile was created going three levels deep and in the sample group, this generated more complete captures.
Results by Tool – Blue Crab (Mac Based)

- Seventy three percent of the captures were high content captures (80% or higher).
- Average capture time was 47 seconds for just the one page capture. There was a high of 260 seconds and a low of zero seconds in capture. Length of time was not an indicator for capture success.
Thank you for visiting RadioShack. If you need assistance with shopping on our site, please call us at 800-843-7422 and a customer care representative will be happy to assist you. Please inform the Customer Service representative that you require assistance.

- 24/7 Customer Service (800) 843-7422
- My Store:
  - Select Your Store
- Recently Viewed
- My Cart 0 items $0.00
- My Account
- Store Finder
- Deals
  - What's On Sale
  - Weekly Ad
  - Clearance
- Home Services Setup
  - Screen Replacement
  - Mobile Product Support
  - Trade-In Service
  - Mobile Security & Virus Removal
  - IT Happens Protection Plan
  - RadioShack Credit Card
  - DSM Network
- Weekly Ad

- CELL PHONES & ACCESSORIES
  - Cell phones & plans
    - Apple iPhone
    - No contract cell phones
    - No contract SIM cards
    - Unlocked cell phones
    - Cell phone accessories
    - Trade & Save Program
- ELECTRONICS & ACCESSORIES
  - Phones & Radio Communications
    - TV & Video
    - Music & Audio
    - Fitness & Health
  - Computers & Tablets
    - GPS & Car
    - Camera & Camcorders
    - Home & Office
    - Video Games & Toys
- HOBBY & DO-IT-YOURSELF
  - DIY Projects
File not found

Firefox can't find the file at
/www.facebook.cor
/plugins
/likebox.php?ref=
/www.facebook.cc
/su.ischool&w
idth=300&
height=388&
colorscheme=light&
show_faces=false&
show_border=false
stream=true&
header=false.

MORE POSTS...
BlueCrab quit unexpectedly.
Click Reopen to open the application again. Click Report to see more detailed information and send a report to Apple.
California Digital Library - WAS

This Web Archiving Service “The University of California’s Web Archiving Service (WAS) helps a wide variety of institutions—from small institutes to large public universities—create enduring digital archives of fragile web resources and safeguard them in long-term private or publicly accessible.”

http://webarchives.cdlib.org/about
Why Use Web Archiving Service?

- Archive your institution’s website
- Build new collections
- Capture political & social events
- Save at-risk websites

Archive today's news for continuing study and analysis.

Learn more
California Digital Library - WAS

- Of the 95 sites listed, 12 were in WAS
- Major search issue was the folder structure for examining sites.
- Via an email, we received an inventory list to assist in findability as the site is organized by files and the internal search box for each site does not do boolean.
- Dates of capture were primarily 2012 and 2013, with one being from 2014.
One of the most useful aspects of this format was the header. There is sometimes a category included like:
California Digital Library - WAS

- Archive is arranged by folders which is fine for browsing not for searching.
Archives

Web Archiving Service (WAS) subscribing institutions have created more than 60 public archives.

These unique collections of websites are organized around a variety of diverse topics, historical events, geography, and more. They provide researchers with lasting access to ephemeral web content.

2003 California Recall Election
2007 Southern California Wildfires Web Archive
2009 H1N1 Influenza A (Swine Flu) Outbreak
2010 Winter Olympics
AFL-CIO - Change to Win: the open web archive
African Politics Web Archive
Alternative Mass Media / News Web Site Web Archive
Anarchism Web Archive
Animal Rights
Arts and Cultural Left Web Archive
At Your Service Web Archive
California Aggie, UC Davis Student Newspaper
California Political Blogs
California State Government: .ca.gov Web Archive
California Tobacco Control Web Archive
California Water Districts Web Archive
Chinese Environmental NGOs
Michigan Historical Collections Web Archives
Bentley Historical Library, University of Michigan

Description
The Bentley Historical Library's Michigan Historical Collections (MHC) is committed to documenting the history and everyday life of the State by preserving material in a variety of formats, including materials available on the Web. As part of the Bentley Historical Library's Web Archiving initiative, the MHC Web Archives preserves, describes, and provides access to websites of historically-significant religious, governmental, educational, cultural, and academic organizations and institutions around the State...

more...

Search

Quick Facts
Sites: 179
Oldest site: 08/04/10
Most recent site: 01/15/14

Powered by the Web Archiving Service from the California Digital Library.
Materials in these web archives are archived copies for private study, scholarship and research.
Copyright © 2007-2014 The Regents of The University of California.
Detroit Digital Justice Coalition Web Archives

Latest Starting URL
http://detroitdc.org/ (live link)

Description
Detroit Digital Justice Coalition (DDJC) web site contains information about local events and workshops, employment opportunities in the field of electronic technology and media, articles and commentaries, as well as DDJC online magazine "Communication is a Fundamental Human Right." DDJC is comprised of people and organizations in Detroit who believe that communication is a fundamental human right through activities that are grounded in the digital justice principles of: access, participation, common ownership, and healthy communities.

Captured
06/03/11 01:22 AM
12/02/11 01:05 PM
05/01/12 10:18 PM
05/01/13 03:51 PM

Creator:
Detroit Digital Justice Coalition

Publisher:
Detroit Digital Justice Coalition

Subjects:
Digital divide.
Digital media -- Job vacancies -- Michigan -- Detroit.
Information superhighway.

Geographic coverage:
Detroit (Mich.)

Topics
Commerce and industry
Occupy Web Archive
UCLA Library

Occupy Los Angeles

Latest Starting URL
http://obrag.org/ (live link)

Description
Occupy Los Angeles

Captured
03/18/12 09:47 PM
03/20/12 10:39 PM
04/04/12 12:46 AM
05/02/12 09:15 PM
05/05/12 06:26 PM
07/02/12 07:46 PM
07/27/12 01:53 PM
09/27/12 01:53 PM
09/27/12 01:53 PM
10/27/12 01:53 PM
11/27/12 01:53 PM
12/27/12 01:53 PM
01/27/13 01:53 PM
02/27/13 01:53 PM
09/27/13 01:53 PM
09/27/13 01:53 PM
09/27/13 01:53 PM
09/27/13 01:53 PM
09/27/13 01:53 PM
09/27/13 01:53 PM
09/27/13 01:53 PM
09/27/13 01:53 PM
09/27/13 01:53 PM
09/27/13 01:53 PM
09/27/13 01:53 PM
09/27/13 01:53 PM
09/27/13 01:53 PM
09/27/13 01:53 PM
09/27/13 01:53 PM
09/27/13 01:53 PM

this is NOT the Occupy Los Angeles website
3 Unit Apartment on 4900 Block of West Point Loma Sold for Close to $1 Million

by FRANK GORMLE on JULY 18, 2014 0 COMMENTS
#ECONOMY, ENVIRONMENT, OCEAN BEACH

The Daily Transcript (subscription) / July 16, 2014

A 3-unit apartment on West Point Loma Avenue has just been sold for close to a million bucks. Buyer Robb A. Murphy just spent $950,000 for the 2-story structure and yard at 4960-4962 W. Point Loma Blvd.

The building has 2,016 square feet total and is on a 5,624-square-foot lot (assessor's parcel 448-230-02). Within the 3 units, there's only four bedrooms and three
Of the twelve sites, 74% of the content was captured.

With 58% of those sites being a high capture of 80% of the site or more being captured.

Recency was a 4.0 on a 5.0 scale.
Internet Archive – Wayback Machine

- This well known project describes itself as taking snapshots of the web. “Capture a web page as it appears now for use as a trusted citation in the future.”

  https://archive.org/web/
Internet Archive – Wayback Machine

- Of the 95 sites listed, only one site was not referenced in Wayback Machine in some manner.
- Almost 16% had no or almost no (title only) content captured. The reasons were mainly due to publishers locking down the data or corruption (random symbols) in the capture.
Seventy-four percent were at a high of 80% content captured or higher.

Typically, the results were of a higher capacity captures with 60 percent being graded as having 90% or higher capture completion.

Recency averaged to a 4.2 out of 5 point scale.
<table>
<thead>
<tr>
<th></th>
<th>Completeness Average</th>
<th>High Capture (80% plus)</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Captures Evaluated</td>
<td>61%</td>
<td>58%</td>
<td>Discrepancy can not be assigned to the tools/services at this point. More testing of settings needs to be established in Phase 2. This will enable increased efficiency for using each tool within the diverse world of web design.</td>
</tr>
<tr>
<td>High Average</td>
<td>77%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Average</td>
<td>22%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Average</td>
<td>74%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low Average</td>
<td>42%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completeness Average</td>
<td></td>
<td></td>
<td>Defined as percentage of Homepage Content Captured divided by all sites evaluated.</td>
</tr>
<tr>
<td>High Capture</td>
<td></td>
<td></td>
<td>Defined as percentage of sites that captured content at 80% or higher as assessed by our team.</td>
</tr>
</tbody>
</table>
Web Archiving Project Planning - Lessons Learned

- The complexity of the sites to be captured will affect your choice in tools/services.
- A quantitative and qualitative evaluation is necessary due to the variables in content and within the capabilities of the tools/services.
Web Archiving Project Planning - Lessons Learned

- If you have the staff, look for a more advanced tool/service that allows for options to nurse the content through the system rather than accepting incomplete or failed captures.
- Understand your tolerance for failure and what you define as failure.
- Develop relationships with the web publishers to break through robots.txt limitations.
“We have updated our model for capture: Previously, we captured the same sites quarterly. What we have done is made it more flexible so that we alternate sites that instead of capturing all of the same sites quarterly, we now capture a number of different sites each quarter. So, for the same price,..” instead of capturing one site four times per year, we now capture 4 different sites in that same calendar year.
Web Archiving Lessons Learned – Coca Cola

“The other thing we found worked better on our social media captures was instead of capturing quarterly to scroll back 120 days, the technology captures better if we go monthly to scroll back 60 days. Due to the way old content is pushed down on social media sites, the crawlers do better with more frequent, shorter scroll back crawls.”
Would have liked a cleaner way to differentiate between Archive Team and IA captures. We were careful but do not feel 100% sure that an IA capture may have been attributed to AT and vice versa.

Archive Team in IA
Disclaimers

- Only looked at Homepages, would like to get into greater depth to assure content capture is accurate overall.
  - This will take processing power and time as well as greater storage.
  - We were all learning so this is just a first step.
Phase 2 - Next Steps

- More widespread and greater depth of evaluation.
- Deeper consultation with experts.
- Open it to the professional community.
- Better clarify classifications such as “site sucker”, “web freezer”, “web dump”, “web capture tool”, “offline browser”, etc.
- Establish templates for setting up Institutional Web Archiving Programs.
Phase 2 - Next Steps

- Look at diversity of captures and how each works on the back-end to understand why.
- Ex. Blue Crab was able to capture this site, IA and HTTrack did not.
Additional Sources

- https://www.google.com/search?q=web+archiving+basics&oq=web+archiving+basics&aqs=chrome..69i57.6455j0j1&sourcei=d=chrome&ie=UTF-8#
- http://www.slideshare.net/eibeed/ch-4-381
The Team

We have a fantastic group of students that assisted in evaluating the sites:
- Lauren Schroeder - Michigan
- Courtney Whitmore - Michigan
- Laura Gentry - Alabama
- Aubrey Maynard - Iowa
- Margaret Diaz - Arizona
Additional Thanks

- Jason Scott and his programmers at Archive Team were a wealth of assistance and openness!
- Kevin Barton, our SLIS Technology GSA who made sure that our technology kept running!
Contact

- If you are interested in the next phase, please contact me at:

- Kim Schroeder  
  Coordinator, Archival Program  
  Lecturer and Career Advisor  
  Wayne State University  
  School of Library and Information Science  
  Faculty Advisor for National Digital Stewardship Alliance  
  [http://wsustudentndsa.wordpress.com/](http://wsustudentndsa.wordpress.com/)  
  ag1797@wayne.edu  
  313 577-9783