

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>
 - 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.

<http://www.bso.org/> [history]

e8bf10emlqaacx6kjbw63dlj5

```
200 OK http://www.bso.org/Content/themes/base/images/ui-icons_cd0a0a_256x240.png
200 OK http://www.bso.org/Content/themes/base/images/ui-icons_cd0a0a_256x240.png
200 OK http://www.bso.org/scripts/mediacenter/font/fontawesome-webfont.ttf?v=4.0.3
200 OK http://www.bso.org/scripts/mediacenter/font/fontawesome-webfont.svg?v=4.0.3
200 OK http://www.bso.org/scripts/mediacenter/font/fontawesome-webfont.woff?v=4.0.3
200 OK http://www.bso.org/scripts/mediacenter/font/fontawesome-webfont.eot
200 OK http://www.bso.org/scripts/mediacenter/font/fontawesome-webfont.eot?v=4.0.3
200 OK http://www.bso.org/Content/images/icons/spinner.gif
200 OK http://www.bso.org/Content/images/icons/loading.gif
200 OK http://www.bso.org/Content/Fonts/FunctionPro_Book/FunctionPro-Book-webfont.ttf
200 OK http://www.bso.org/Content/Fonts/FunctionPro_Medium/FunctionPro-Medium-webfont.ttf
200 OK http://www.bso.org/Content/Fonts/FunctionPro_Book/FunctionPro-Book-webfont.eot?iefix
200 OK http://www.bso.org/Content/Fonts/FunctionPro_Book/FunctionPro-Book-webfont.svg
200 OK http://www.bso.org/Content/Fonts/FunctionPro_Bold/FunctionPro-Bold-webfont.eot
200 OK http://www.bso.org/Content/Fonts/FunctionPro_Book/FunctionPro-Book-webfont.woff
200 OK http://www.bso.org/Content/Fonts/FunctionPro_Medium/FunctionPro-Medium-webfont.eot
```

3.65 MB 1 minute 2.45 resps/sec 1xx: 0 2xx: 131 3xx: 0 4xx: 3 5xx: 0 Unknown: 0

Show ignores Pause output

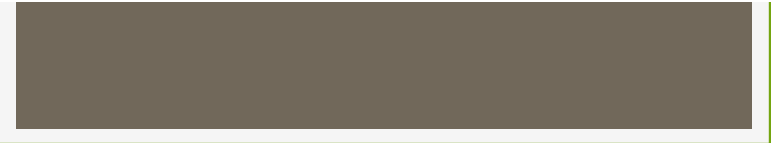
<http://www.chinadaily.com.cn/> [history]

505np8kj6q2schd1cbmr0a7ro

```
200 OK http://www.chinadaily.com.cn/image_e/2013/ico_poll.png
200 OK http://www.chinadaily.com.cn/image_e/2013/ico_fdj.png
200 OK http://www.chinadaily.com.cn/image_e/2013/ico_mail.png
200 OK http://www.chinadaily.com.cn/image_e/2013/ico_com.png
200 OK http://www.chinadaily.com.cn/js/jquery-1.6.js
200 OK http://www.chinadaily.com.cn/image_e/2013/blue_k.gif
200 OK http://www.chinadaily.com.cn/image_e/2013/red_k.gif
403 OK http://www.chinadaily.com.cn/js/2013e/
200 OK http://www.chinadaily.com.cn/image_e/2013/arr_bot.gif
200 OK http://www.chinadaily.com.cn/image_e/2013/ico2.gif
200 OK http://www.chinadaily.com.cn/image_e/2013/line1.gif
404 OK http://www.chinadaily.com.cn/image_e/2013/scrBg.gif
200 OK http://www.chinadaily.com.cn/image_e/2013/ico_poll.gif
200 OK http://www.chinadaily.com.cn/image_e/2013/ico_plan.gif
200 OK http://www.chinadaily.com.cn/image_e/2013/bi.png
200 OK http://www.chinadaily.com.cn/image_e/2013/l_arr.png
```

1.95 MB a few seconds 11.01 resps/sec 1xx: 0 2xx: 103 3xx: 0 4xx: 3 5xx: 0 Unknown: 0

Show ignores Pause output



<http://www.chinadaily.com.cn/> [history]

505np8kj6q2schr1cbmr0a7ro

```

200 OK http://www.chinadaily.com.cn/image_e/2013/ico_poll.png
200 OK http://www.chinadaily.com.cn/image_e/2013/ico_fdj.png
200 OK http://www.chinadaily.com.cn/image_e/2013/ico_mail.png
200 OK http://www.chinadaily.com.cn/image_e/2013/ico_com.png
200 OK http://www.chinadaily.com.cn/js/jquery-1.6.js
200 OK http://www.chinadaily.com.cn/image_e/2013/blue_k.gif
200 OK http://www.chinadaily.com.cn/image_e/2013/red_k.gif
403 OK http://www.chinadaily.com.cn/js/2013e/
200 OK http://www.chinadaily.com.cn/image_e/2013/arr_bot.gif
200 OK http://www.chinadaily.com.cn/image_e/2013/ico2.gif
200 OK http://www.chinadaily.com.cn/image_e/2013/line1.gif
404 OK http://www.chinadaily.com.cn/image_e/2013/scr8g.gif
200 OK http://www.chinadaily.com.cn/image_e/2013/ico_poll.gif
200 OK http://www.chinadaily.com.cn/image_e/2013/ico_plan.gif
200 OK http://www.chinadaily.com.cn/image_e/2013/bi.png
200 OK http://www.chinadaily.com.cn/image_e/2013/l_arr.png

```

1.95 MB
 a few seconds
 11.01 resps/sec
 1xx: 0 2xx: 103 3xx: 0 4xx: 3 5xx: 0 Unknown: 0

Show ignores



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.



SAME-DAY CARE
Call, click or come in.

LEARN HOW »



Notable deaths of 2014



75° Detroit
Few Clouds

Site Web

Search **GO**

Home News Consumer Weather Sports Lifestyle Entertainment Contests About Local 4 Seen On

Top Stories

July 13, 2014



Mild Sunday night, more showers Monday

Lower temperatures Tuesday, Wednesday [More](#)

- Off-duty officer shoots man with ax outside church
- ◀ Mild Sunday night, more showers Monday
- Livonia man shot while taking a walk
- Husband threatens to kill wife in Rochester Hills 49 m
- Business owner sues police over raid
- World Cup final: Germany defeats Argentina
- Tigers finish first half against Royals Sunday
- Detroit City FC beats Fort Pitt, 3-1
- Grilled & topped: Get your hands on a gourmet hot dog

[View More](#)

Coming Up On Local 4 News ...



Are all sunscreens safe for children?

Sunday at 11 p.m. - We'll show you what you need to look for and what to avoid.

StormPins APP
SEE & SHARE WEATHER AND MORE IN REAL TIME

SPONSORED BY **TOM HOLZER**

Video & Live Events

- Local
- National
- Live Feeds
- Entertainment



Watch Local 4 News here



Off-duty officer shoots man with ax outside church



Flashpoint 7/13/14

A Special Getaway

PETOSKEY AREA.COM

Petoskey • Harbor Springs • Bay Harbor • Boyne City

Most Popular on ClickOnDetroit

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](#)



[Back To Mobile Site](#)



Most powerful celebrities of 2014

Site Web

Search

- [Home](#)
 - [Ford Fireworks](#)
 - [Mackinac Policy Conference](#)
 - [Most Popular](#)
 - [Sign Up For Email Newsletters](#)
 - [Get Mobile Text Alerts](#)
 - [ConnecTV](#)
 - [Contact Us](#)
- [News](#)
 - [National News](#)
 - [Live in the D](#)
 - [GM Investigation](#)
 - [Detroit Bankruptcy](#)
 - [Video](#)
 - [Politics](#)
 - [Flashpoint](#)
 - [Local 4 Defenders](#)
 - [Technology](#)
 - [Trending News Channel](#)
- [Weather](#)
 - [Local 4Casters](#)
 - [Radar](#)

Results by Tool – HTTrack(PC Based)

- Self-described “offline browser utility tool”
 - <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:



file:///C:/My%20Web%20Sites/disney%202/disney.com/index.html

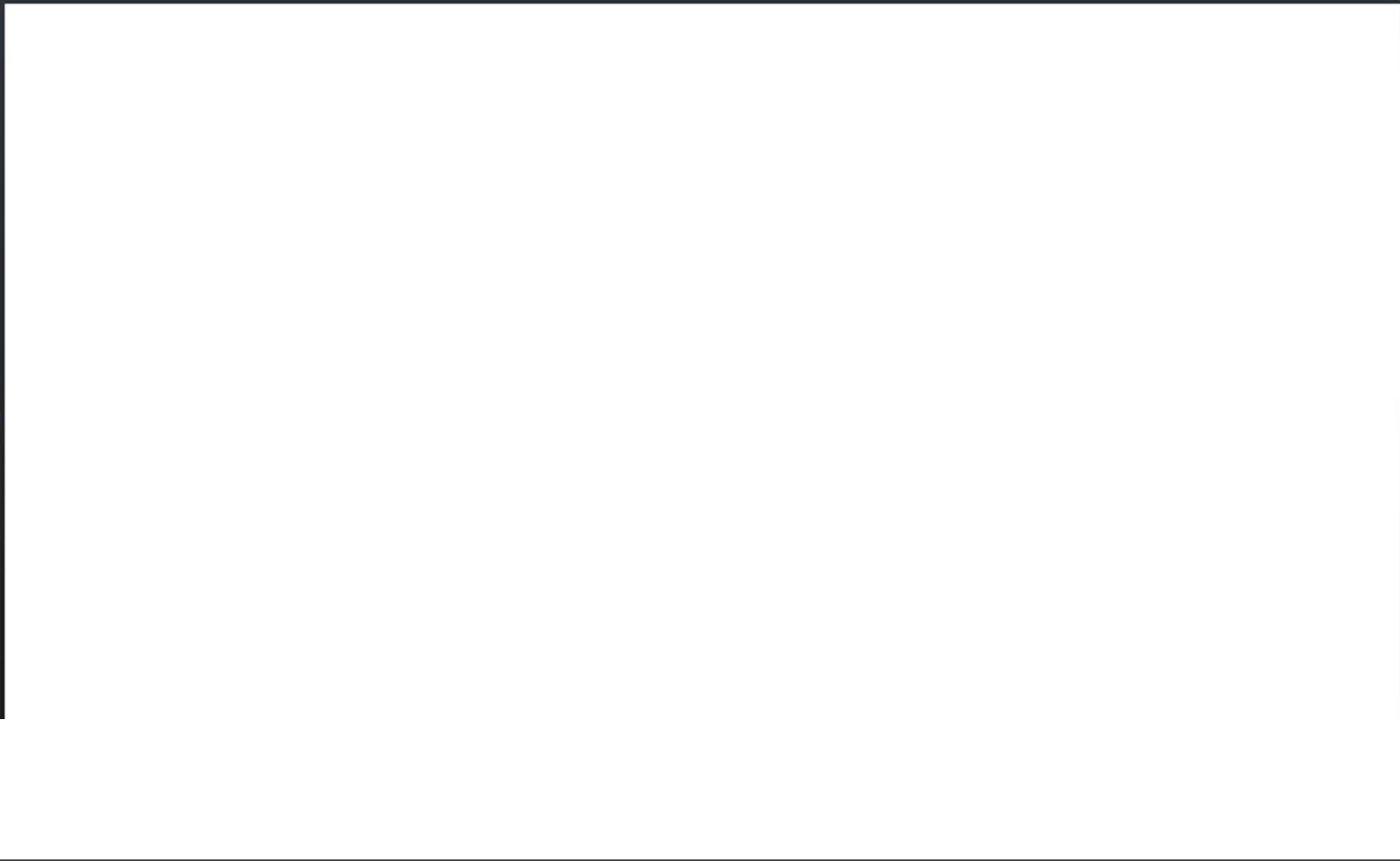


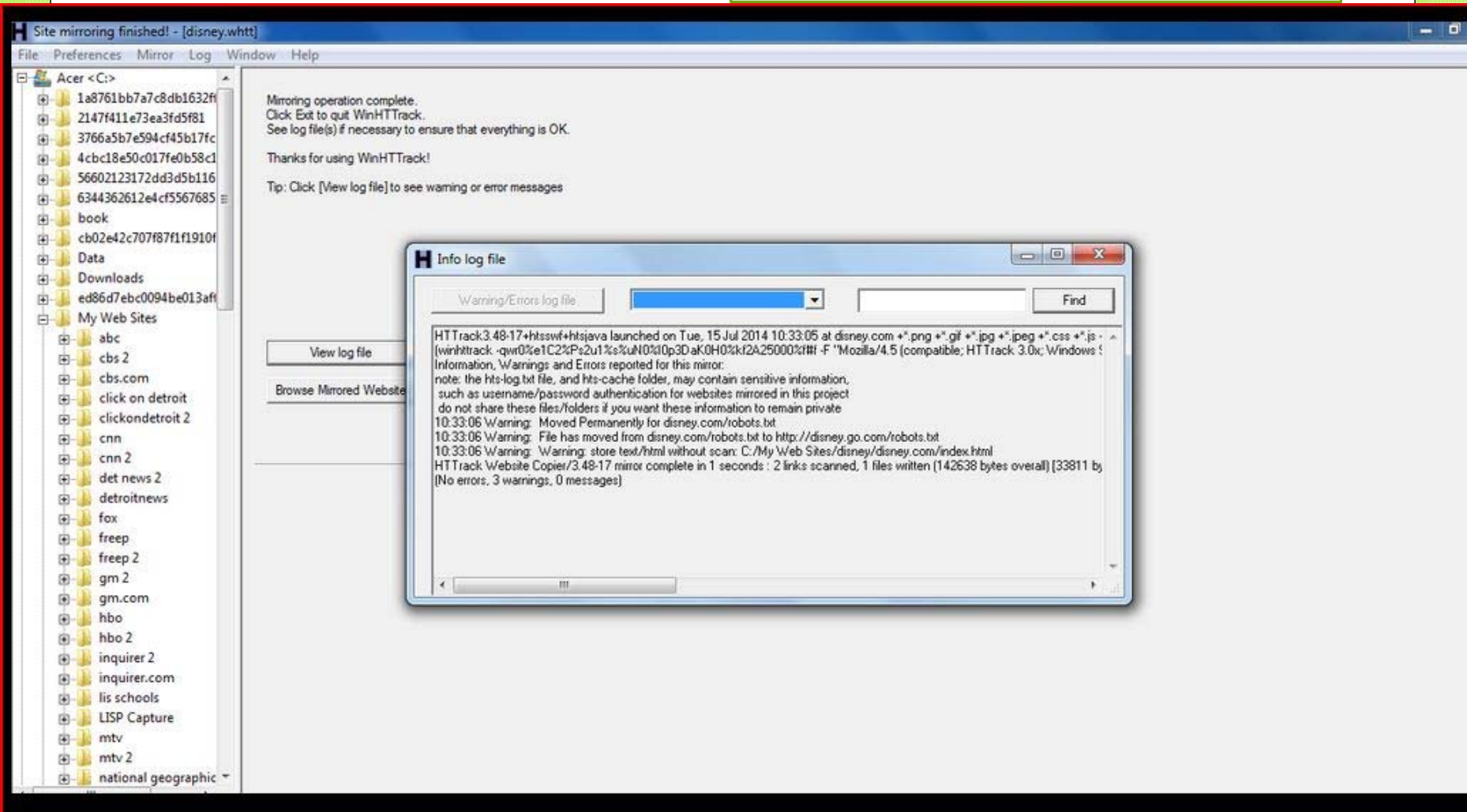
[Disney.com](#) [Store](#) [Parks & Travel](#) [Video](#) [Movies](#) [Shows](#) [Music](#) [Games](#) [Books](#) [Live Shows](#)

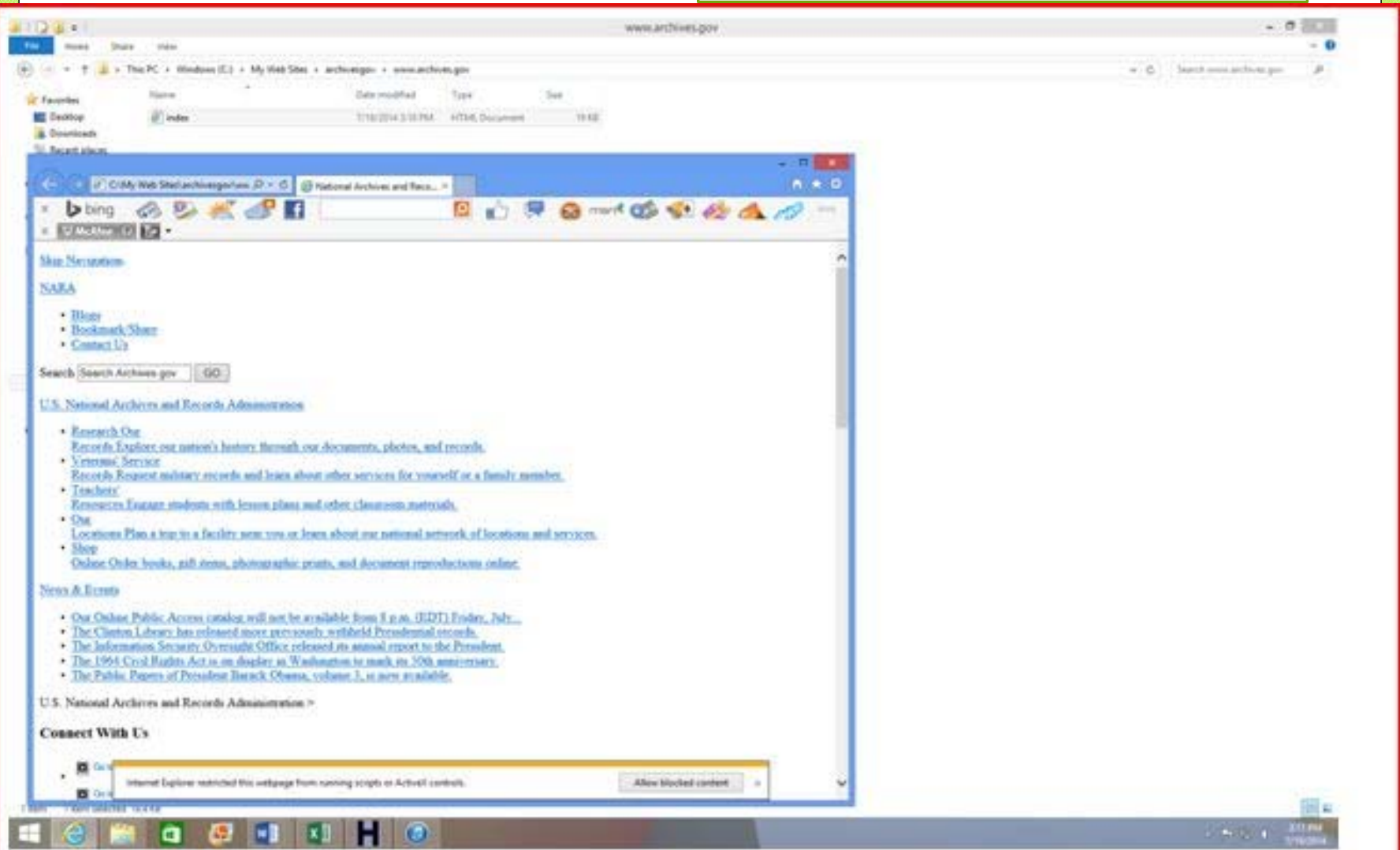
Search



[Games](#) [Video](#) [Blogs](#) [Store](#) [Parks & Travel](#) [TV](#) [Movies](#) [Music](#)







Screenshot of captured website 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

HTTrack3.48-13+htsswf+htsjava launched on Sat, 12 Jul 2014 19:29:00 at
http://occupyphx.org/ +*.png +*.gif +*.jpg +*.jpeg +*.css +*.js -ad.doubleclick.net/* -
mime:application/foobar
(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 -->" -%l "en, *"
http://occupyphx.org/ -O1 "C:\My Web Sites\occupyphx" +*.png +*.gif +*.jpg
+*.jpeg +*.css +*.js -ad.doubleclick.net/* -mime:application/foobar)
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
19:29:01 Warning: Warning: store text/html without scan: C:/My Web
Sites/occupyphx/occupyphx.org/index.html
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.



Visit Site

External Download Site

Average User Rating:



out of 9 votes

[See all user reviews](#)

Quick Specs

Version:

5.0.07

File Size:

7.83MB

Date Added:

January 18, 2014

Price:

Purchase; \$24.99 to buy
[\(Buy it now\)](#)

Operating Systems:

Mac OS X
10.6/10.7/10.8/10.9

Total Downloads:

4,681

Downloads Last Week:

6

Product ranking:

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

dw.cbsi.com/redir?tag=offer_click&lop=link&ptid=3000&pagetype=product...ring.)

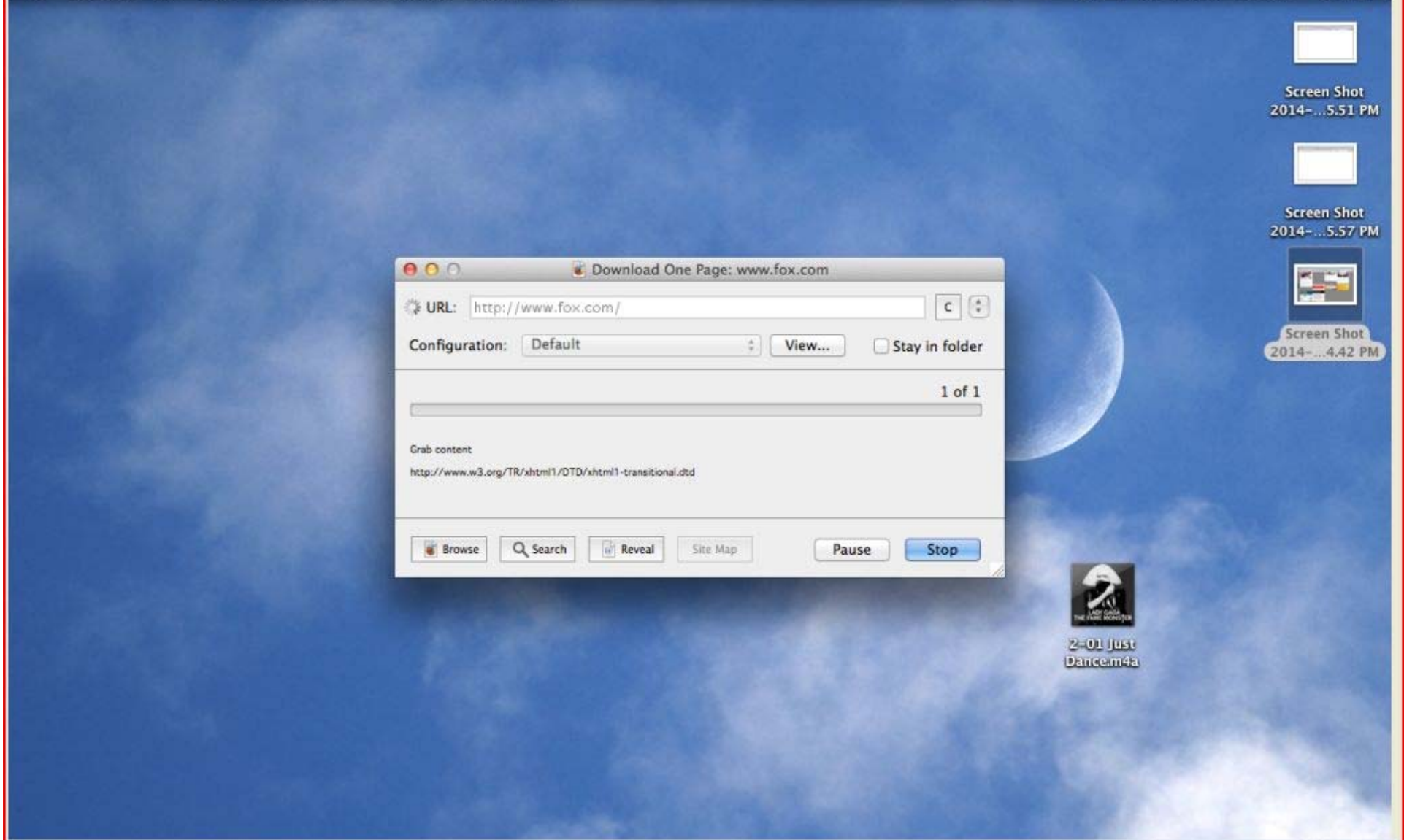


Blue Crab for Mac - ...

Microsoft PowerPoi...

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.



Download One Page: www.fox.com

URL:

Configuration: Default View... Stay in folder

1 of 1

Grab content
<http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd>

Browse Search Reveal Site Map Pause Stop



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.



Firefox File Edit View History Bookmarks Tools Window Help

Sat 2:13 PM faculty

Committed to you, your p... Mozilla Firefox Start Page Welcome - School of Library a... WSU Libraries University of Michigan Sch...

file:///Users/faculty/Library/Application Support/Limit Point Software/Blue Crab/Grabbed Files/www.si.umich.edu/Index.html Google

- Prospective students
- Faculty and staff
- Prospective faculty
- Media
- Current students



Informatics program

Associate Professor Kai Zheng is named acting director of the Health Informatics program, offered jointly by UMSI and the School of Public Health.

[Read more](#)



Social Media Feeds

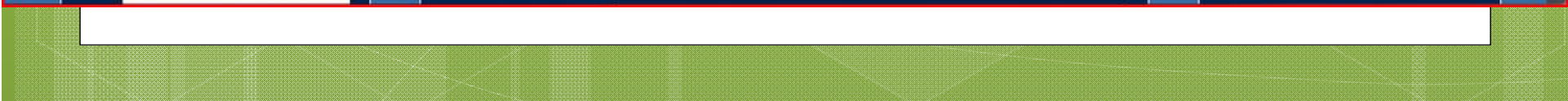
News

Kai Zheng to serve as interim director of Health Informatics program

Associate Professor Kai Zheng will serve as acting director of U-M's Health Informatics program, a graduate program offered by the School of Information and School of Public Health.

Videos

Each student has been evaluating an information system over the course of the semester



TextEdit File Edit Format View Window Help index.jsp — Locked

Times Regular 12 B I U 1.0

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17

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:**
 - [\(Select Your Store\)](#)
- [Español](#)
- [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 & Save](#)
 - [Mobile Security & Virus Removal](#)
 - [IT Happens Protection Plan](#)
 - [RadioShack Credit Card](#)
 - [DISH 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](#)
 - [Cameras & Camcorders](#)
 - [Home & Office](#)
 - [Video Games & Toys](#)
- **HOBBY & DO-IT-YOURSELF**
 - [DIY Projects](#)



Archives & Digital Content Management

Have you ever considered what archaeologists, anthropologists, sociologists and historians will study about people in the future?

[Read more](#)



Library Services

Are you curious about how people seek information?

[Read more](#)



Each graduate course costs about \$2000. All online students pay pay in-state tuition.

[Read more about tuition](#)

Contact

Get an MLIS degree in 36 credits, 4-6 semesters.

APPLY NOW!

Attend an information meeting

INFORMATION MANAGEMENT



WHAT WE OFFER

SLIS offers a 36-credit Master of Library and Information Science (MLIS) graduate degree. We also offer four pre/post Master's certificates, a joint MLIS/MA with the Department of History and the School Library...

[Read more](#)



Tweets by @iSchoolSU

MORE TWEETS...



File not found



Firefox can't find the file at
`/www.facebook.com/plugins/likebox.php?href=//www.facebook.com/su.ischool&width=300&height=388&colorscheme=light&show_faces=false&show_border=false&stream=true&header=false.`

MORE POSTS...





Screen Shot 2014-07-18 at 5:47:53 PM.png
Type: PNG Image
Size: 500 KB
Dimension: 1280 x 800 pixels

On Tonight

M T W T S S

8/7c **MASTERCHEF**

The remaining 14 home cooks split into two teams

9/8c **24 - LIVE ANOTHER DAY**

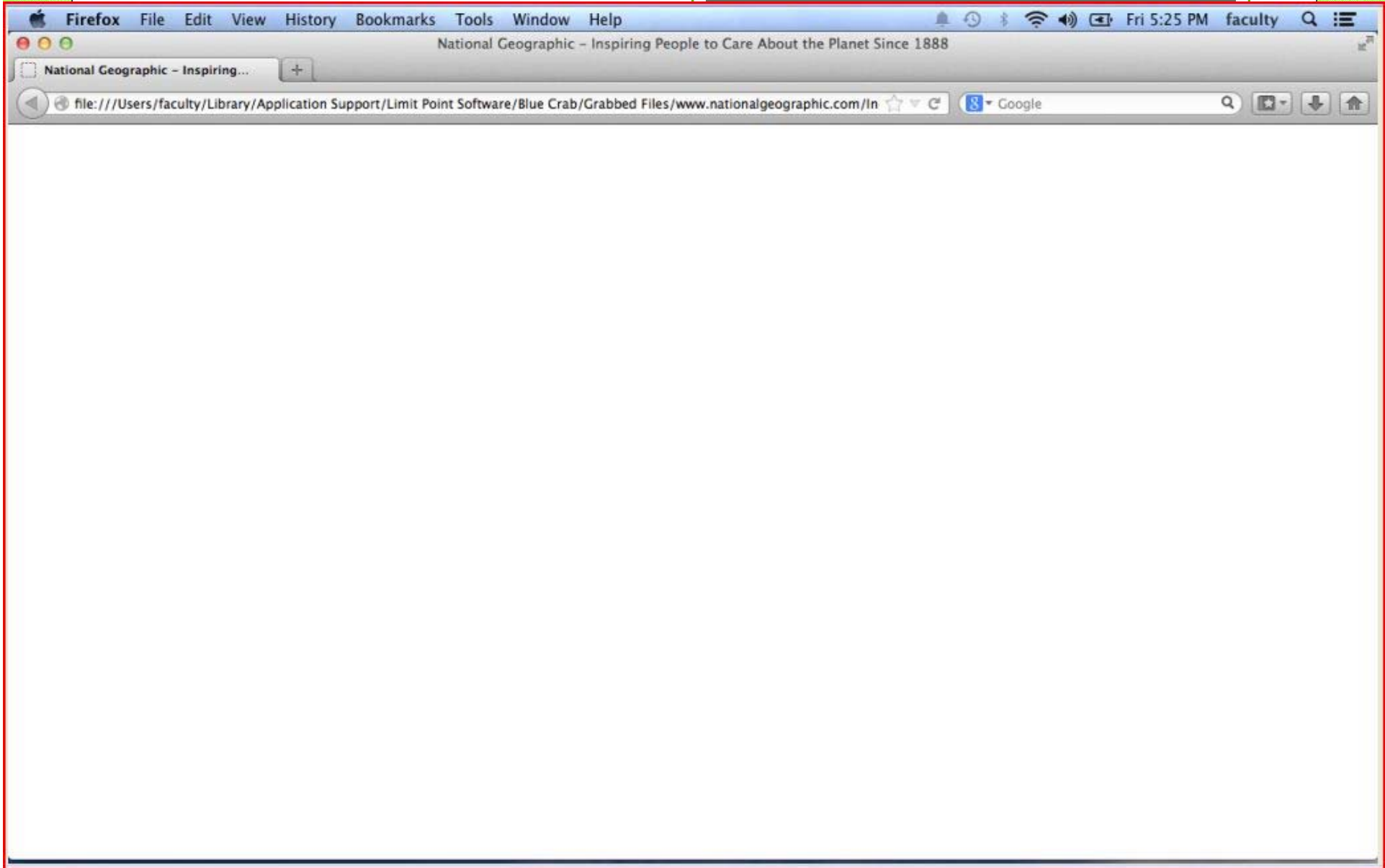
Jack is on a mission to locate the target before it's too late and the world is changed forever

Featured Videos



RED BAND SOCIETY Covers 'Be Okay' Music Video

Full Schedule




 **BlueCrab quit unexpectedly.**
Click Reopen to open the application again. Click Report to see more detailed information and send a report to Apple.



- Screen Shot 2014-...1:55 PM
- Screen Shot 2014-...5:51 PM
- Screen Shot 2014-...8:32 PM
- Screen Shot 2014-...5:57 PM
- Screen Shot 2014-...4:42 PM
- Screen Shot 2014-...6:15 PM
- Screen Shot 2014-...7:53 PM
- Screen Shot 2014-...2:37 PM



2-01 Just Dance.m4a

 HTTrack menu.docx added
"HTTrack menu.docx" was added to your Dropbox (click to view). ✕

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 publically accessible.”
- <http://webarchives.cdlib.org/about>



Web Archiving Service

Collect, manage & preserve websites

[>> CONTACT US](#) [>> LOGIN](#)

- About
- Learning Center
- FAQ
- Features
- Institutions
- Archives

SIGN UP NOW FOR A FREE TRIAL

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](#)

Name

Organization

Email Address

[▶ SUBMIT](#)

WHAT OUR USERS ARE SAYING



Explore Recent Archives

[SEE ALL ▶](#)



MAYOR OF LOS ANGELES



UNIVERSITY OF MICHIGAN



CALIFORNIA COASTAL COMMISSION



NYU LIBRARIES

Recent News & Events

WAS Service Update: June 2014

WAS Service Update: March 2014

[SEE MORE NEWS ▶](#)

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.

California Digital Library - WAS

- One of the most useful aspects of this format was the header
- There is sometimes a category included like:



Title: MalcolmX.com
Archival URL: <http://webarchives.cdlib.org/sw1vm4455/http://www.malcolmx.com/index.html>
Date captured: 08/10/12 07:45 PM
[About this archive](#)

[show document](#) [show details](#) [help](#)

This document is an archived copy for study and research. The original may be available at <http://www.malcolmx.com/index.html>



Home

ABOUT MALCOLM X

The Official Web Site of Malcolm X has everything you want to know about this historical figure. Read his biography and read inspirational quotes from this talented speaker. Browse the photo gallery for pictures of Malcolm X throughout his life! [Click here for more!](#)

COMMUNITY

The Community section is a great place to download FREE desktop wallpapers and screen savers. Visit other Web sites dedicated to Malcolm X or submit a link of your own. [Click here now!](#)

BUSINESS

CMG Worldwide is the exclusive business representative for the Estate of Malcolm X. We work with companies around the world who wish to use the name or likeness of Malcolm X in any commercial fashion. [Click here for more information.](#)

SHOPPING

Visit the Official Store for books and video! Start your own Malcolm X collection today. [Start shopping now!](#)

HOME

ABOUT

- NEWS
- BIOGRAPHY
- CHRONOLOGY
- ACHIEVEMENTS
- PHOTOS
- QUOTES
- FAST FACTS
- EULOGY

COMMUNITY

- DOWNLOADS
- LINKS

BUSINESS

- INQUIRIES

SHOPPING

- STORE
- POSTERS
- BOOKS & VIDEOS
- AUCTIONS

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 Sites 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](#)

Federal Regional Agencies in California Web Archive

Feminism and Women's Movements Web Archive

Grateful Dead Web Archive

Guantanamo Bay Detention Camp & War Crimes (U.S.) Web Archive

Housing Web Archive

International Development Organizations Web Archive

Internet/Cyberspace Democracy

Jewish American Progressive & Left Activity Web Archive

Labor Unions and Organizations (U.S.) Web Archive

Left Academia and Theory, Intellectuals and Other Notables

LGBT Rights

Los Angeles Local Government Web Archive

Mexican Presidential Election 2012 web archive

MHC Web Archives

Michigan Historical Collections Web Archives

Middle East Political Sites Project

Monterey Bay Area Local Government Web Archive

Myanmar Cyclone 2008 Web Archive

Northwestern University Web Archive

Notable Individuals

Occupy Web Archive

Orange County Government Information Web Archive

Other Left Activism Web Archive

Bentley Historical Library Web Archives

<http://bentley.umich.edu> bhlwebarchive@umich.edu (734)764-3482 | 150 Beal Ave. Ann Arbor, MI 48109

Michigan Historical Collections Web Archives

Bentley Historical Library, University of Michigan

[Home](#) [About](#) [Site List](#) [Search](#) [Help](#) [Contact Us](#)

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 activist organizations and institutions around the S...

[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.

Refine site list

lookup by site name

Go

Clear

Site list by topic:

Commerce and industry
Ethnic communities
Lesbian, Gay, Bisexual, and
Transgender Community
Natural resources
Politics and public policy
Religion
Social justice
Women in Michigan

[Automation Alley Web Archives](#)
[Bay Mills Chippewa Indian Community Web Archives](#)
[Benton Spirit Community Newspaper Web Archives](#)
[Black Autonomy Network Community Organization \(BANCO\) Web Archives](#)
[Boyne USA Resorts Web Archives](#)
[Bridge Magazine Web Archives](#)
[Carl Levin - United States Senator for Michigan Web Archives](#)
[Center for Automotive Research \(CAR\) Web Archives](#)
[Center for Michigan Web Archives](#)
[Center for Military Readiness Web Archives](#)
[Chaldean American Ladies of Charity \(CALC\) Web Archives](#)
[Chippewa Ottawa Resource Authority Web Archives](#)
[Chrysler LLC Bankruptcy Collection Web Archives](#)
[Citizens Research Council of Michigan Web Archives](#)
[Cityscape Detroit Web Archives](#)
[Clonlara School Web Archives](#)
[Council of Michigan Foundations Web Archives](#)
[Council on American-Islamic Relations Michigan \(CAIR Michigan\) Web Archives](#)
[Covey's Corner by Craig Covey Web Archives](#)
[Dan Pliskow Jazz Archives](#)
[Debbie Stabenow Senator for Michigan Web Archives](#)
[Detroit Association of Black Organizations \(DABO\) Web Archives](#)
[Detroit Cant Wait Web Archives](#)
[Detroit Digital Justice Coalition Web Archives](#)
[Detroit Tree of Heaven Woodshop Web Archives](#)

« Previous 1 2 3 4 5 6 7 8 Next »

<http://bentley.umich.edu> bhlwebarchive@umich.edu (734)764-3482 | 150 Beal Ave. Ann Arbor, MI 48109

Michigan Historical Collections Web Archives

Bentley Historical Library, University of Michigan

[Home](#) [About](#) [Site List](#) [Search](#) [Help](#) [Contact Us](#)

Detroit Digital Justice Coalition Web Archives

Latest Starting URL

<http://detroitdjc.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



web archives
yesterday's web; today's archives

Occupy Web Archive

UCLA Library

[Home](#) [About](#) [Site List](#) [Search](#) [Help](#) [Contact Us](#)

Occupy Los Angeles

Latest Starting URL

<http://obrag.org/> (live link)

← this is NOT the Occupy Los Angeles website

Description

Occupy Los Angeles

Captured

- 03/18/12 09:47 PM
- 03/20/12 10:39 PM
- 04/04/12 12:48 AM
- 05/02/12 09:15 PM
- 05/25/12 05:26 PM
- 07/02/12 07:46 PM
- 07/27/12 01:53 PM
- 08/27/12 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
- 03/27/13 01:53 PM
- 04/27/13 01:53 PM
- 05/28/13 08:07 PM
- 06/27/13 01:53 PM
- 07/27/13 01:53 PM
- 08/27/13 01:53 PM
- 09/27/13 01:53 PM
- 10/27/13 01:53 PM
- 11/27/13 01:53 PM
- 12/27/13 01:53 PM



Ocean Beach California 92107
Serving OB, the Peninsula and San Diego Beaches

Say **Keep it goin'** to the OB Rag with your donation today!



Secure Donations through PayPal

HOME ABOUT CONTACT 1ST RAG LINKS EVENTS COLUMNS OCEAN BEACH POPULAR CLASSIFIEDS SD FREE PRESS

3 Unit Apartment on 4900 Block of West Point Loma Sold for Close to \$1 Million

by FRANK GORMLIE on JULY 18, 2014 0 COMMENTS
in 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



OB NOODLE HOUSE Bar 1502

4993 Niagara Avenue
(between Cable and Bacon)
619-255-9858

Mention This Ad For \$1 OFF Your Beer!

Mention This Ad For 5% Off!

Bone Appétit!
Food • Treats • Toys • Provisions
Ocean Beach Pet Supply

California Digital Library - WAS

- 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.

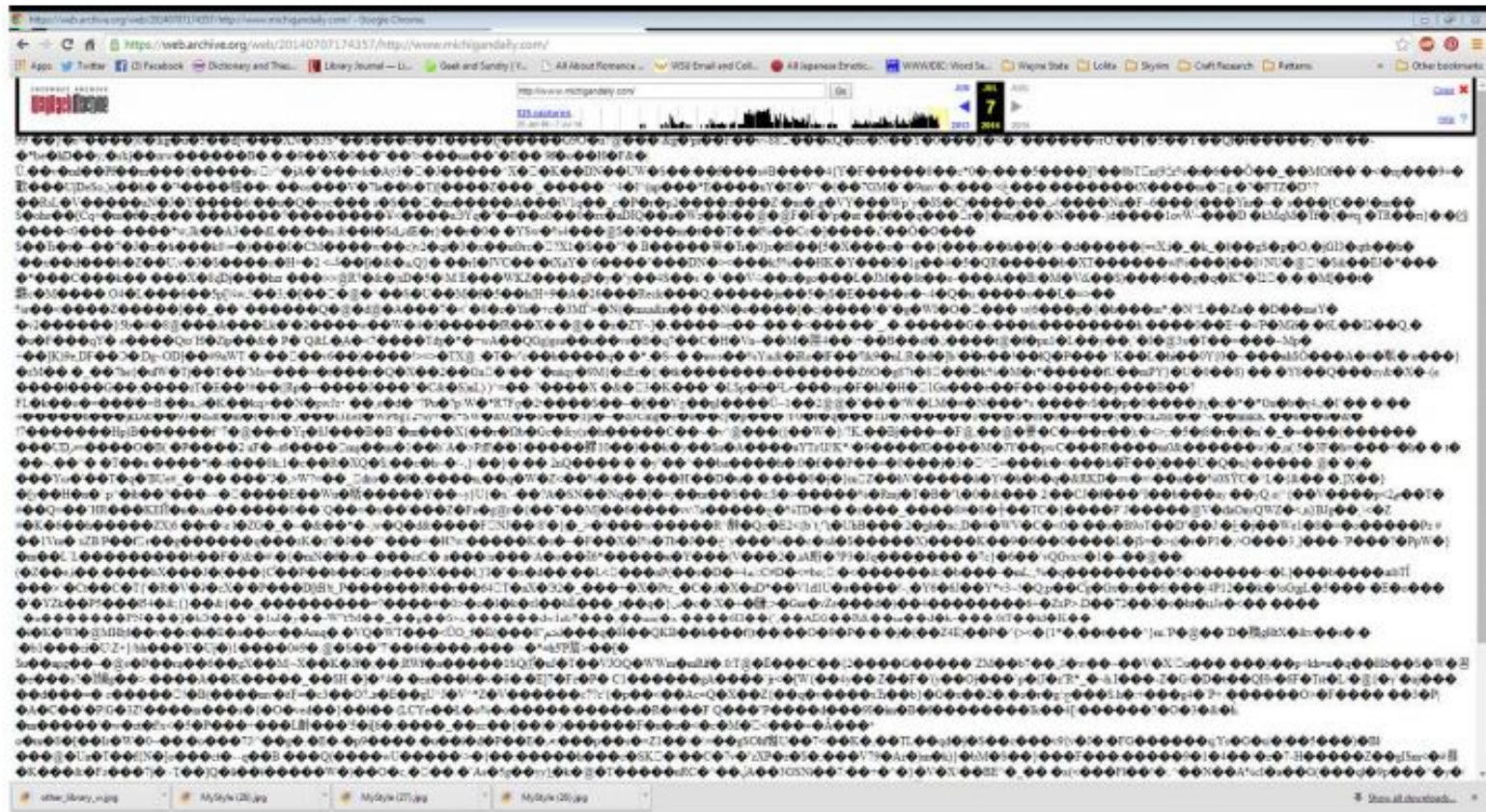
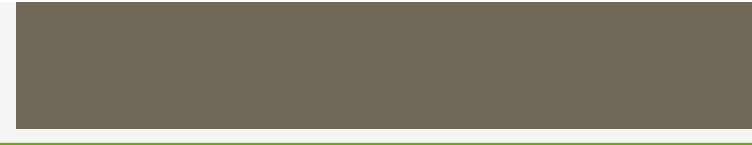
Internet Archive – Wayback Machine

- 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.



Internet Archive – Wayback Machine

- 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.

	Completeness Average	High Capture (80% plus)	Note
All Captures Evaluated	61%	58%	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.
High Average	77%		
Low Average	22%		
High Average		74%	
Low Average		42%	
Completeness Average	Defined as percentage of Homepage Content Captured divided by all sites evaluated.		
High Capture	Defined as percentage of sites that captured content at 80% or higher as assessed by our team.		

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.

Web Archiving Lessons Learned –

Coca Cola Archives, Jamal Booker

- “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.”

Disclaimers

- 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.
- <https://archive.org/search.php?query=%28collection%3Aarchiveteam-fire%20OR%20mediatype%3Aarchiveteam-fire%29%20AND%20-mediatype%3Acollection&sort=-reviewdate>

Archive Team in IA

The screenshot shows a web browser window with the Internet Archive search results page. The search query is `collection:archiveteam-fire OR mediatype:archiveteam-fire AND -mediatype:collection`. The page displays five search results, each with a title, description, keywords, and download count. On the right side, there are three panels: 'Advanced search', 'Sort results by', and 'Refine your search'.

Search Results
Results: 1 through 50 of 1,124 (5.493 secs)
You searched for: (collection:archiveteam-fire OR mediatype:archiveteam-fire) AND -mediatype:collection

[1] 2 3 4 5 6 7 8 9 10 11 Next Last

1 [Usenet Archive of UTZOO Tapes](#)
This is a **collection** of .TGZ files of very early USENET posted data provided by a number of driven and brave individuals: Lance Bailey, Bruce Jones, Bob Webber, Brewster Kahle, and Sue Thielen. This mirror is, unfortunately, incomplete in some aspects that are not currently available generally. We welcome contributions of those missing files...
Keywords: [usenet](#); [utzoo](#)
Downloads: 1,347 ★★★★★ (2 reviews)

2 [www.theregister.co.uk/2010_201210_panic_download](#) - theregister.co.uk
This is panic download of all [www.theregister.co.uk/2010](#) urls as of 2012-10-07. This has a wget.log and a cdx and a list of urls.
Keywords: [theregister.co.uk](#); [theregister](#); [archiveteam](#)
Downloads: 194 (1 review)

3 [torrentfreak.com 20121002_panic_download](#) - torrentfreak.com
This is a panic download of [torrentfreak.com](#) as of 2012-10-02. This is just the warc.gz and cdx files. I also included a list of urls.
Keywords: [torrentfreak.com](#); [torrentfreak](#); [archiveteam](#)
Downloads: 279 ★★★★★ (1 review)

4 [Archive Team: The Proust.Com Panic Download](#)
PROUST.COM, a **collection** of memories and records of human lives, announced it was closing down and giving up its domain. The site has since announced it will likely live on through "anonymous" benefactors, but the Archive Team has this **collection** of public-facing Proust accounts to keep for history's sake.
Downloads: 24 ★★★★★ (1 review)

5 [occupywallst.org 20120822_panic_download](#) - occupywallst.org
This is a panic download of [occupywallst.org](#) as of 2012-08-22. I grabbed all images that are linked on occupywallst.org.

Advanced search

Sort results by:
[Relevance](#)
[Average rating](#)
[Download count](#)
[Date](#)
[Date added](#)

Group results by:
 Relevance
 Mediatype
 Collection

Refine your search:

Collection
[archiveteam-fire](#)
[archiveteam](#)
[web](#)
[usenet](#)
[data](#)

Creator
[theregister.co.uk](#)
[torrentfreak.com](#)
[occupywallst.org](#)
[www.theguardian.com](#)
[g4tv.com](#)

Disclaimers

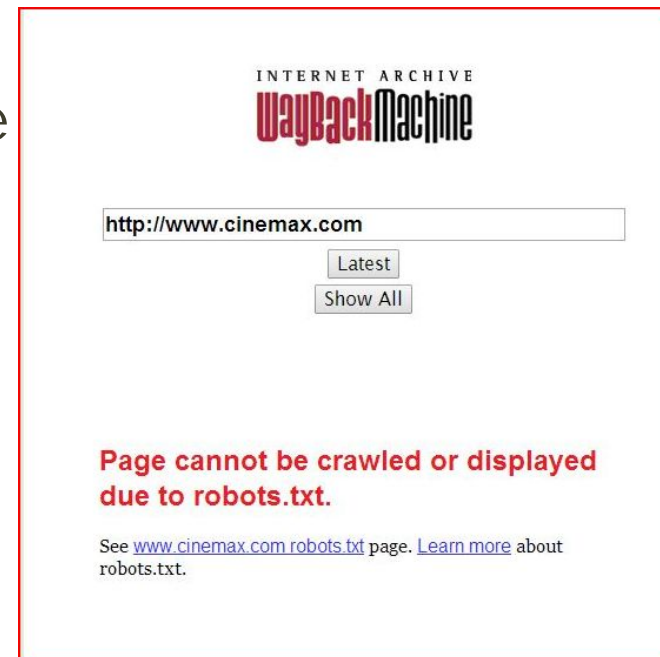
- Only looked at Homepages, would like to get into greater depth to assure content capture is accurate overall.
 - This will take processing power and time as well as greater storage.
 - We were all learning so this is just a first step.

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 capture this site, IA and HTTrack did not.



Additional Sources

- <http://ws-dl.blogspot.com/2014/07/2014-07-14-archival-acid-test.html>
- <https://www.google.com/search?q=web+archiving+basics&oq=web+archiving+basics&aqs=chrome..69i57.6455j0j1&sourceid=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/>
ag1797@wayne.edu
[313 577-9783](tel:3135779783)