Building Archivable Websites

Nicholas Taylor
Web Archiving Service Manager
Digital Library Systems and Services

Drupal Camp
April 19, 2014
Why Build ARCHIVABLE WEBSITES?
future users are users, too
maintain web usability
improve temporal web usability
improve temporal web usability

Internet Archive: "Wayback Machine"

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.
recover your lost website
refer to earlier website versions
institutional history
websites are cultural artifacts

The World Wide Web project

WORLD WIDE WEB

The WorldWideWeb (W3) is a wide-area hypermedia[1] information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary[2] of the project, Mailing lists[3], Policy[4], November’s W3 news[5], Frequently Asked Questions[6].

What’s out there?[7] Pointers to the world’s online information, subjects[8], W3 servers[9], etc.

Help[10] on the browser you are using

Software

Products[11] A list of W3 project components and their current state. (e.g. Line Mode[12], X11 Viola[13], NeXTStep[14], Servers[15], Tools[16], Mail robot[17], Library[18])

Technical[19] Details of protocols, formats, program internals etc

<ref.number>, Back, <RETURN> for more, or Help:  

“The World Wide Web project”
facilitate compliance
optimize for other crawlers
How to IMPROVE ARCHIVABILITY

“metal web” by paul-74 under CC BY-NC-SA 2.0
follow web standards and accessibility guidelines

“The nice thing about web standards is, there are so many to choose from.”

“Web Standards Fortune Cookie” by Matt Herzberger under CC BY-SA 2.0
use a site map, transparent links, and contiguous navigation
maintain stable URLs and redirect when necessary
use semantically-meaningful URLs
be careful w/ robot exclusion rules

```
Disallow: /scripts/
Disallow: /themes/
```

"drupal/robots.txt at 7.x"
minimize reliance on external assets necessary for presentation
minimize reliance on external assets necessary for presentation
serve reusable assets from a single, common location
specify HTTP response headers for caching and content encoding
embed metadata, especially character encoding
use durable data formats
prefer responsive design over user-agent personalization
examine your site in the Internet Archive Wayback Machine

Greetings & Welcome to
A Multidimensional Perception ~/\=
& PCGuru

Last Updated December 5, 1996
Best viewed with either

Netscape Navigator or Internet Explorer

We are constantly growing in this cyberdimensional reality.
The following list of skills & talents are currently available.

- Audio Ambient/Soundboard Mastering.
- Video Producing/Directing/Editing.
- 35mm Still Photography.
- PC Consulting/Training/Upgrade/Repair/Installation.
- HTML3 Home Page Programming & Designing.
- Internet Consulting/Training.
- Technical Advising.
Web Archiving

TOOLS AND SERVICES

“giant mechanical spider & crowd” by mjtmail (tiggy) under CC BY 2.0
Heritrix

---

<table>
<thead>
<tr>
<th>Jobs</th>
<th>Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Running: 150-sites-declining-10MB</td>
<td>102217 KB used</td>
</tr>
<tr>
<td>0 pending, 9 completed</td>
<td>117056 KB current heap</td>
</tr>
<tr>
<td>Alerts: 0 (0 new)</td>
<td>236224 KB max heap</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Job Status: RUNNING</th>
<th>Load</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rates</td>
<td>15 active of 50 threads</td>
</tr>
<tr>
<td>25.5 URLs/sec (32.27 avg)</td>
<td>1.0 congestion ratio</td>
</tr>
<tr>
<td>555 KB/sec (545 avg)</td>
<td>399 deepest queue</td>
</tr>
<tr>
<td>Time</td>
<td>79 average depth</td>
</tr>
<tr>
<td>4m43s elapsed</td>
<td></td>
</tr>
<tr>
<td>3m34s remaining (estimated)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Totals</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>56%</td>
<td>6874 queued</td>
</tr>
<tr>
<td>downloaded 9108</td>
<td>15997 total downloaded and queued</td>
</tr>
<tr>
<td>151 MB uncompressed data received</td>
<td></td>
</tr>
</tbody>
</table>
Wget

adam@a:~$ wget https://en.wikipedia.org/wiki/Wget
Resolving en.wikipedia.org (en.wikipedia.org)... 208.80.154.224, 2620:0:861:ed1a::1
Connecting to en.wikipedia.org (en.wikipedia.org)|208.80.154.224|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: unspecified [text/html]
Saving to: `Wget'

[ <=> ] 91,646 583K/s in 0.2s
2014-02-09 07:38:43 (583 KB/s) - `Wget' saved [91646]
adam@a:~$
HTTrack

Welcome

HTTrack is a free (GPL, libre/free software) and easy-to-use offline browser utility.

It allows you to download a World Wide Web site from the internet to a local directory, building recursively all directories, getting HTML, images, and other files from the server to your computer. HTTrack arranges the original site's relative link-structure. Simply open a page of the "mirrored" website in your browser, and you can browse the site from link to link, as if you were viewing it online. HTTrack can also update an existing mirrored site, and resume interrupted downloads. HTTrack is fully configurable, and has an integrated help system.

WinHTTrack is the Windows 2000/XP/Vista/Seven release of HTTrack, and WebHTTrack the Linux/Unix/BSD release. See the download page.
Wayback

INTERNET ARCHIVE
Wayback Machine

407 Billion web pages saved over time.

Tools
- Wayback Machine Availability API
- Build your own tools.
- WordPress Broken Link Checker
- Banish broken links from your blog.
- 404 Handler for Webmasters
- Help users get where they were going.

Subscription Service
- Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit Archive-It to build and browse the collections.

Save Page Now
- Capture a web page as it appears now for use as a trusted citation in the future. Only available for sites that allow crawlers.

"Internet Archive Wayback Machine"
Web Archiving Integration Layer

What Is It?

Web Archiving Integration Layer (WAIL) is a graphical user interface (GUI) atop multiple web archiving tools intended to be used as an easy way for anyone to preserve and replay web pages.

Tools included and accessible through the GUI are Heritrix 3.1.2, Wayback 1.7, and ware-proxy. Support packages include Apache Tomcat, phantomjs, and pyinstaller.

WAIL is written mostly in Python and a small amount of JavaScript.

Screenshots
Memento
assess archivability w/ Archive Ready

Is your website Archive Ready?

What is ArchiveReady?
An online tool which evaluates if your website will be archived correctly by web archives, such as the Internet Archive.

Why bother?
Because not all websites are archive ready and this may result in invalid web archives and ultimately in information loss.

Who is it for?
Web professionals who need to check if their websites are archive ready. Web archive engineers who need to evaluate target websites before harvesting and ingestion.

How does it work?
ArchiveReady analyses your website (i.e. HTML, Images, CSS, JS, Sitemaps) and calculates a set of Archivability Facets: Accessibility, Cohesion, Metadata, Performance & Standards Compliance.

How much does it cost?
ArchiveReady.com is completely FREE for personal use. Large scale use of the ArchiveReady.com API as well as technical and scientific support is available on a fee. Please contact the author to learn more.

Do you need more information?

Service started in 2012/10/01, websites checked: 37018.
thank you!