



STANFORD UNIVERSITY LIBRARIES

Building Archivable Websites

Nicholas Taylor

Web Archiving Service Manager

[Digital Library Systems and Services](#)

[Drupal Camp](#)

April 19, 2014

Why Build

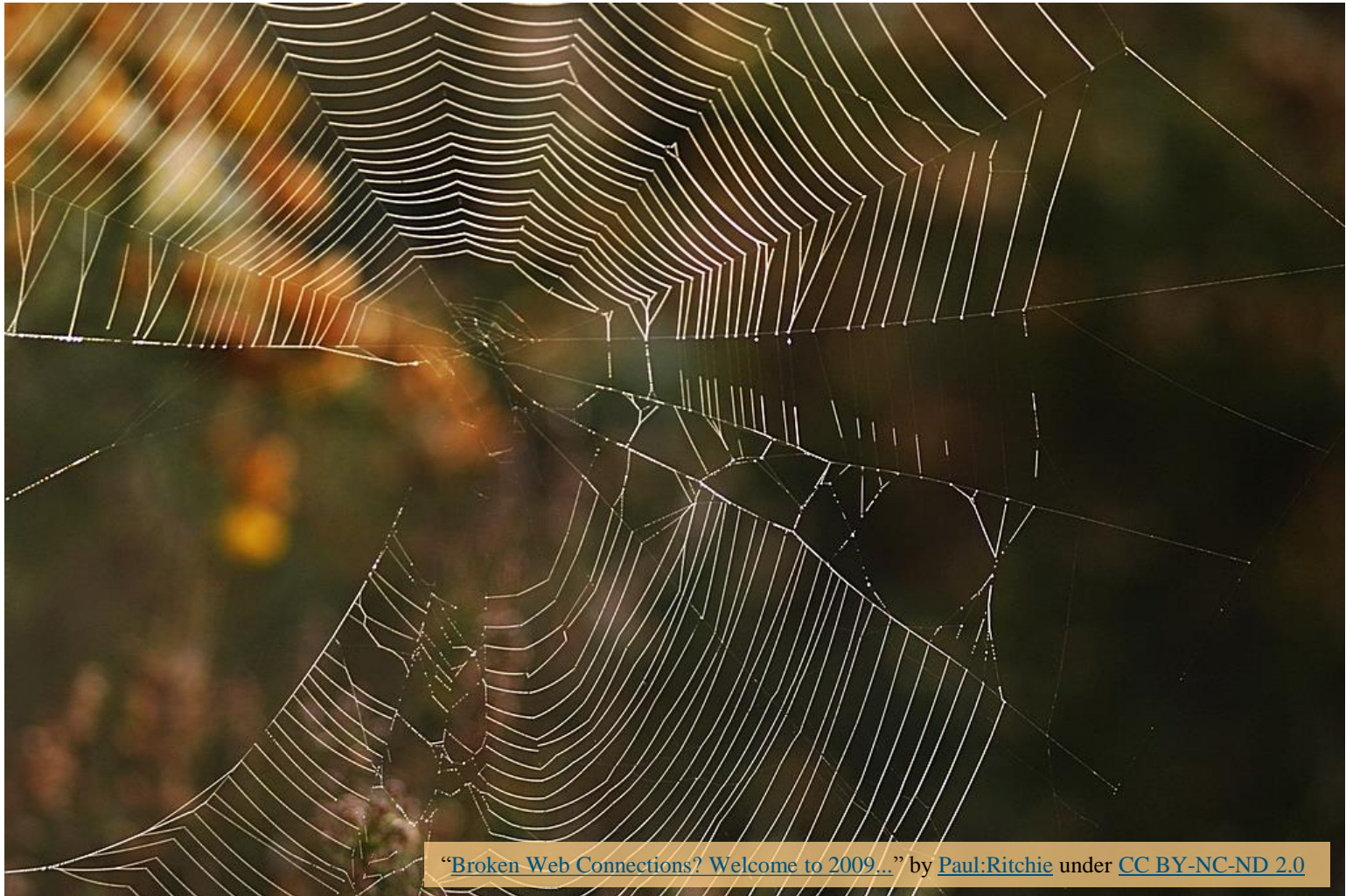
ARCHIVABLE WEBSITES?

future users are users, too



“a connection between past and future” by [Gioia De Antoniis](#) under [CC BY-NC-ND 2.0](#)

maintain web usability



improve temporal web usability



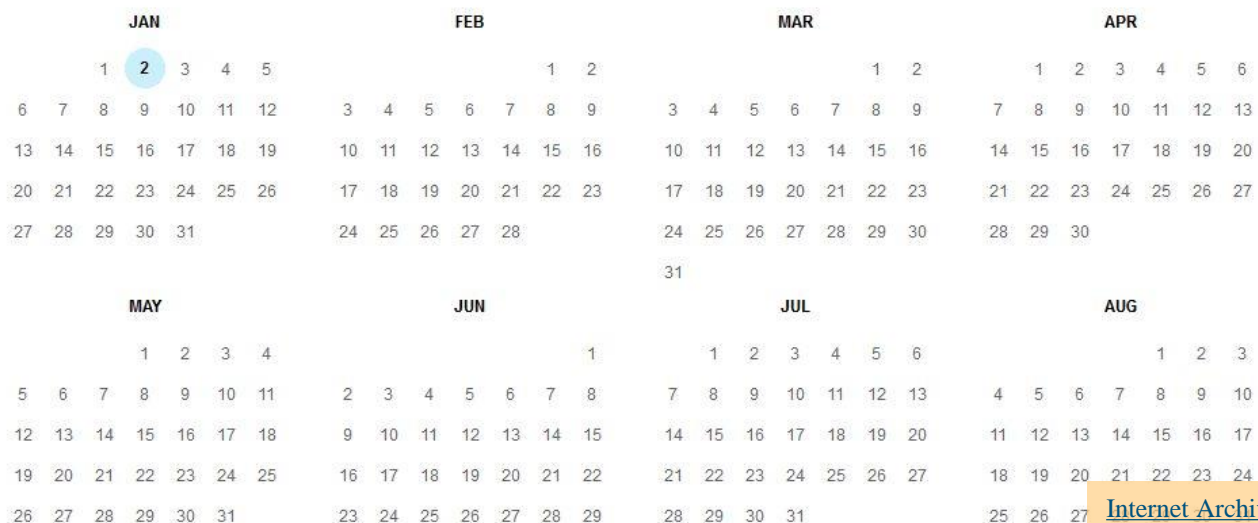
<http://bono.house.gov/>

BROWSE HISTORY

<http://bono.house.gov/>

Saved **425 times** between September 21, 2006 and January 2, 2013.

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.



Internet Archive: "Wayback Machine"

improve temporal web usability



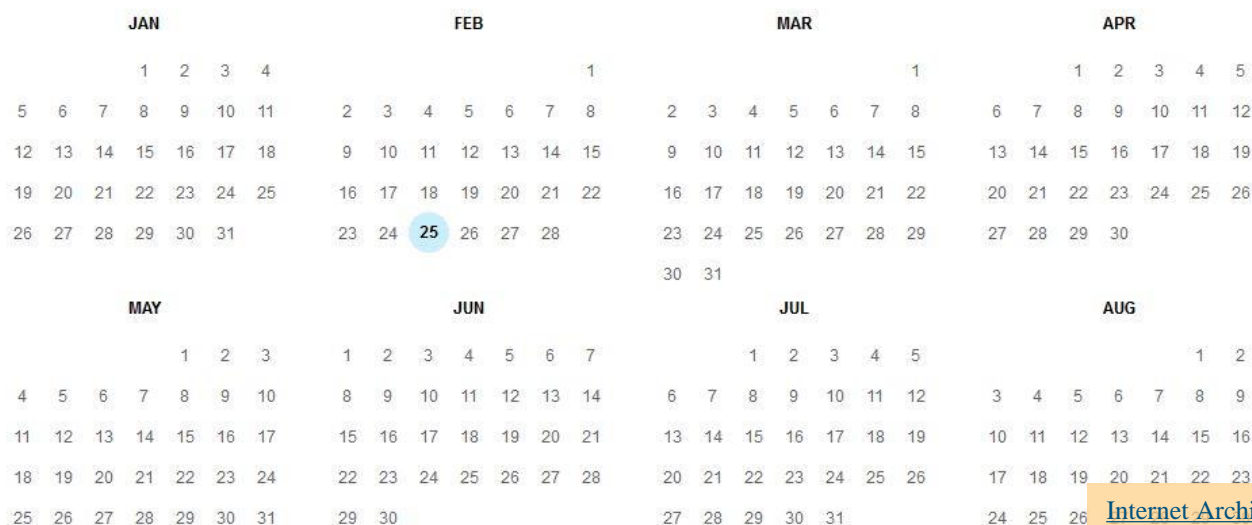
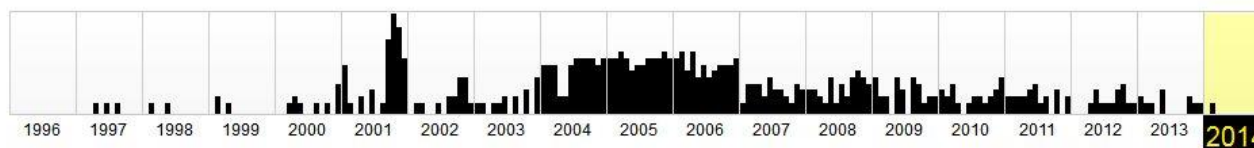
<http://house.gov/bono/>

BROWSE HISTORY

<http://house.gov/bono/>

Saved **891 times** between April 28, 1997 and February 25, 2014.

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.



Internet Archive: "Wayback Machine"

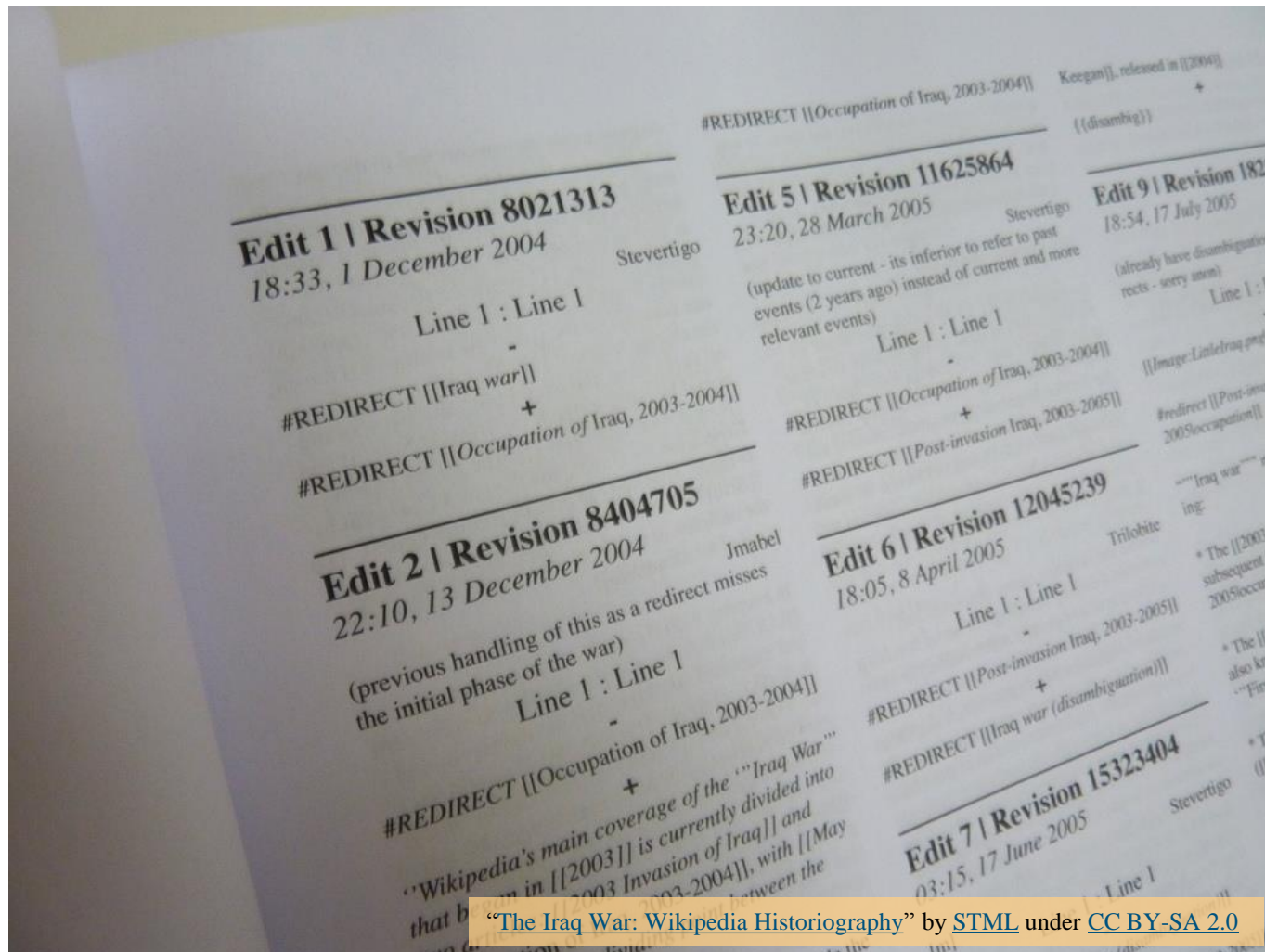
recover your lost website



Old Dominion University
Computer Science Department
Norfolk, VA

“Warrick”

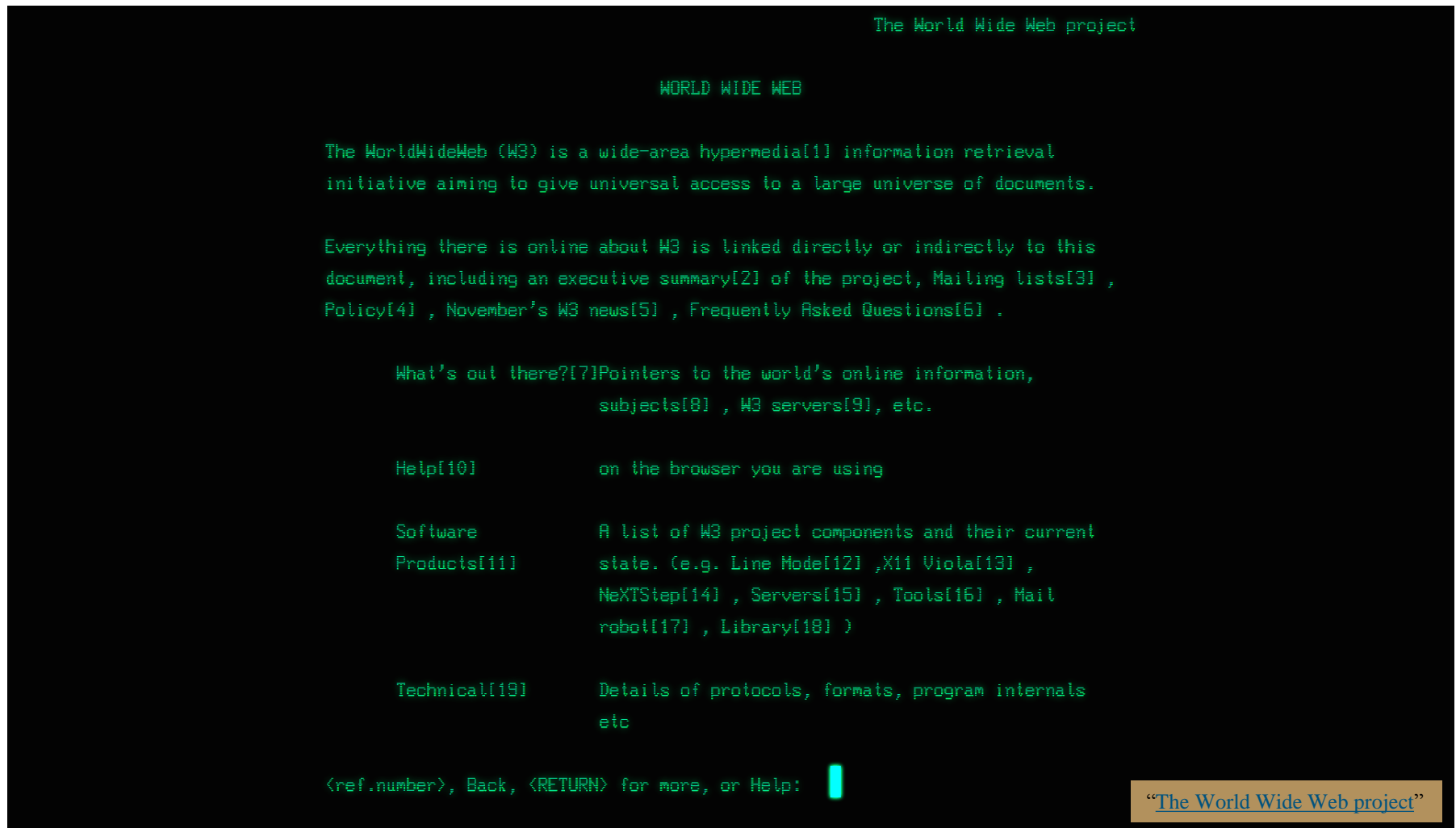
refer to earlier website versions



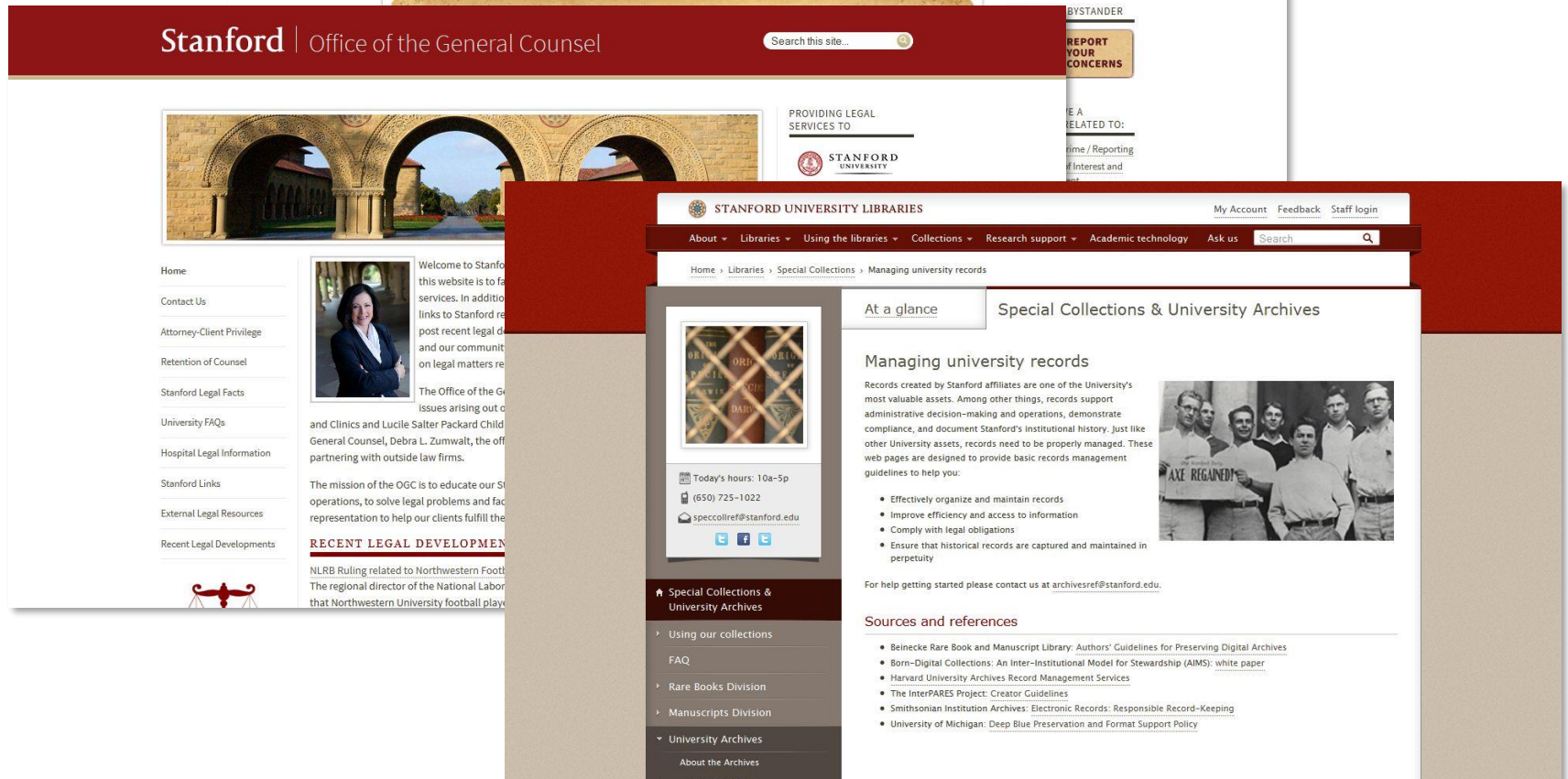
institutional history



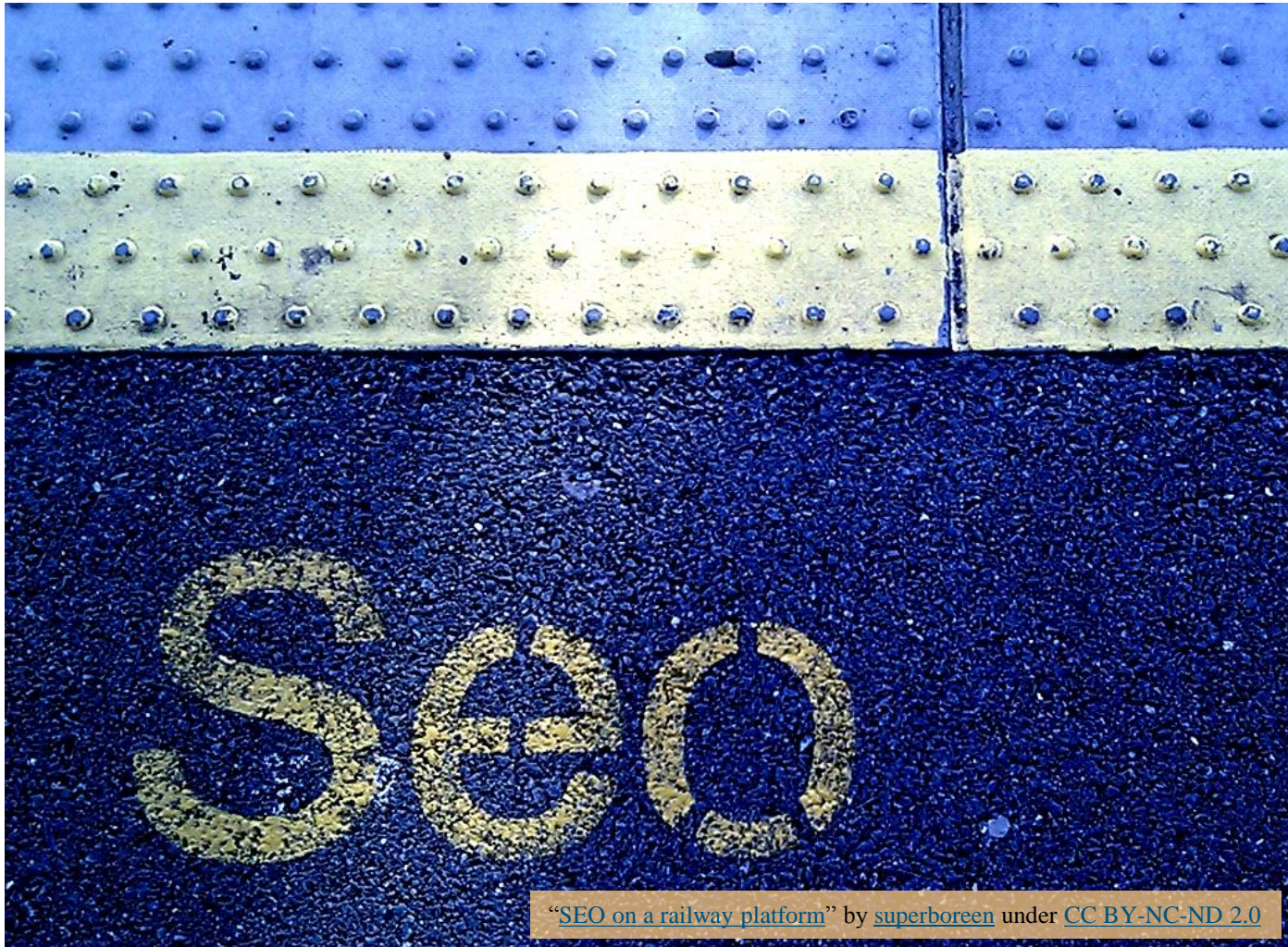
websites are cultural artifacts



facilitate compliance



optimize for other crawlers



“SEO on a railway platform” by [superboreen](#) under [CC BY-NC-ND 2.0](#)



How to

IMPROVE ARCHIVABILITY

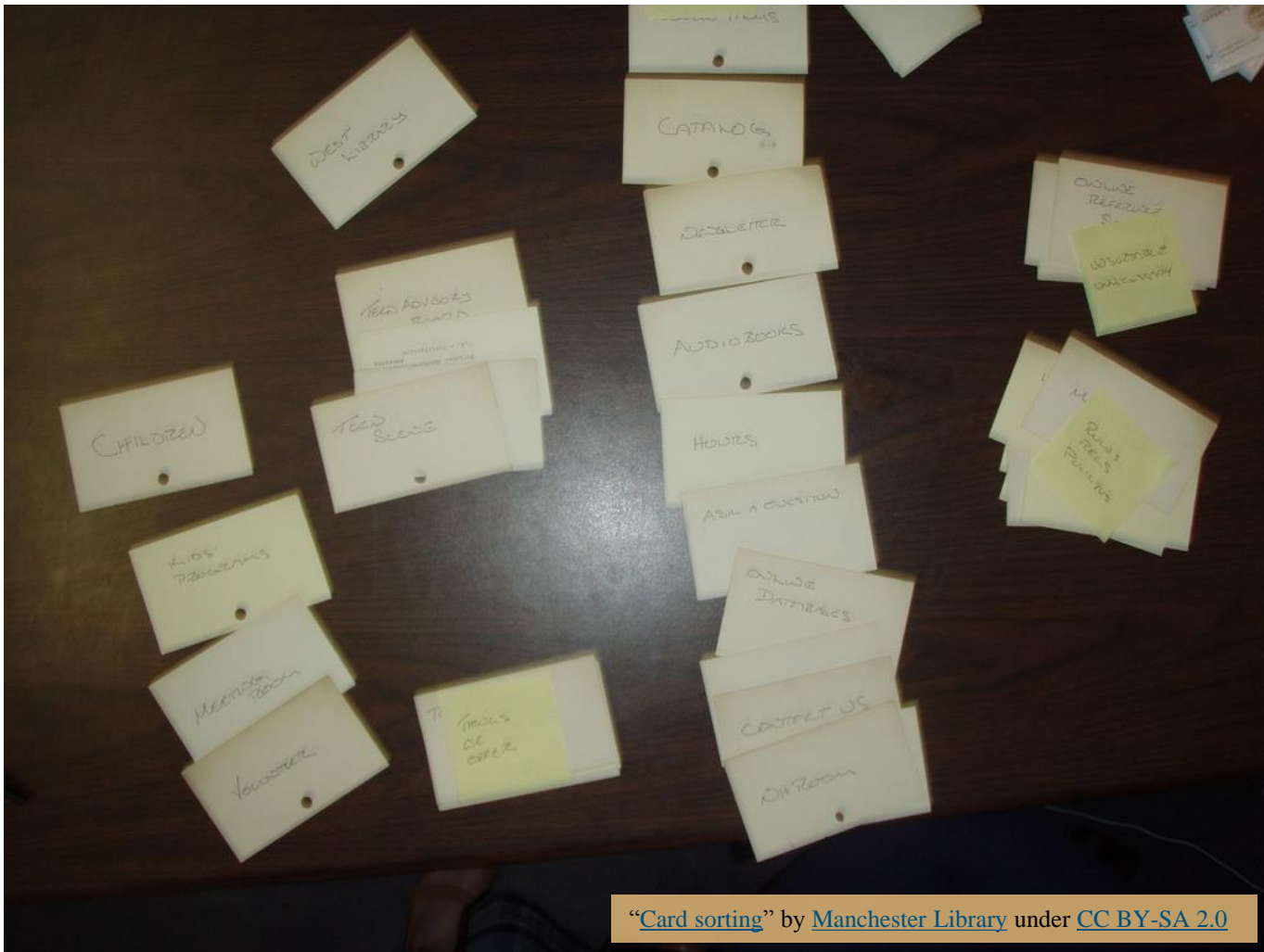
“[metal web](#)” by [paul:74](#) under [CC BY-NC-SA 2.0](#)

follow web standards and accessibility guidelines



“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



“San Francisco-Oakland Bay Bridge 1442a” by [Don Barrett](#) under [CC BY-NC-ND 2.0](#)

use semantically-meaningful URLs

The screenshot shows a web browser window with the address bar displaying `http://w63.org`. The page content includes the text "Welcome to w63.w63.org!" and a large block of text consisting of many repeated "w" characters, which is highlighted with a red box. Below the page content, the browser's developer tools are open, showing the "Net" tab. A red box highlights the "200 OK" status in the request list. Another red box highlights the "200 OK" text in the "Response Headers" section. The "Request Headers" section shows various headers, including "Accept", "Accept-Encoding", "Accept-Language", "Cache-Control", "Connection", "DNT", "Host", and "User-Agent".

200 OK

200 OK

Access-Control-Allow-Origin: *

Connection: keep-alive

Content-Encoding: gzip

Content-Type: text/html

Date: Mon, 14 Apr 2014 17:51:37 GMT

Server: nginx

Transfer-Encoding: chunked

Vary: Accept-Encoding

Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8

Accept-Encoding: gzip, deflate

Accept-Language: en-US,en;q=0.5

Cache-Control: max-age=0

Connection: keep-alive

DNT: 1

Host: w63.w63.org

User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64; rv:24.0) Gecko/20100101 Firefox/24.0

0 requests

“w63.w63.org”

be careful w/ robot exclusion rules

GitHub This repository Search or type a command Explore Features Enterprise Blog Sign up Sign in

drupal / drupal
mirrored from <http://git.drupal.org/project/drupal.git> ★ Star 1,286 🍴 Fork 465

branch: 7.x drupal / robots.txt

webchickenator 2 years ago Issue #1249132 by aspiclicious: Add INSTALL.sqlite.txt in robots.txt.
3 contributors

file | 61 lines (59 sloc) | 1.561 kb Open Edit Raw Blame History Delete

```

1 #
2 # robots.txt
3 #
4 # This file is to prevent the crawling and indexing of certain parts
5 # of your site by web crawlers and spiders run by sites like Yahoo!
6 # and Google. By telling these "robots" where not to go on your site,
7 # you save bandwidth and server resources.
8 #
9 # This file will be ignored unless it is at the root of your host:
10 # Used: http://example.com/robots.txt
11 # Ignored: http://example.com/site/robots.txt
12 #
13 # For more information about the robots.txt standard, see:
14 # http://www.robotstxt.org/wc/robots.html
15 #
16 # For syntax checking, see:
17 # http://www.sxw.org.uk/computing/robots/check.html
18
19 User-agent: *
20 Crawl-delay: 10
21 # Directories
22 Disallow: /includes/
23 Disallow: /misc/
24 Disallow: /modules/
25 Disallow: /nodefiles/
26 Disallow: /scripts/
27 Disallow: /themes/
28
29 # Files
30 Disallow: /CHANGELOG.txt
31 Disallow: /cron.php
32 Disallow: /INSTALL.mysql.txt
33 Disallow: /INSTALL.pgsql.txt
34 Disallow: /INSTALL.sqlite.txt
35 Disallow: /install.php
36 Disallow: /INSTALL.txt
37 Disallow: /LICENSE.txt
38 Disallow: /MAINTAINERS.txt
39 Disallow: /update.php
40 Disallow: /UPGRADE.txt
41 Disallow: /xmlrpc.php

```

Disallow: /scripts/
Disallow: /themes/

“drupal/robots.txt at 7.x”

minimize reliance on external assets necessary for presentation

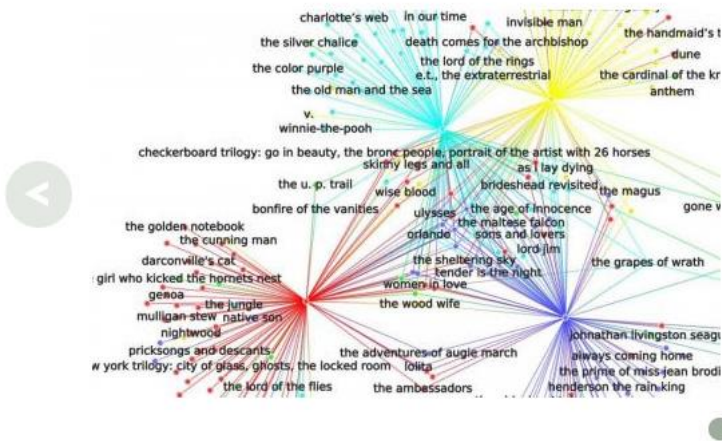
Stanford University

SUNetID Login

Stanford

Department of English

[Home](#)
[About](#)
[People](#)
[Courses](#)
[Degree Programs](#)
[Department Bookshelf](#)
[News](#)
[Events](#)



English Today

“In the long run, a people is known, not by its statements or its statistics, but by the stories it tells.” —Flannery O’Connor

Discover the power of English to expand your horizons.

Prospective student?

Recent News

APR 11 2014 | THE BOOK HAVEN

Pulitzer prize-winning Adam Johnson takes home another big prize

Upcoming Events

APR 17

Ian Watt Lecture: Michael McKeon

Affiliated Programs

Creative Writing Program

[Internet Archive Wayback Machine: “Stanford Department of English”](#)

minimize reliance on external assets necessary for presentation

Stanford University

SUNetID Login

Stanford
Department
of English

Search this site...

Home About People Courses Degree Programs Department Bookshelf News

Events



English Today

"In the long run, a people is known, not by its statements or its statistics, but by the stories it tells."
—Flannery O'Connor

Discover the power of English to expand your horizons.

Prospective student?



Recent News

MAR 25 2014 | STANFORD REPORT



Teaching tech and trees brings

Upcoming Events

APR 10

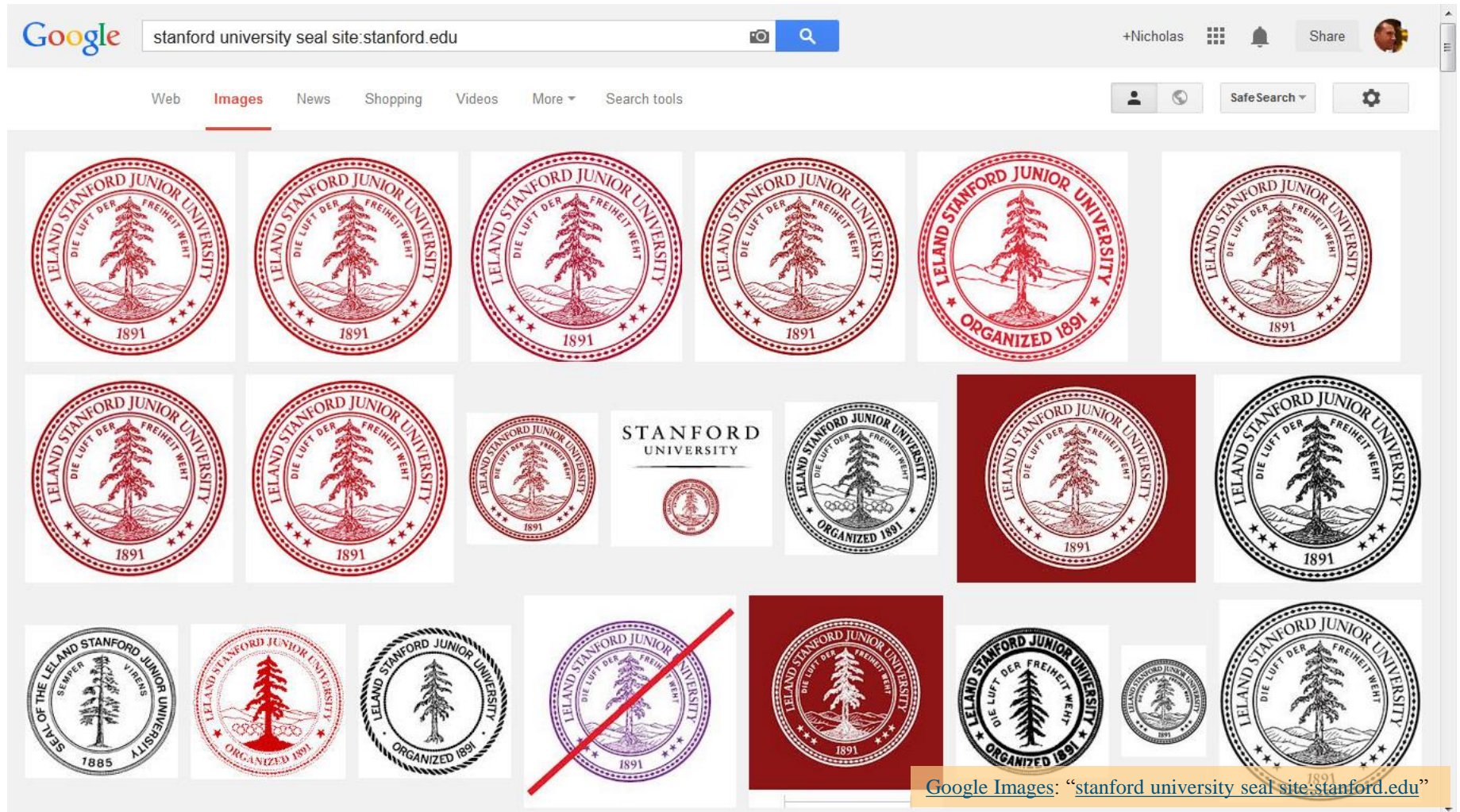
The Social Network of Benjamin Franklin, Printer

Affiliated Programs

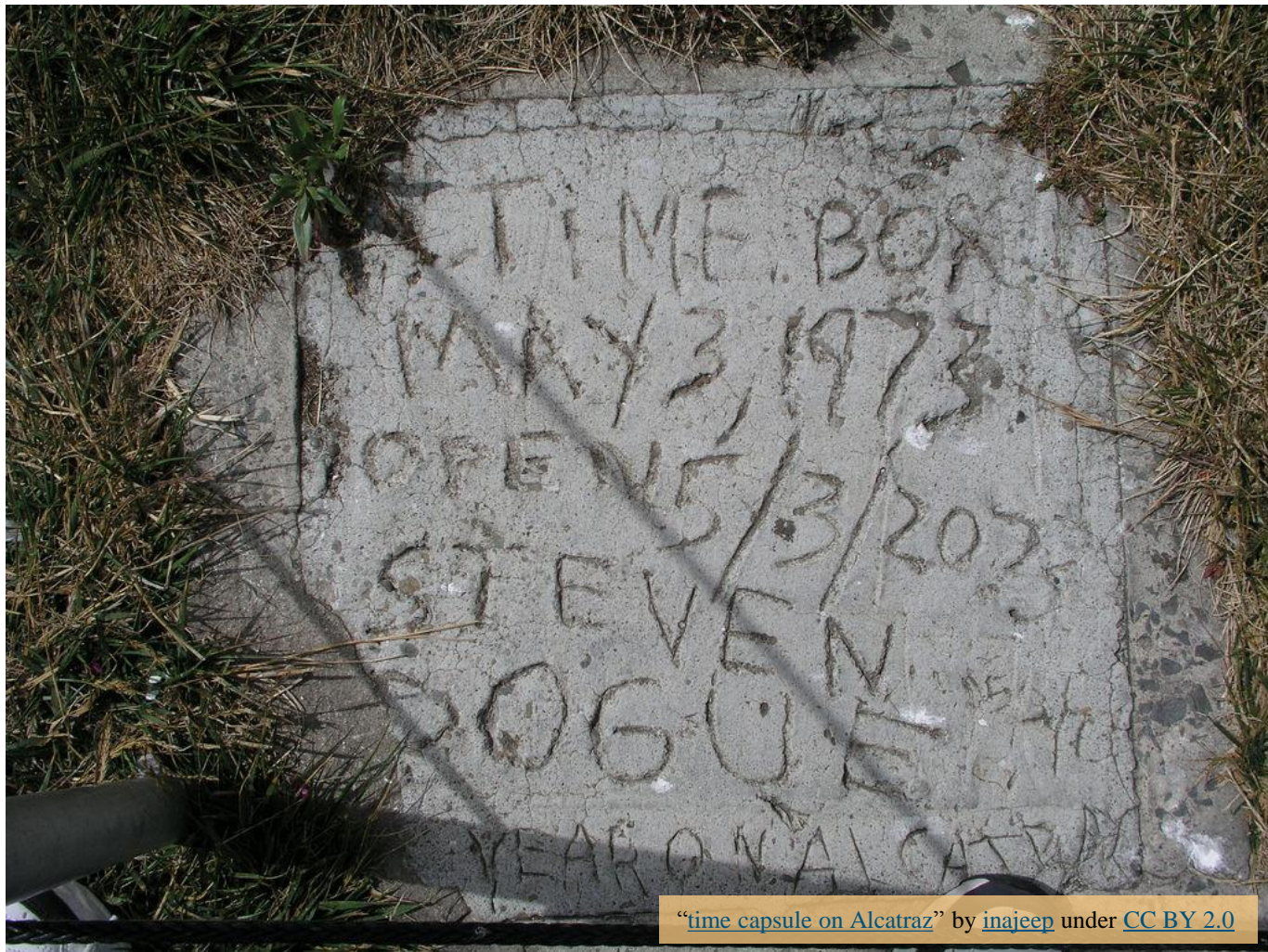
Creative Writing Program

"Stanford Department of English"

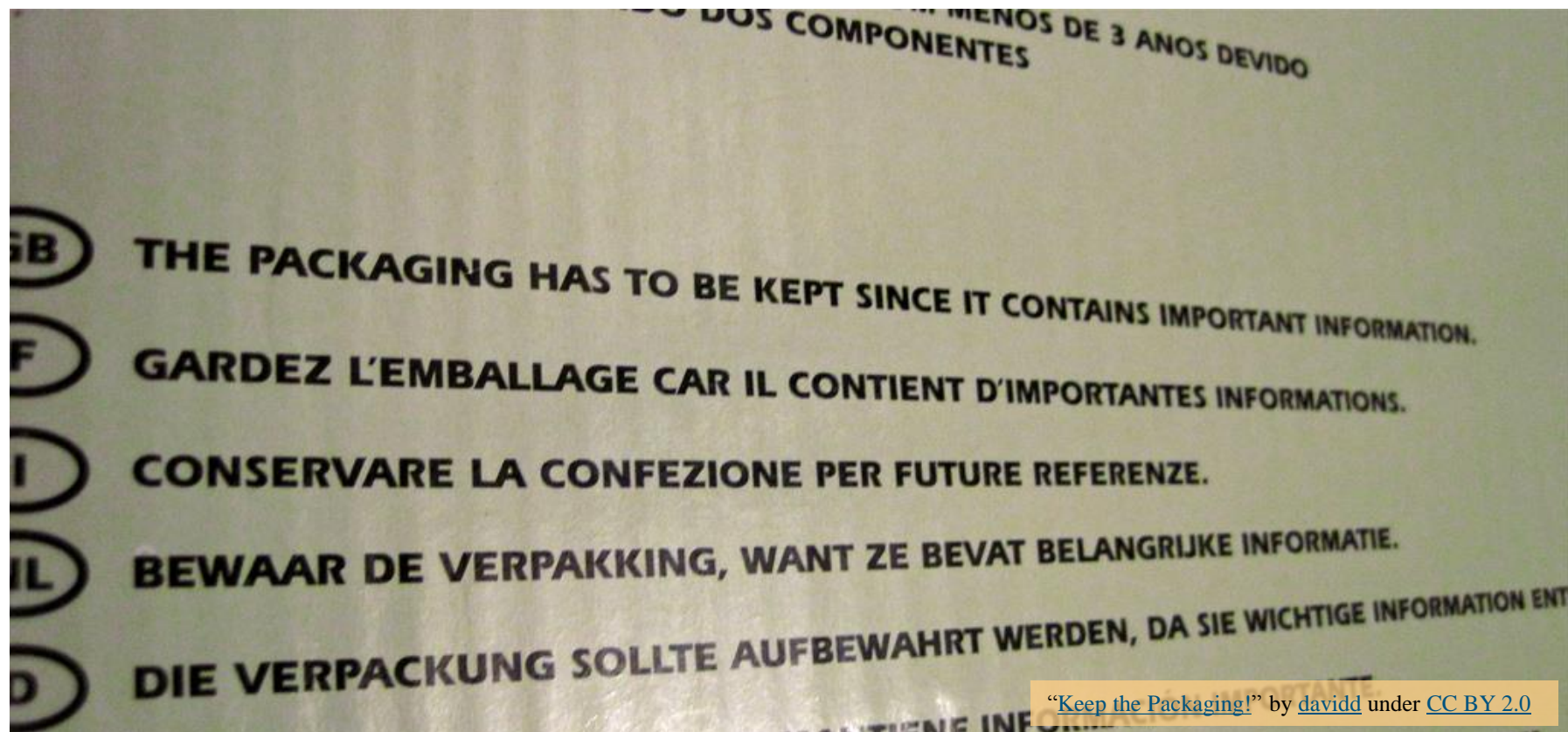
serve reusable assets from a single, common location



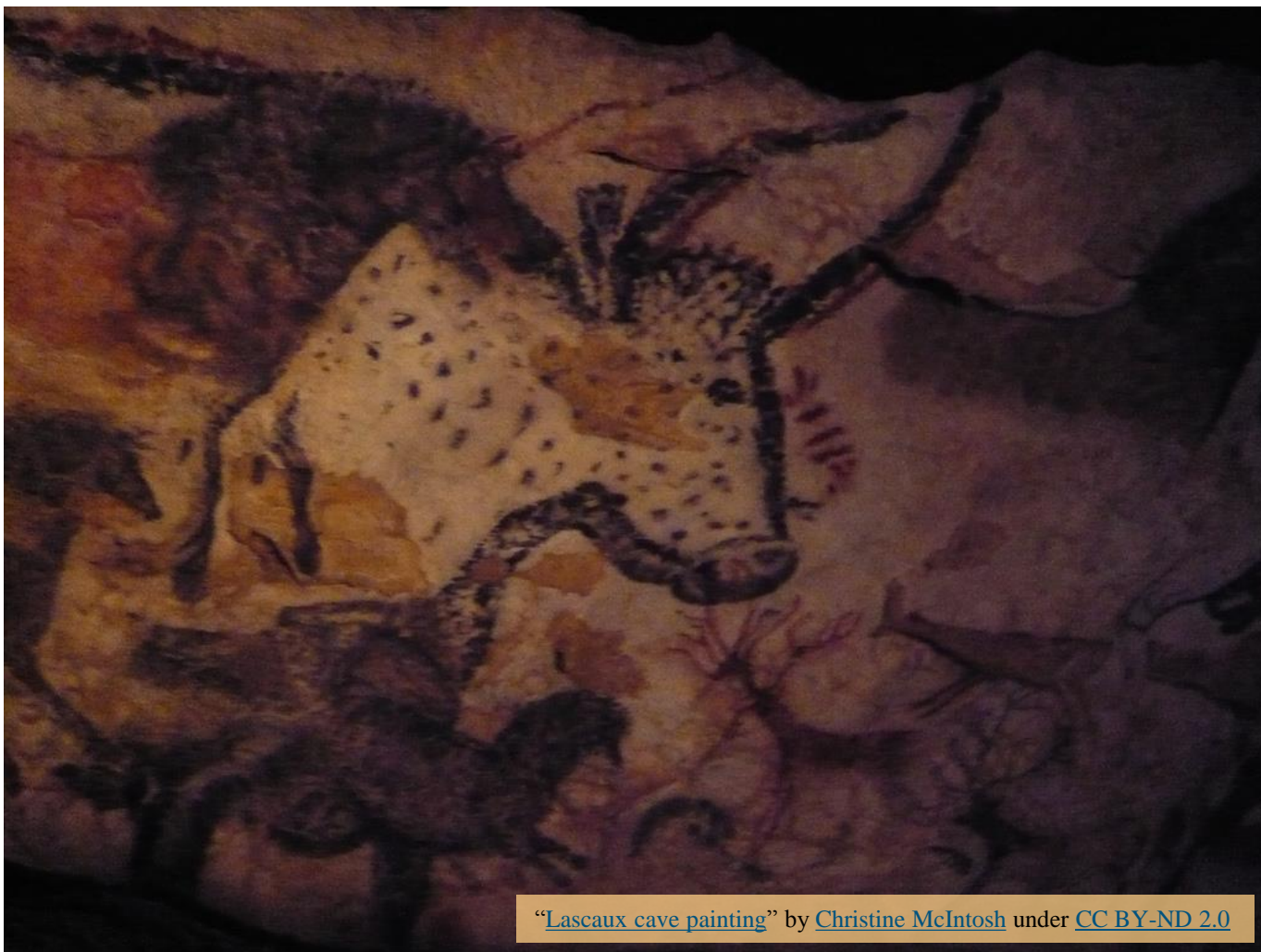
specify HTTP response headers for caching and content encoding



embed metadata, especially character encoding



use durable data formats



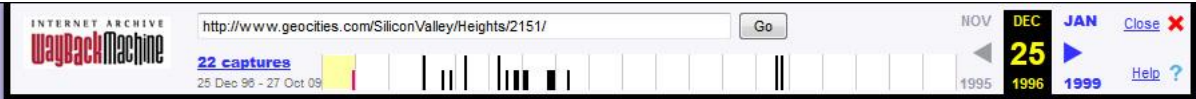
“Lascaux cave painting” by [Christine McIntosh](#) under [CC BY-ND 2.0](#)

prefer responsive design over user-agent personalization





“«Responsive web design» - 217/366” by [Roger Ferrer Ibáñez](#) under [CC BY-NC-SA 2.0](#)

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](#) [Internet Explorer](#)
or
 


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.
- Desktop Publishing, Graphics & Designing.
- PC Consulting/Training/Upgrade/Repair/Installation.
- HTML3 Home Page Programing & Designing.
- Internet Consulting/Training.
- Technical Advising.

Internet Archive Wayback Machine: “Welcome to A Multidimensional Perception ~/*\= & PCGuru”



Web Archiving

TOOLS AND SERVICES


“giant mechanical spider & crowd” by [mjtmail \(tiggy\)](#) under [CC BY 2.0](#)

Heritrix

Heritrix: Admin Console - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

[←](#) [→](#) [↺](#) [✕](#) [🏠](#) [http://foo.edu:8080/index.jsp](#) [Go](#) [G|](#)


Status as of Jun. 23, 2006 19:20:37 GMT Alerts: no alerts
CRAWLING JOBS **RUNNING job: 150-sites-deciding-10MB**
Admin Console 0 jobs pending, 9 completed 9108 URIs in 4m43s (25.5/sec)

[Console](#) [Jobs](#) [Profiles](#) [Logs](#) [Reports](#) [About](#) [Help](#)

Crawler Status: CRAWLING JOBS | [Hold](#)

Jobs
Running: 150-sites-deciding-10MB
0 pending, 9 completed
Alerts: 0 (0 new)

Memory
102217 KB used
117056 KB current heap
236224 KB max heap

Job Status: RUNNING | [Pause](#) | [Checkpoint](#) | [Terminate](#)

Rates
25.5 URIs/sec (32.27 avg)
595 KB/sec (545 avg)

Load
15 active of 50 threads
1.0 congestion ratio
399 deepest queue
79 average depth

Time
4m43s elapsed
3m34s remaining (estimated)

Totals
downloaded 9108 56% 6874 queued
15997 total downloaded and queued
151 MB uncompressed data received

[Refresh](#)

[Shut down Heritrix software](#) | [Logout](#)

Done

[Wikimedia Commons: "File:Heritrix-screenshot.png"](#)

Wget

```
adam@a: ~
adam@a:~$ wget https://en.wikipedia.org/wiki/Wget
--2014-02-09 07:38:43-- 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

HTTrack WEBSITE COPIER

Free software offline browser

[About](#)

[Download](#)

[Manual](#)

[Forum](#)

[Blog](#)

[Information](#)

[Français](#)

Version 3.48-3 (04/11/2014)

Engine fixes (keep-alive, redirects, new hashtables, unit tests)

Installing HTTrack: Go to the [download section](#) now!

For help and questions: [Visit the forum](#), [Read the documentation](#), [Read the FAQs](#), [Browse the sources](#)

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**

[Web](#) [Video](#) [Texts](#) [Audio](#) [Software](#) [About](#) [Account](#) [TVNews](#) [OpenLibrary](#)

[Home](#) [Donate](#) | [Store](#) | [Blog](#) | [FAQ](#) | [Jobs](#) | [Volunteer Positions](#)

Search: All Media Types  [Advanced Search](#)

Anonymous User (login or join us) [Upload](#)

Universal access
to all knowledge

INTERNET ARCHIVE

WayBackMachine

[BROWSE HISTORY](#)

407 Billion web pages saved over time.

[DONATE](#)

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

[SAVE PAGE](#)

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



WAIL

Web Archiving Integration Layer
One-Click User Instigated Preservation

Fork me on GitHub

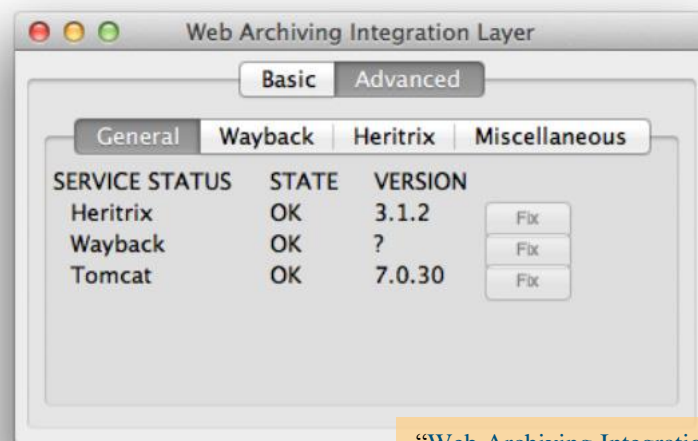
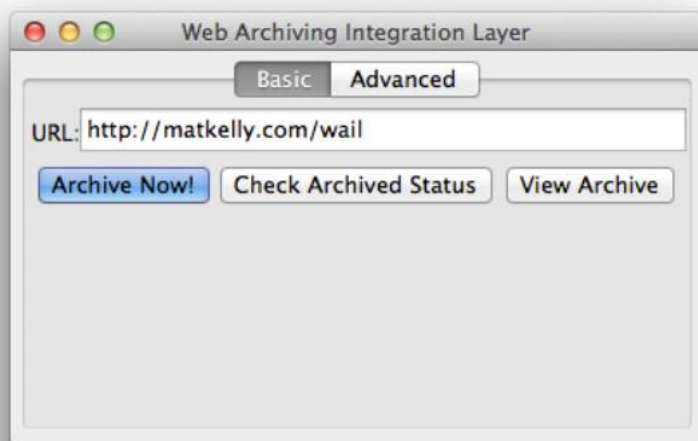
What Is It?

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 [warc-proxy](#). Support packages include [Apache Tomcat](#), [phantomjs](#) and [pvinstaller](#).

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

Screenshots




“Web Archiving Integration Layer”

Memento



[“Memento”](#)

assess archivability w/ Archive Ready


Archive Ready BETA
website archivability testing tool

[Home](#)
[Help](#)
[FAQ](#)
[API](#)

Is your website Archive Ready?

[Check now »](#)

What is ArchiveReady?

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

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 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 [🔗](#).

Why bother?

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


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**.

Do you need more information?

Banos V., Kim Y., Ross S., Manolopoulos Y.: **CLEAR: a credible method to evaluate website archivability**, 10th International Conference on Preservation of Digital Objects (**iPRES'2013**), Lisbon, 2013. [PDF](#) [🔗](#)

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


Archive Ready BETA
website archivability testing tool

About

ArchiveReady was created by [Vangelis Banos](#) [🔗](#) (© 2012-2013). It is a **personal research project** and is not supported by any public or private organization.

Archive Ready

thank you!

