STANFORD UNIVERSITY LIBRARIES

Building Archivable Websites

Nicholas Taylor
Web Archiving Service Manager
Digital Library Systems and Services

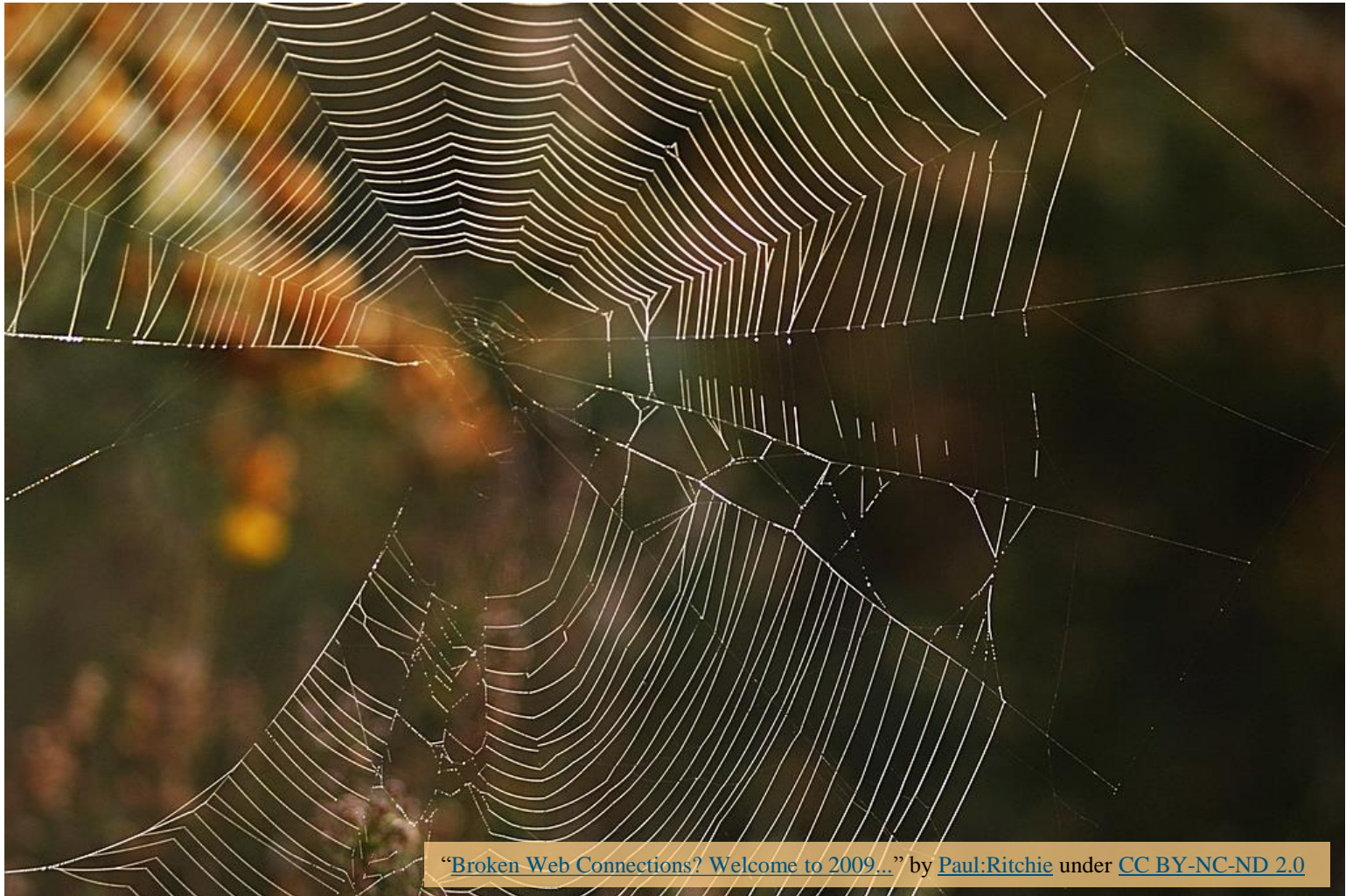Drupal Camp
April 19, 2014

Why Build

ARCHIVABLE WEBSITES?

future users are users, too

maintain web usability

"Broken Web Connections? Welcome to 2009..." by Paul:Ritchie under CC BY-NC-ND 2.0

improve temporal web usability
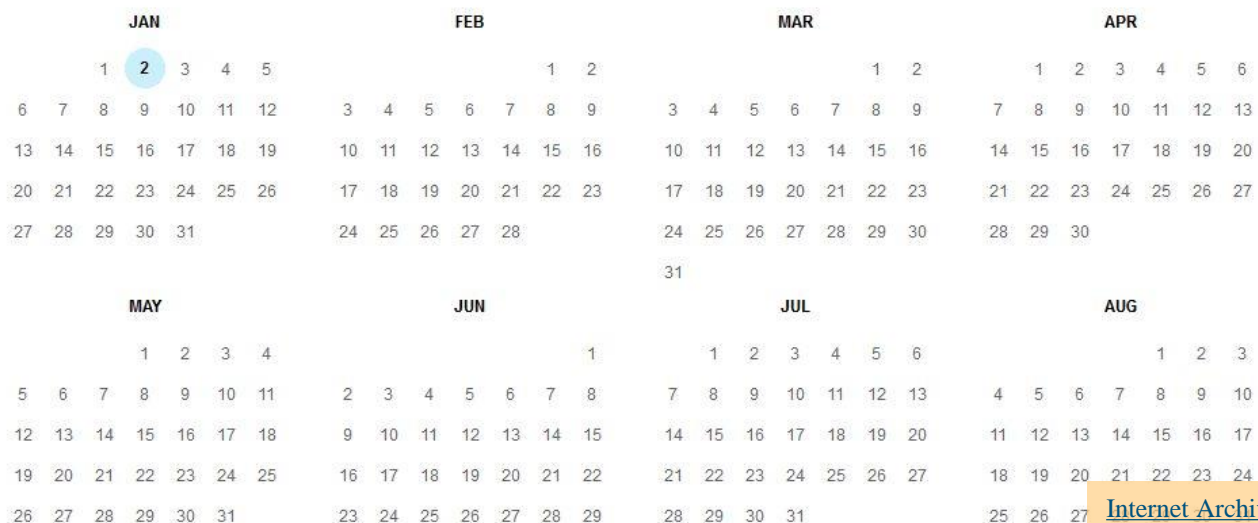
Internet Archive: "Wayback Machine"

improve temporal web usability
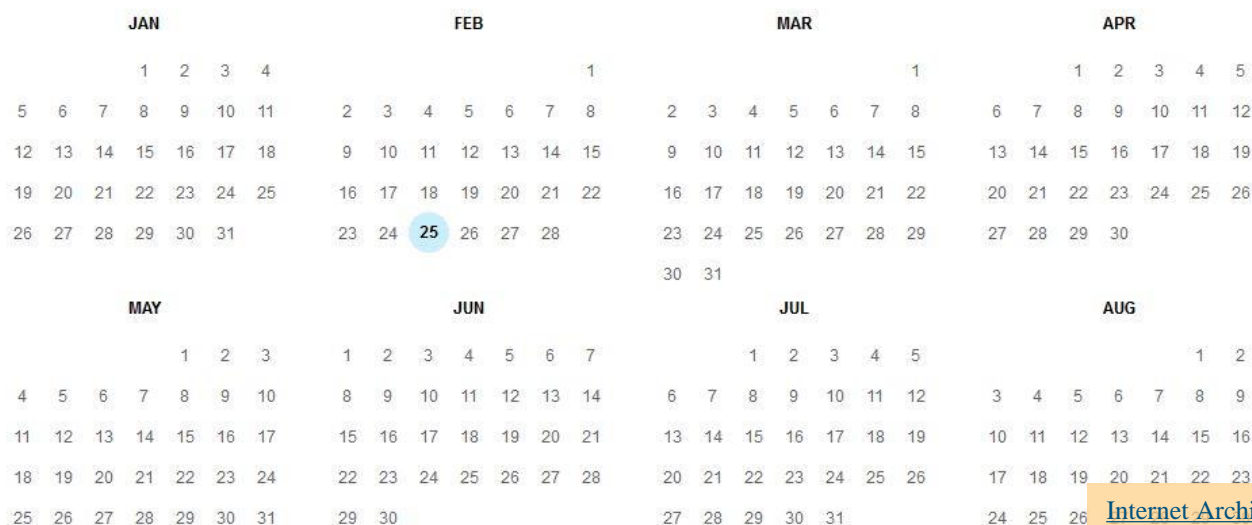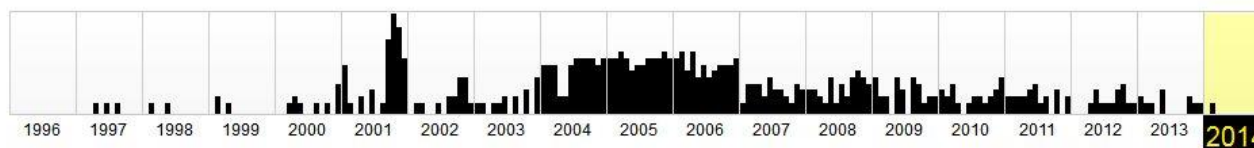
Internet Archive: "Wayback Machine"

recover your lost website

"Warrick"

refer to earlier website versions

institutional history

Internet Archive Wayback Machine: "Stanford University Homepage"
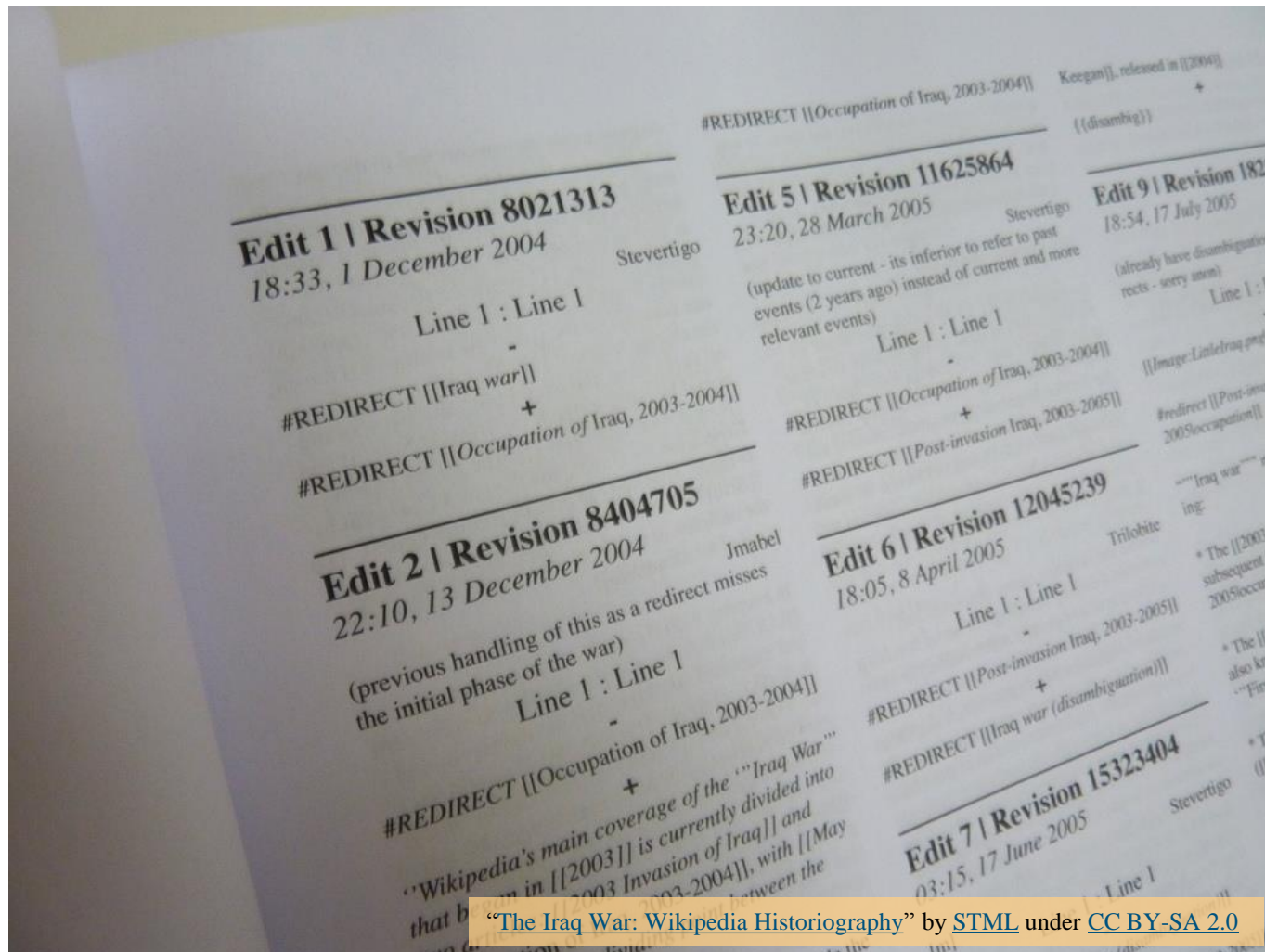
websites are cultural artifacts

```
                                    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            A list of W3 project components and their current
            Products[11]        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"

STANFORD UNIVERSITY LIBRARIES

facilitate compliance
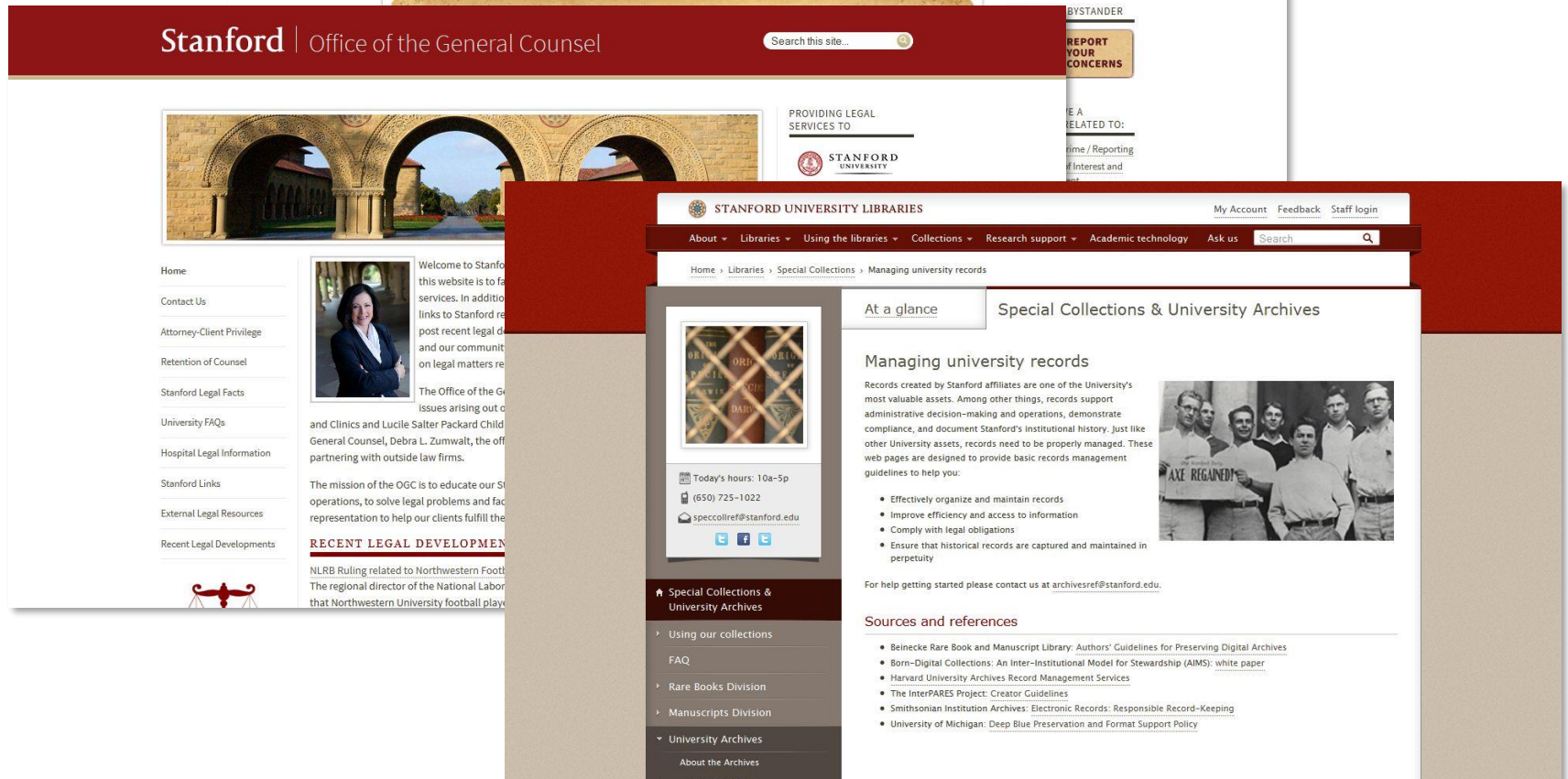
optimize for other crawlers
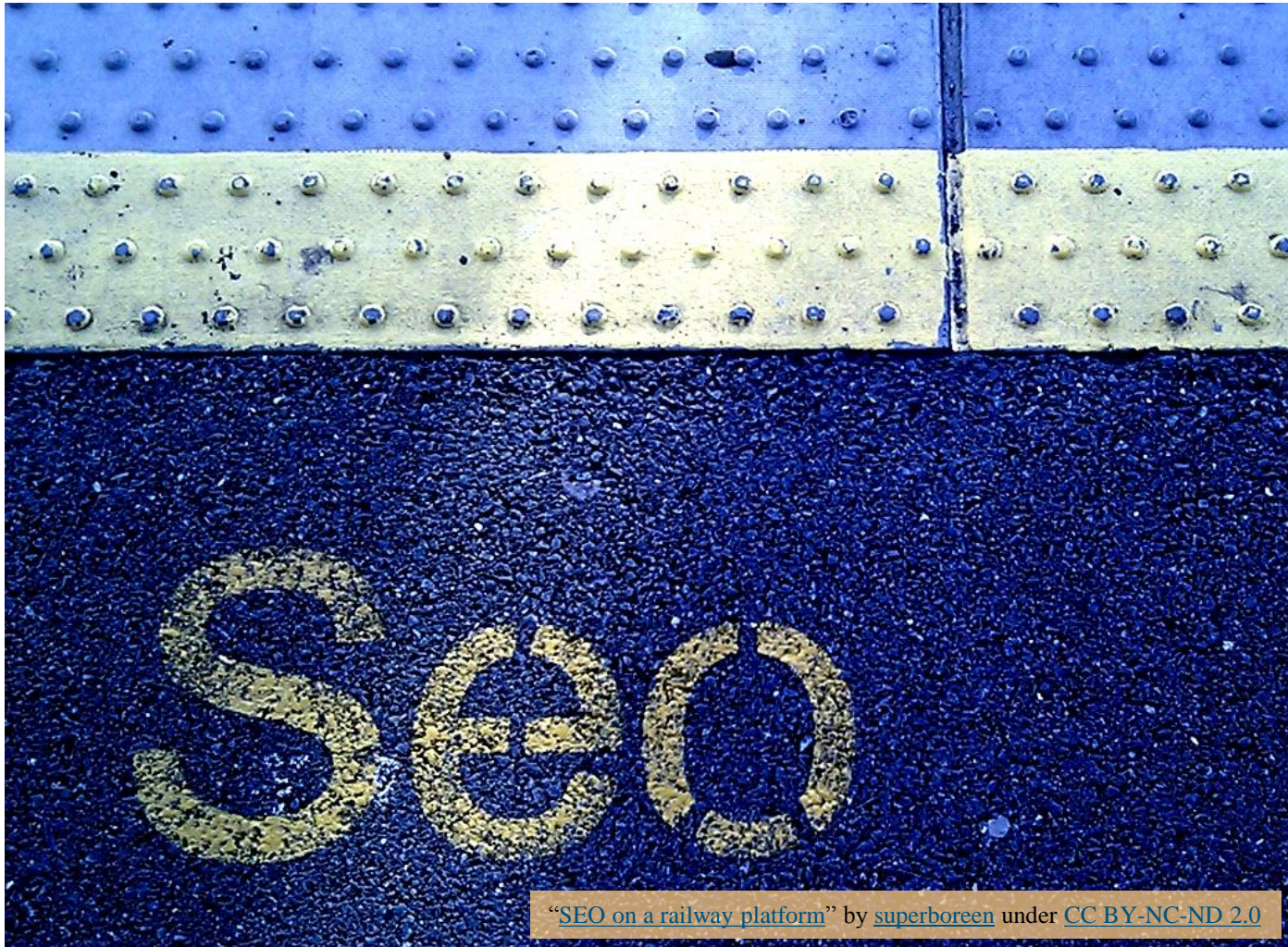
How to

IMPROVE ARCHIVABILITY

follow web standards and accessibility guidelines

use a site map, transparent links, and contiguous navigation
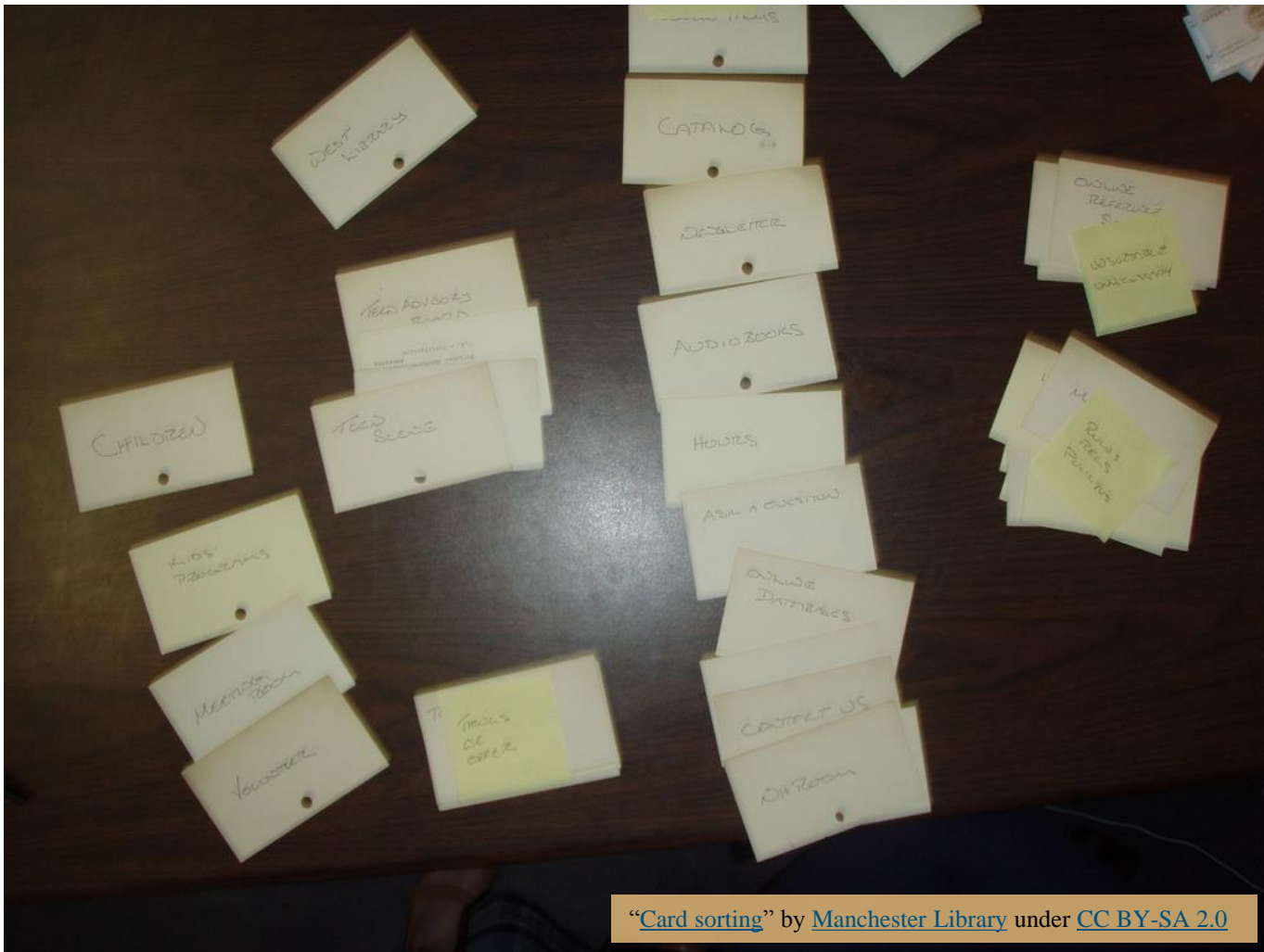
maintain stable URLs and redirect when necessary

"San Francisco-Oakland Bay Bridge 1442a" by Don Barrett under CC BY-NC-ND 2.0

use semantically-meaningful URLs

m9(ˆДˆ)プギャーーーッ

Welcome to w63.w63.org!

wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww.wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww
wwwwwwwwwwwwwwwwwwwwwwwww.org

http://w63.org

Console   HTML   CSS   Script   DOM   Net ▾   Cookies

Clear   Persist   All   HTML   CSS   JavaScript   XHR   Images   Plugins   Media   Fonts

URL	Status	Domain		Size	Remote IP
wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww.wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww.org					
GET ww	200 OK	wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww.wwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwwww.org		1.3 KB	116.58.189.212:80

Headers   Response   HTML   Cache

200 OK

Response Headers   view source

Access-Control-Allow-Orig...   *
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

Request Headers   view source

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

"w63.w63.org"

0 requests

be careful w/ robot exclusion rules

minimize reliance on external assets necessary for presentation

Internet Archive Wayback Machine: "Stanford Department of English"
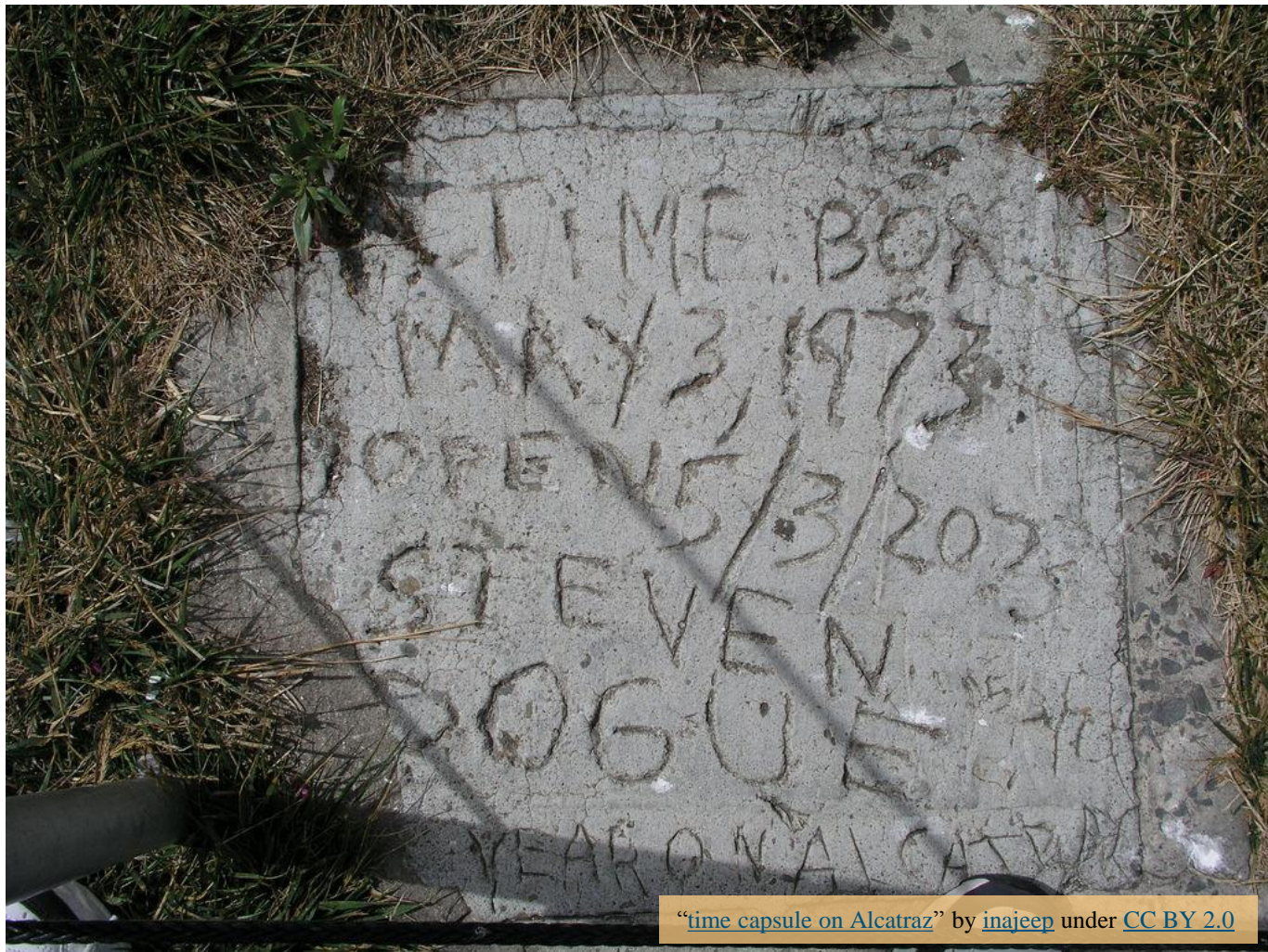
minimize reliance on external assets necessary for presentation

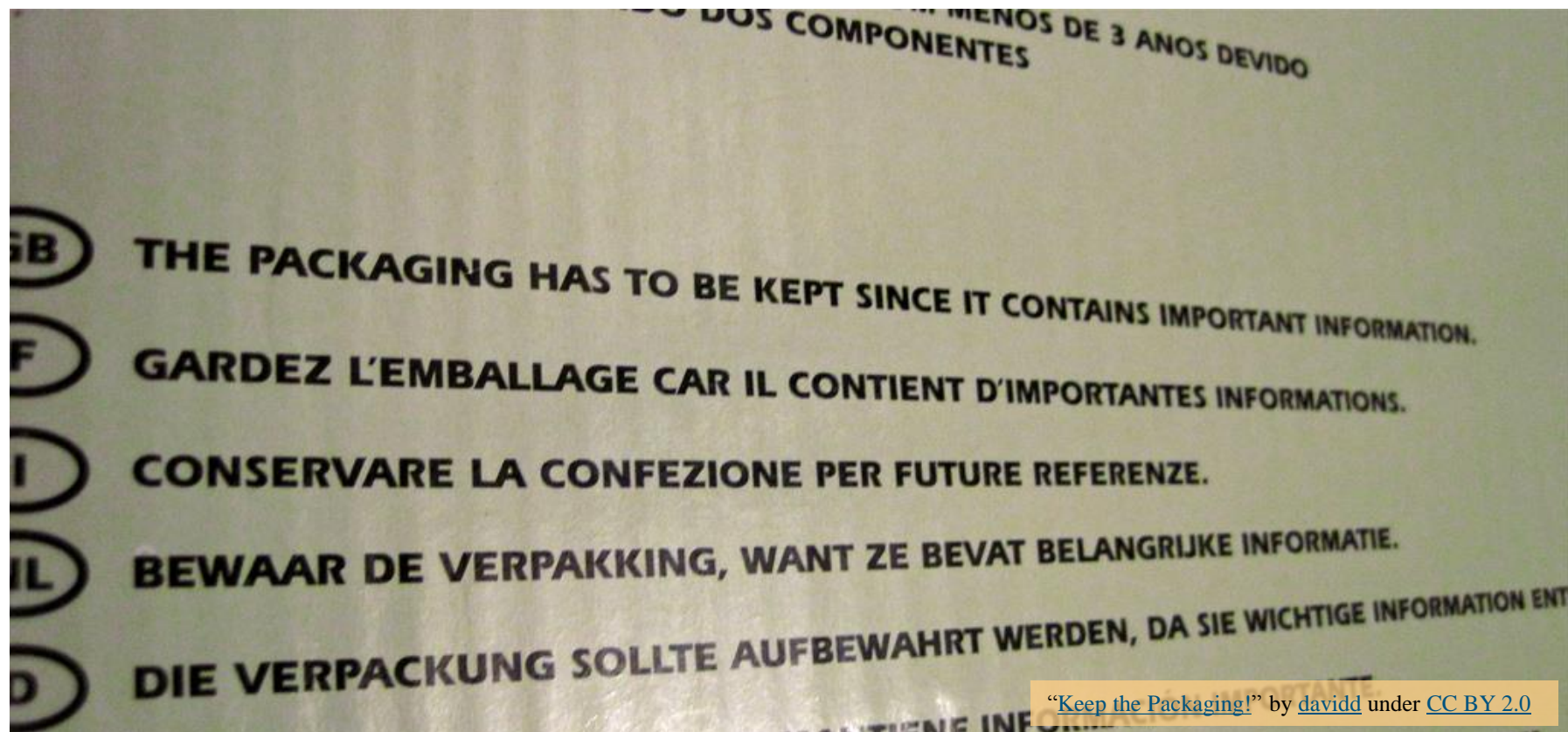"[Stanford Department of English](#)"
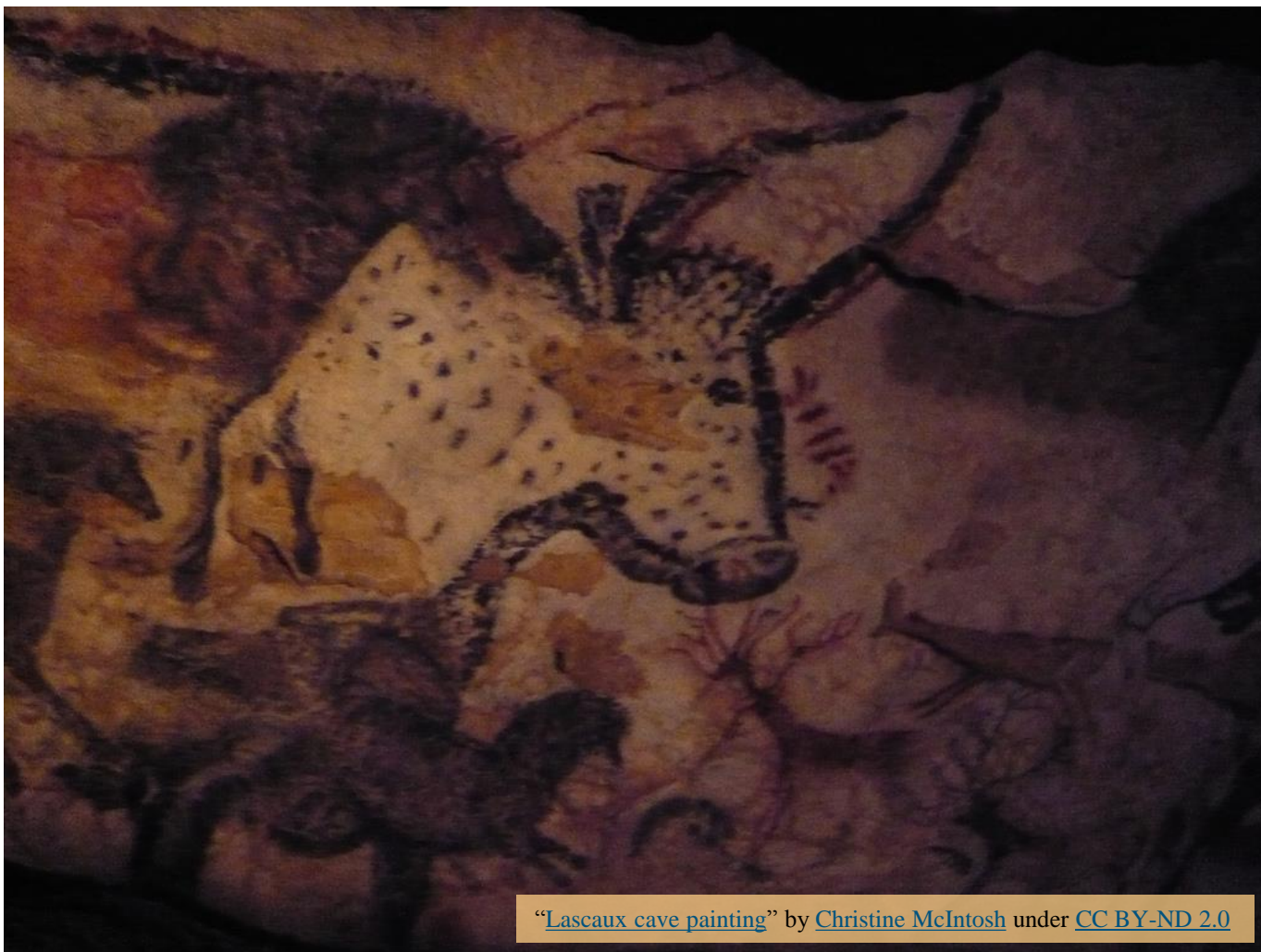
serve reusable assets from a single, common location
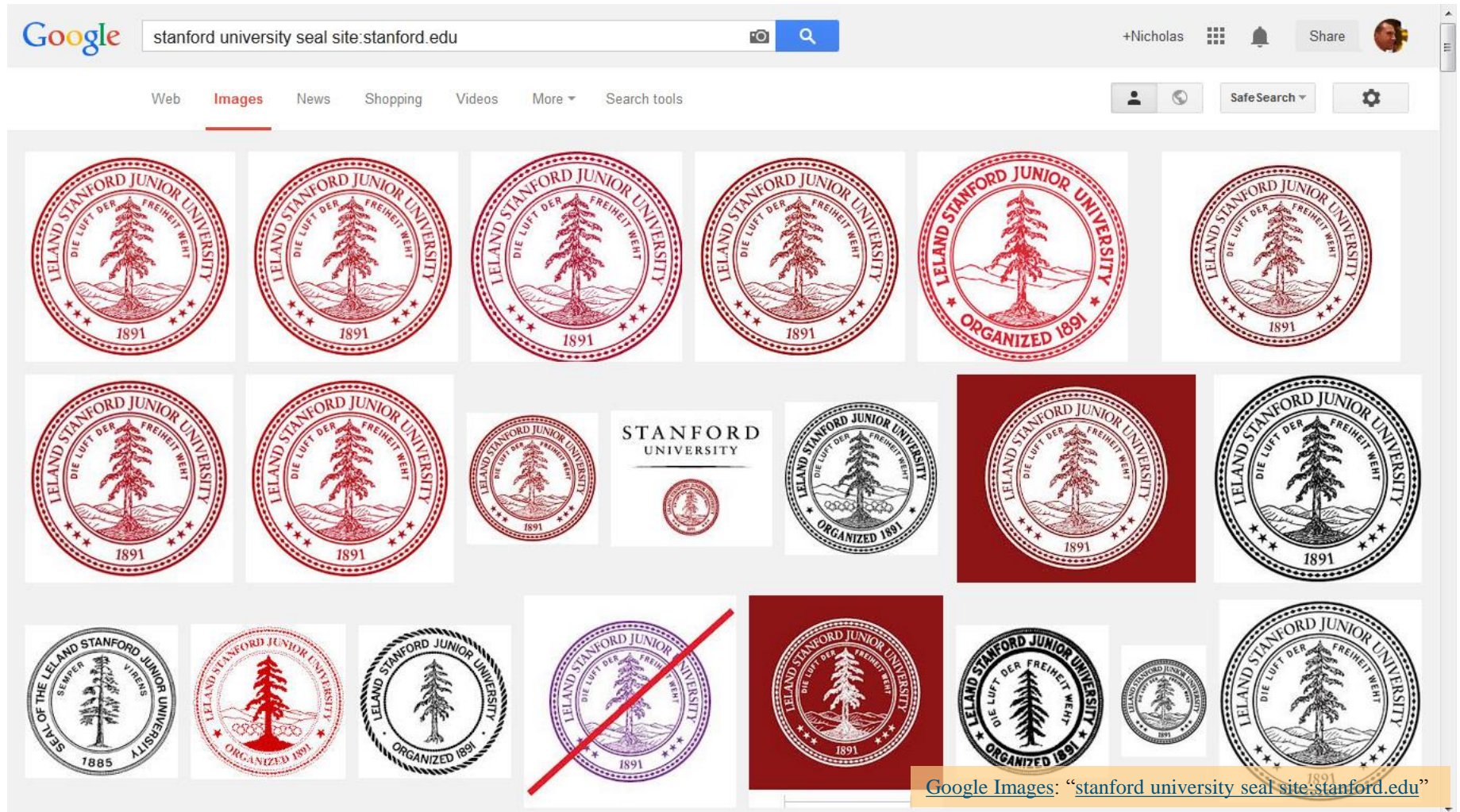
specify HTTP response headers for caching and content encoding

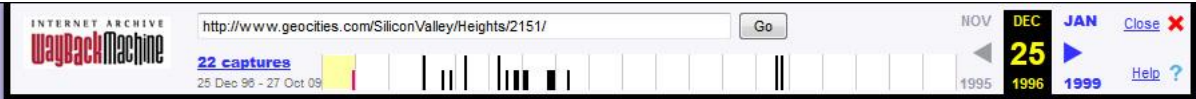# embed metadata, especially character encoding

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

Web Archiving

TOOLS AND SERVICES

Heritrix

Wget

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.

© 2014 Xavier Roche & other contributors - Web Design: Leto Kauler.

"HTTrack Website Copier"

Wayback

"Internet Archive Wayback Machine"

Web Archiving Integration Layer

Memento

"Memento"

assess archivability w/ Archive Ready

STANFORD UNIVERSITY LIBRARIES

Archive Ready BETA
website archivability testing tool

🏠 Home ℹ️ Help ❓ FAQ ⇄ API

Is your website Archive Ready?

http://www.yoursite.com/   [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.

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?
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!

Nicholas Taylor
ntay@stanford.edu