



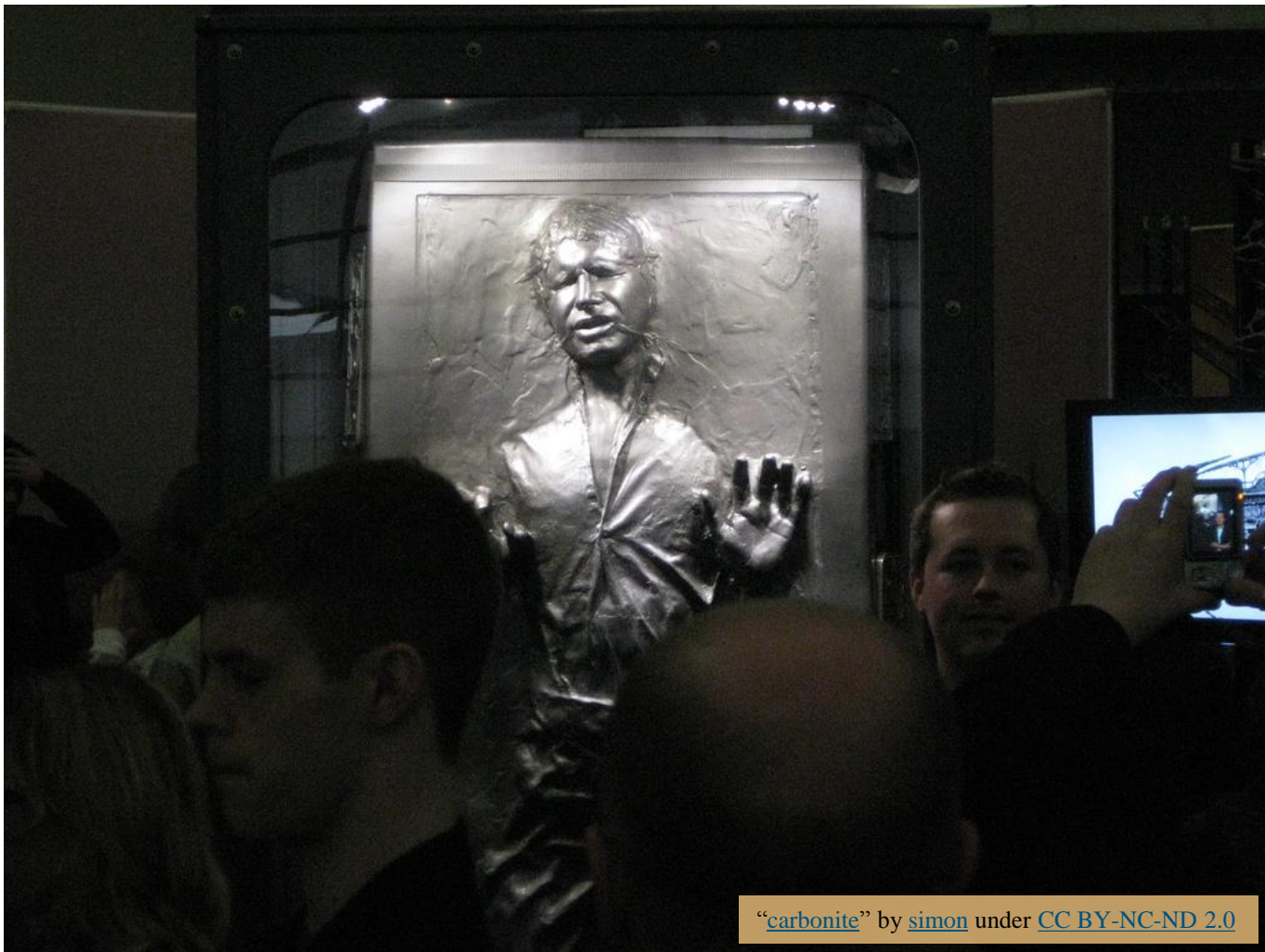
STANFORD UNIVERSITY LIBRARIES

Advocating for Web Archivability

Nicholas Taylor
Web Archiving Service Manager
[Digital Library Systems and Services](#)

[NDSA Content Working Group](#)
May 7, 2014

archivability?



“carbonite” by [simon](#) under [CC BY-NC-ND 2.0](#)

A photograph of a man in a striped shirt and glasses standing at the front of a room, speaking into a microphone. He is holding a glass of beer. The room is filled with rows of industrial washing machines. An audience of several people is seated in the foreground, facing the speaker. The scene is a 'soapbox lecture' in a laundry room.

Why

ADVOCATE?

“Richard Lang of Electric Works - Bernal Bubbles Soapbox lecture” by [Steve Rhodes](#) under [CC BY-NC-SA 2.0](#)

ignorance not indifference



“DSC_9293.jpg” by [Brian Shrader](#) under [CC BY-NC-SA 2.0](#)

good return on investment



“The Art Came Back” by [Michael Reeve](#) under [CC BY-NC-SA 2.0](#)

opens up new opportunities



“Too Many Opportunities” by [Erik Charlton](#) under [CC BY 2.0](#)

Making a

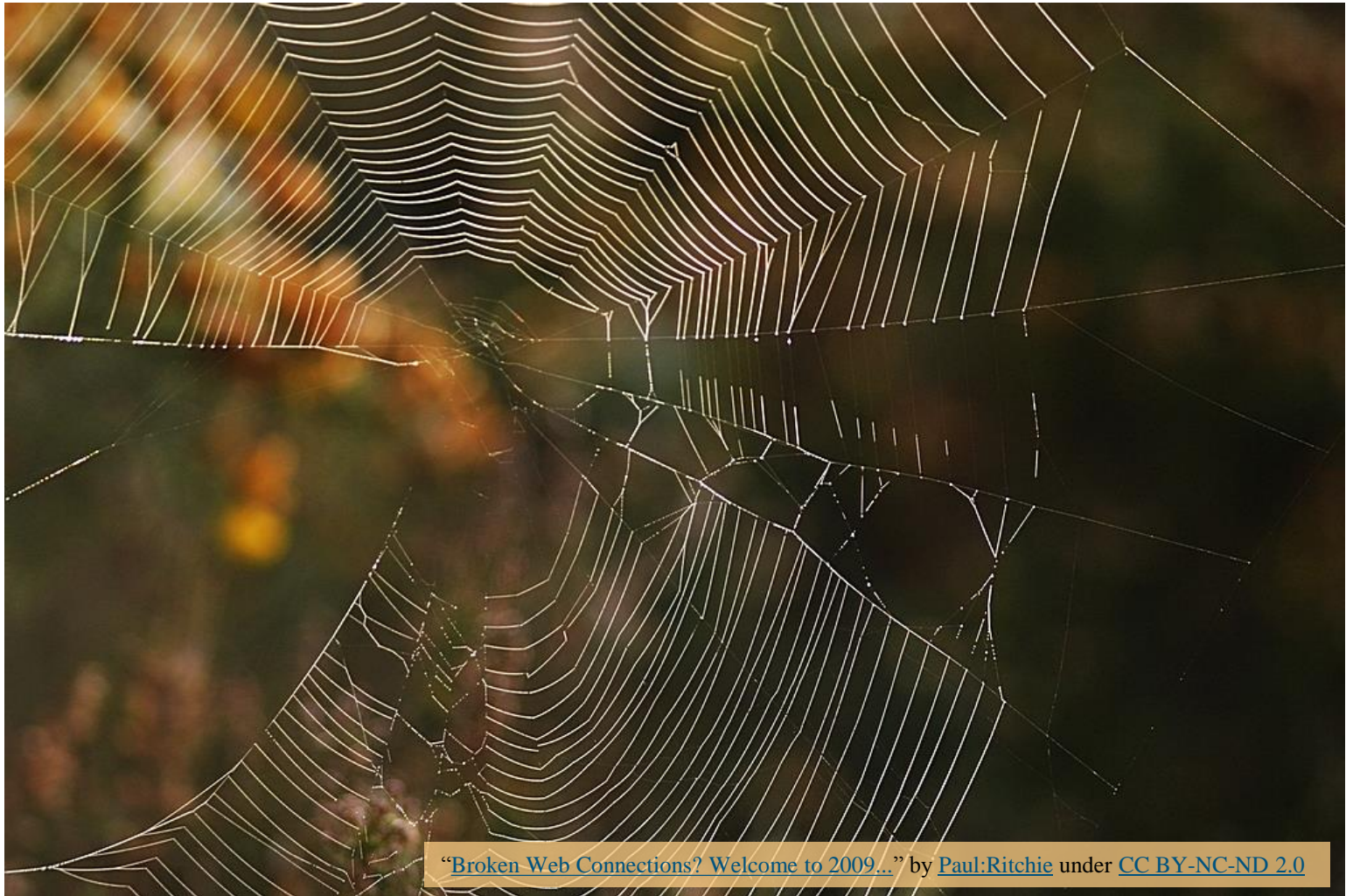
COMPELLING CASE

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



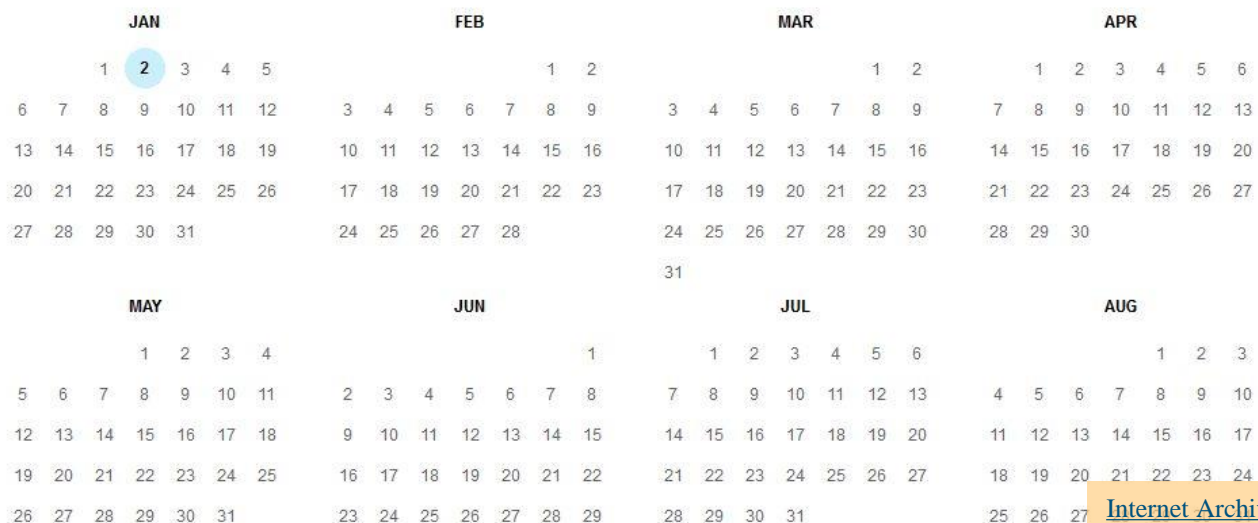
<http://bono.house.gov/>

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



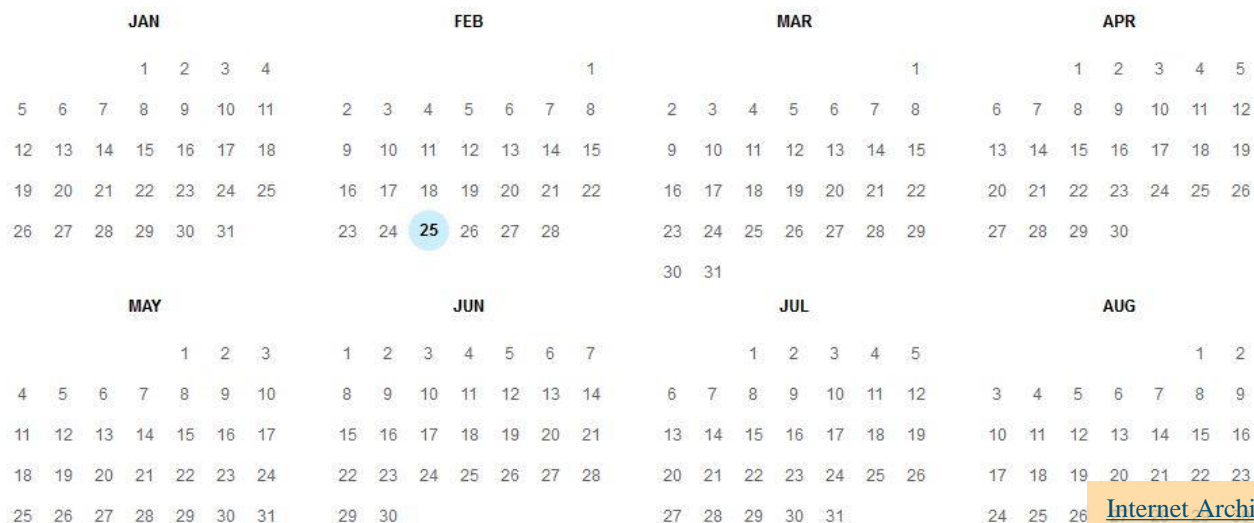
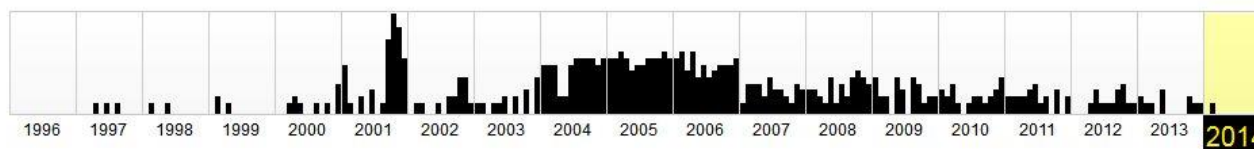
<http://house.gov/bono/>

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"

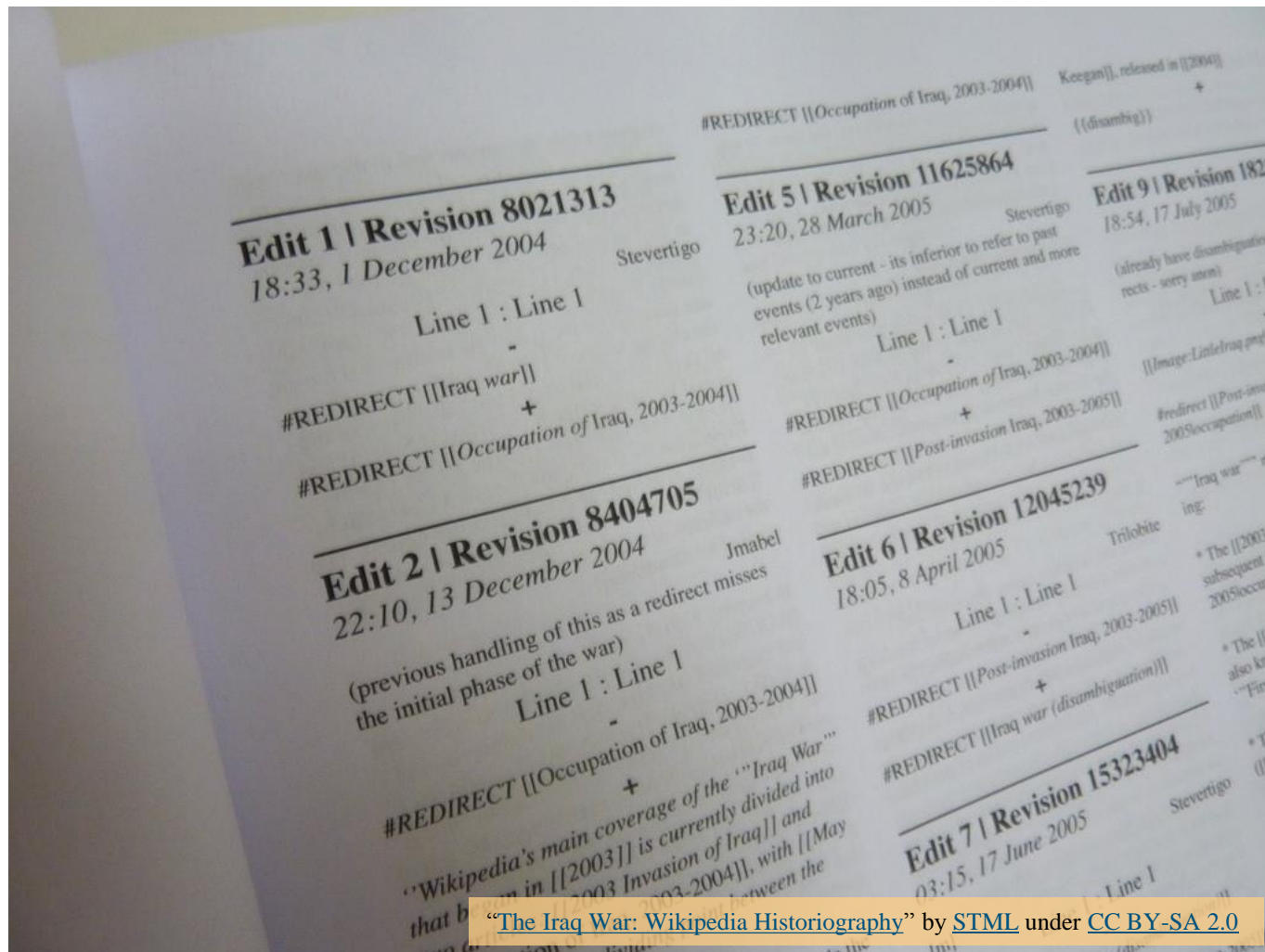
recover your lost website



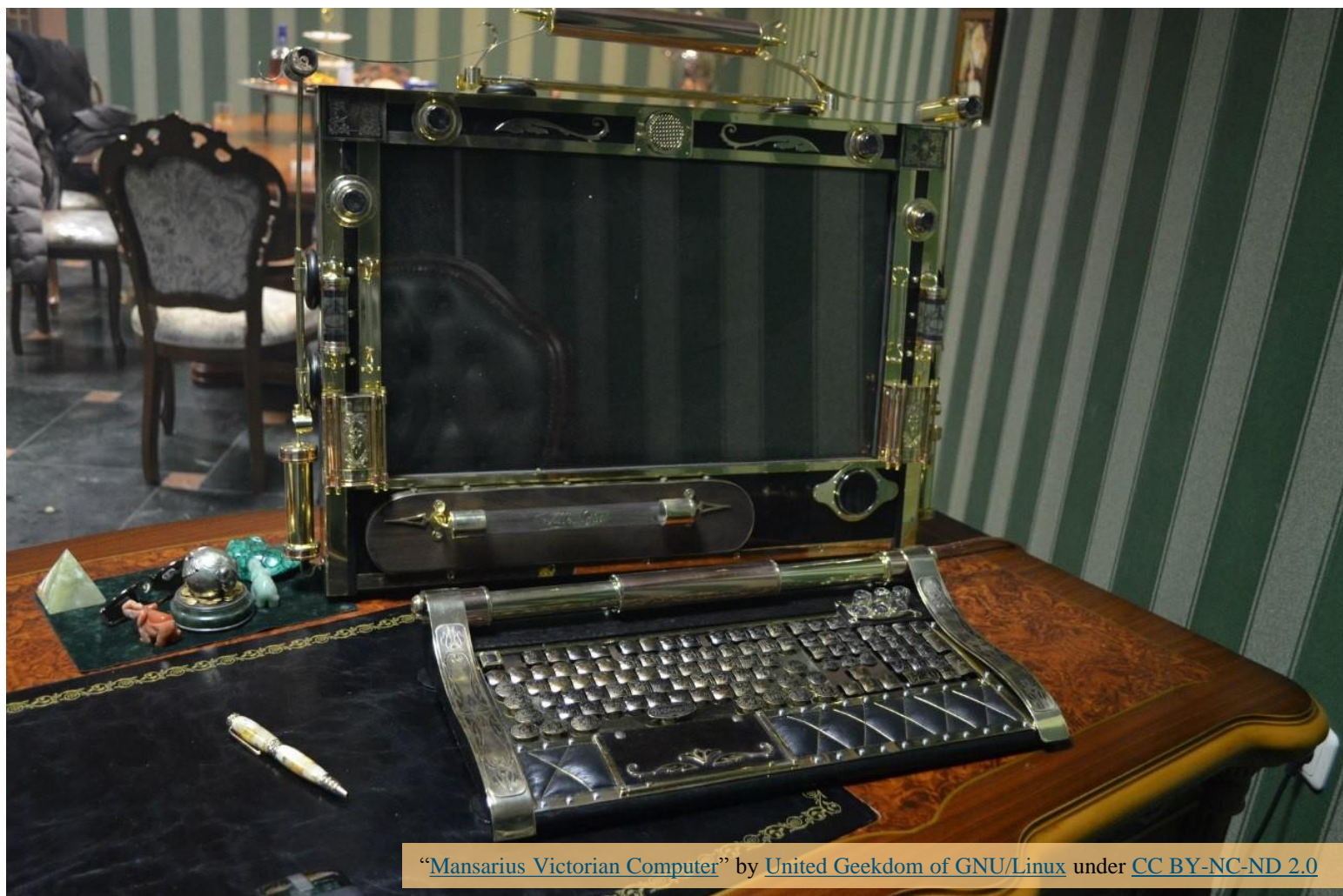
Old Dominion University
Computer Science Department
Norfolk, VA

“Warrick”

refer to earlier website versions



archivability is the new accessibility

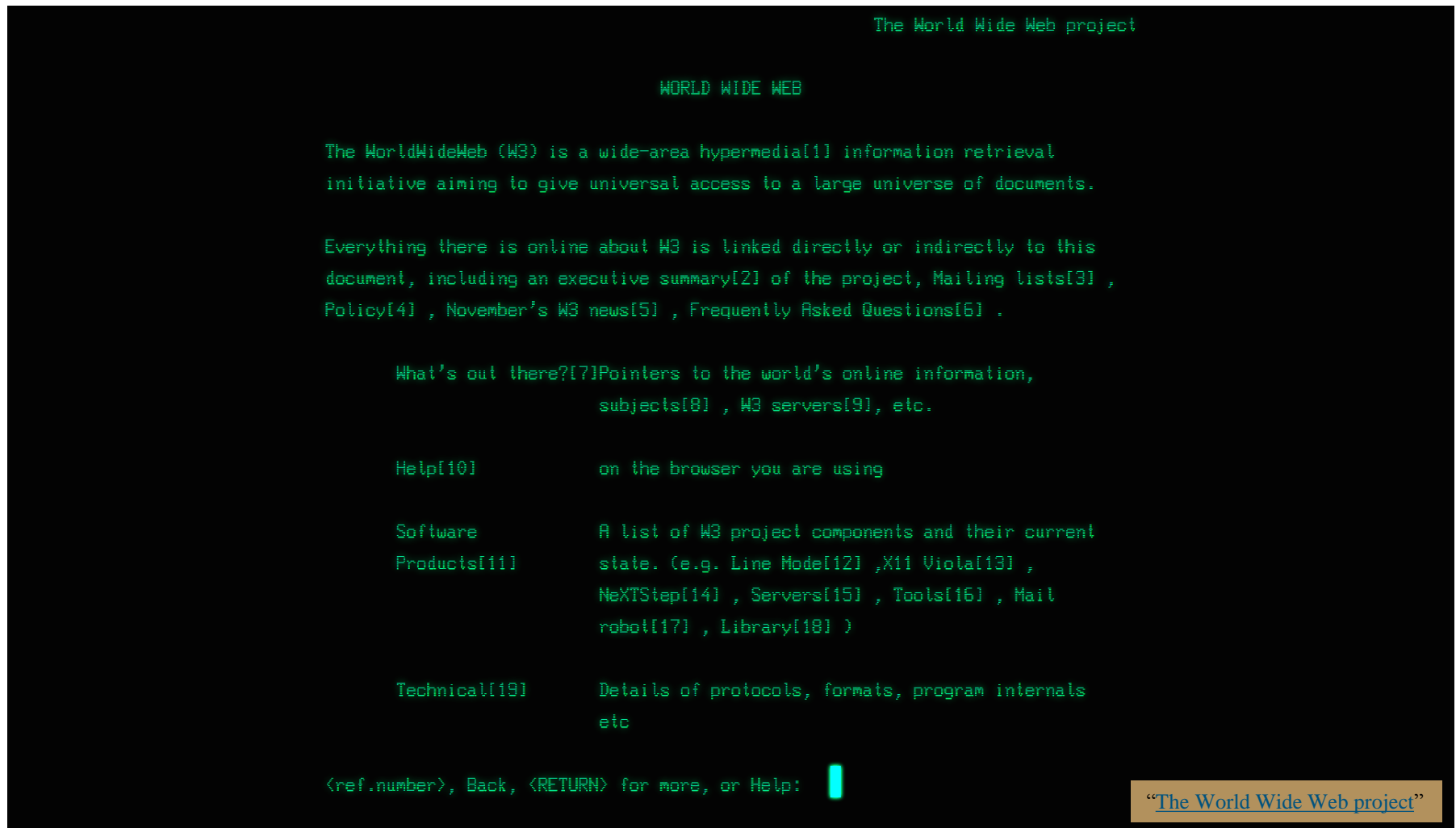


“Mansarius Victorian Computer” by [United Geekdom of GNU/Linux](#) under [CC BY-NC-ND 2.0](#)

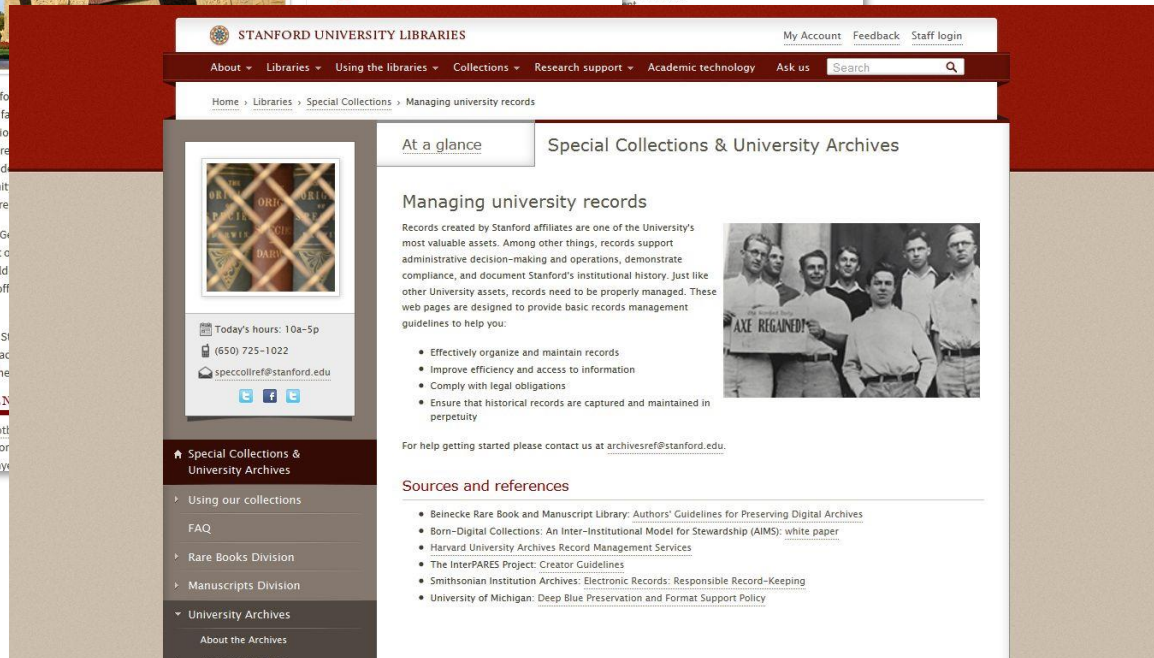
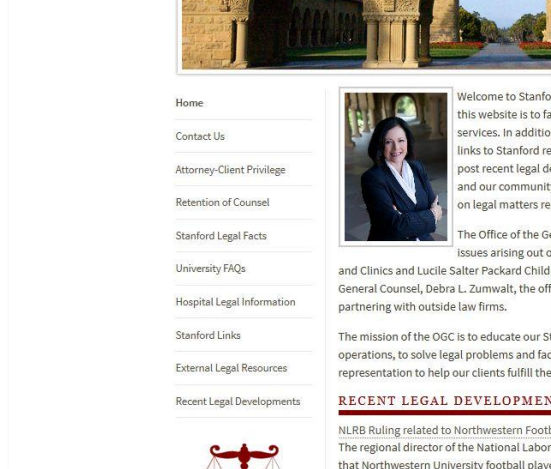
institutional history



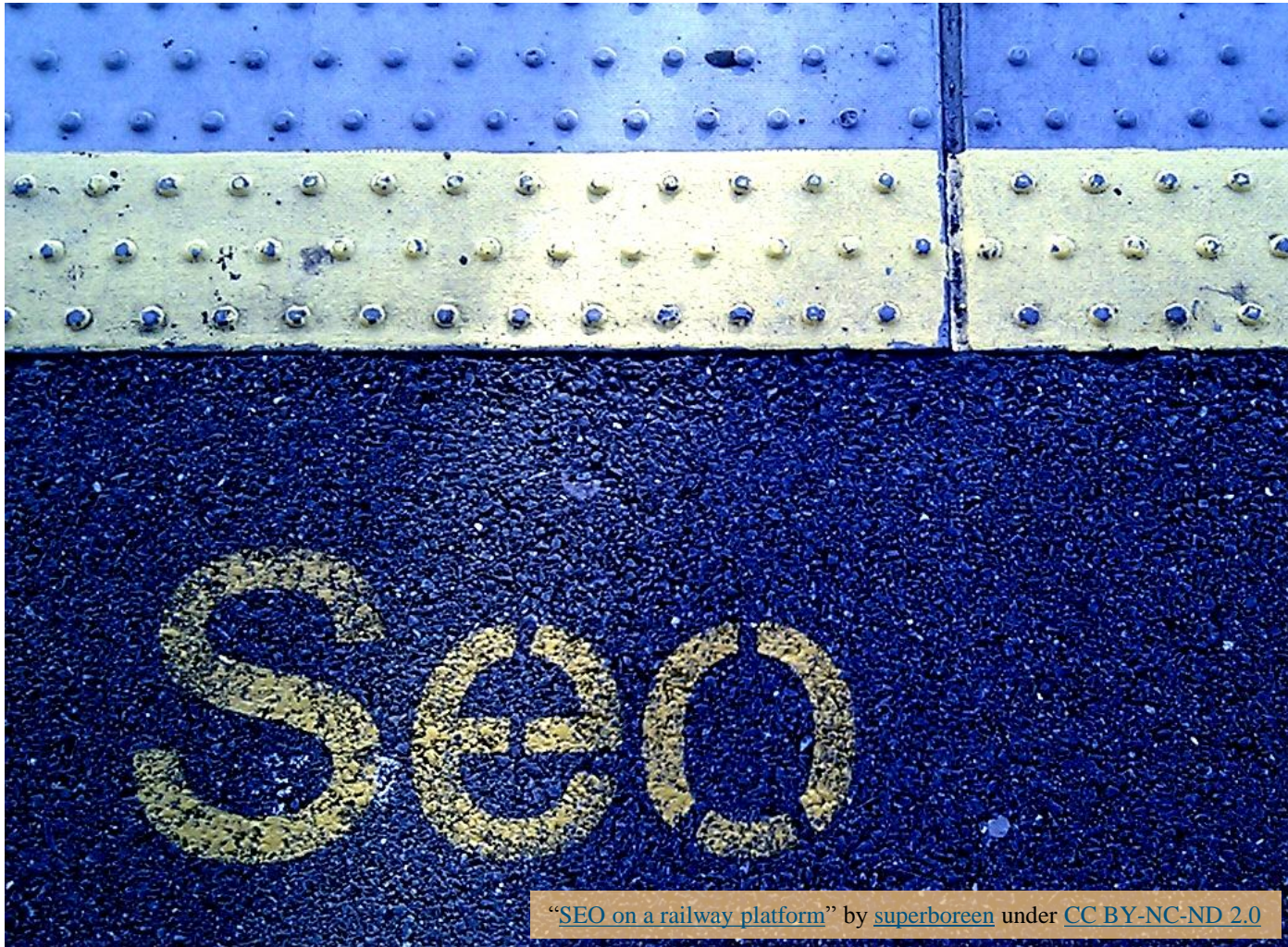
websites are cultural artifacts



facilitate compliance



optimize for other crawlers



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



How to

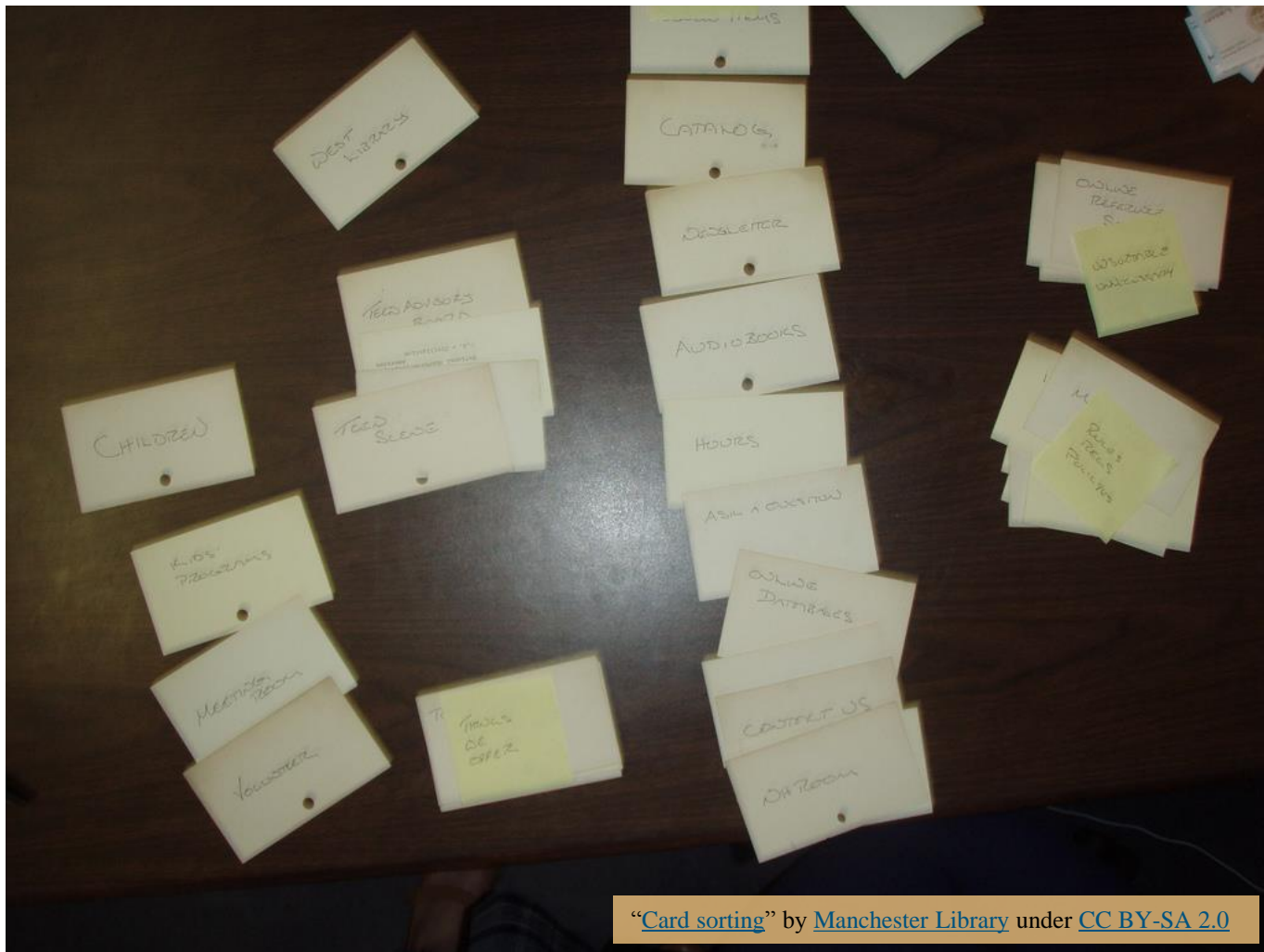
IMPROVE ARCHIVABILITY

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

follow web standards and accessibility guidelines



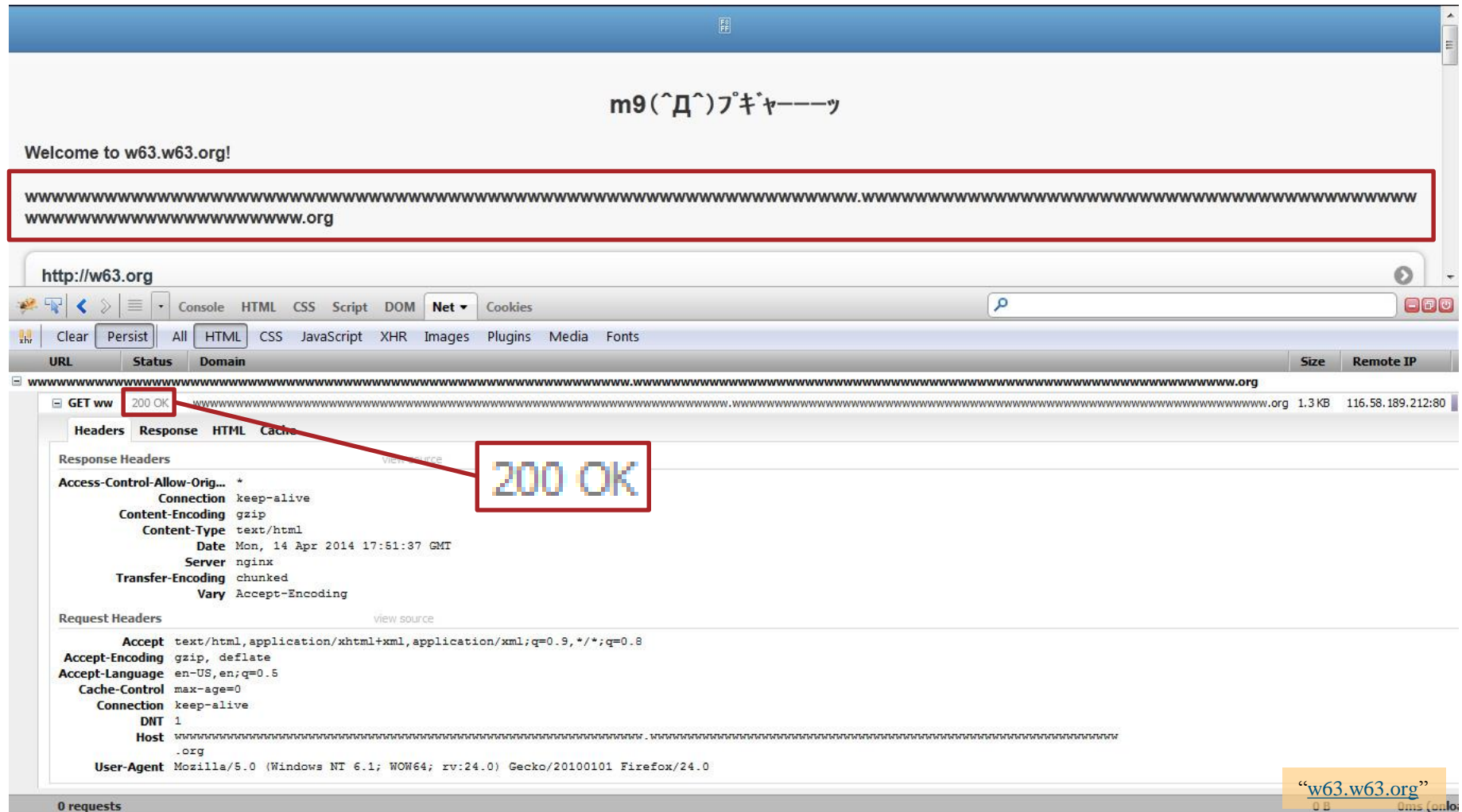
use a site map, transparent links, and contiguous navigation



maintain stable URLs and redirect when necessary



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



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”

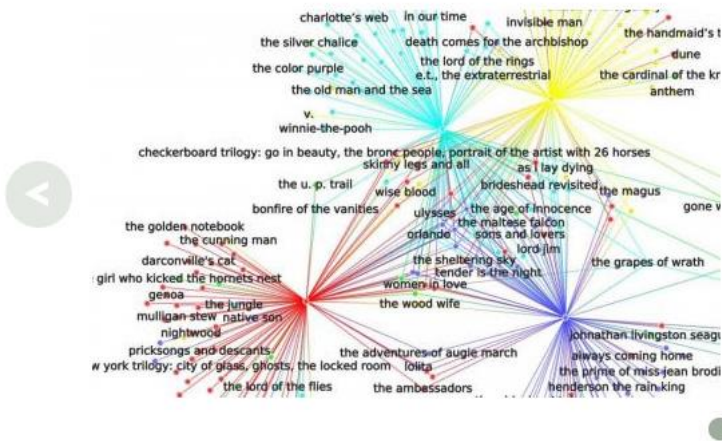
minimize reliance on external assets necessary for presentation

Stanford University
SUNetID Login

Stanford
Department
of English

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

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

Internet Archive Wayback Machine: “Stanford Department of English”

Affiliated Programs

Creative Writing Program



The graph displays numerous literary titles and authors as nodes, interconnected by lines representing relationships or co-occurrences. The nodes are color-coded into three main groups:

- Red Nodes (Left Side):** Includes titles like "the silver chalice", "the color purple", "the old man and the sea", "v.", "winnie-the-pooh", "checkerboard trilogy: go in beauty, the bronze people; portrait of the artist with 26 horses", "the u. p. trail", "bonfire of the vanities", "the golden notebook", "darconville's cat", "girl who kicked the hornets nest", "genoa", "the jungle", "mulligan-stew native son", "nightwood", "pricksongs and descants", and "n york trilogy: city of glass, ghosts, the locked room".
- Blue Nodes (Right Side):** Includes titles like "johnathan livingston seagull", "always coming home", "the prime of miss-jane brodie", "tenderison the ralking", "the grapes of wrath", "lord jim", "the sheltering sky", "tender is the night", "women in love", "the wood wife", "the adventures of augie march", "loila", "the ambassadors", "the lord of the flies", "the cardinal of the kn", "anthem", "gone v", "the maltese falcon", "sons and lovers", "ulysses", "the age of innocence", "orlando", "wise blood", "bridgeshead revisited", "as i lay dying", "skippy legs and all", "death comes for the archbishop", "the lord of the rings e.t., the extraterrestrial", "invisible man", "in our time", "charlotte's web", "the handmaid's tale", "dune", and "the handmaid's tale".
- Yellow Nodes (Top Center):** Includes titles like "the cardinal of the kn", "anthem", "gone v", "the maltese falcon", "sons and lovers", "ulysses", "the age of innocence", "orlando", "wise blood", "bridgeshead revisited", "as i lay dying", "skippy legs and all", "death comes for the archbishop", "the lord of the rings e.t., the extraterrestrial", "invisible man", "in our time", "charlotte's web", "the handmaid's tale", "dune", and "the handmaid's tale".

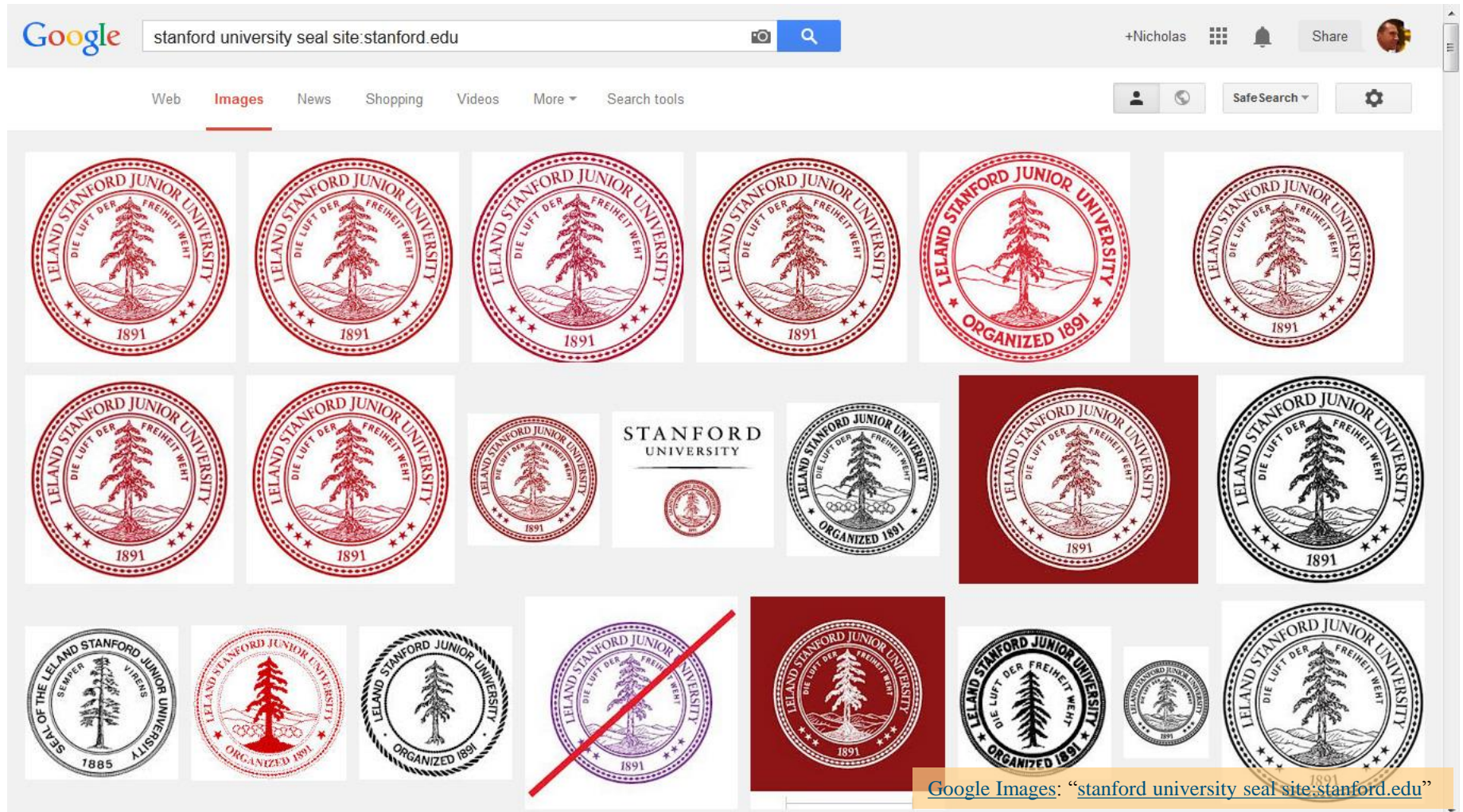
The connections between nodes vary in thickness, indicating the strength or frequency of the relationship. The graph illustrates a complex web of literary associations, with some titles acting as central hubs connecting multiple other works.

APR
10

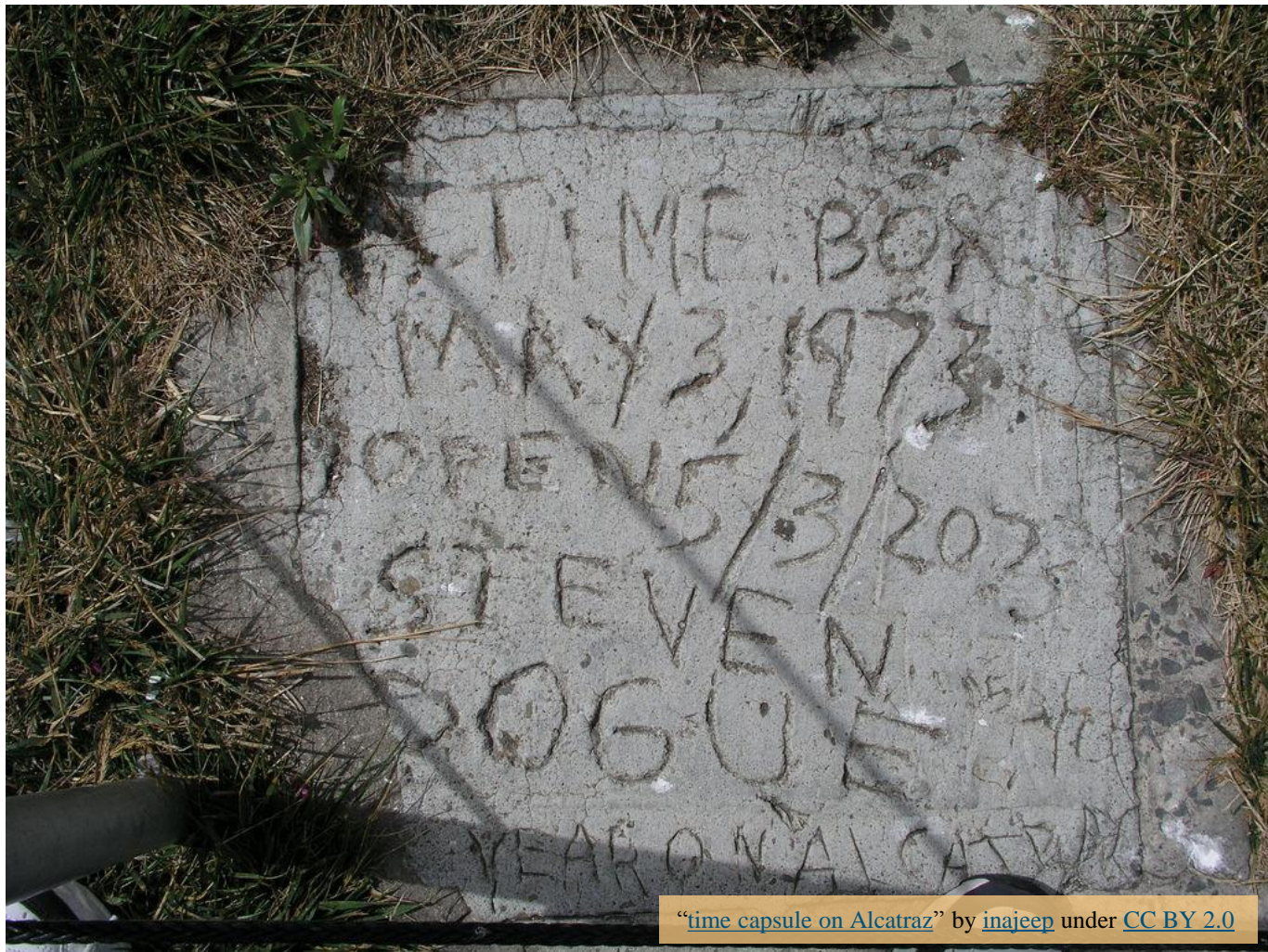
Creative Writing Program

“Stanford Department of English”

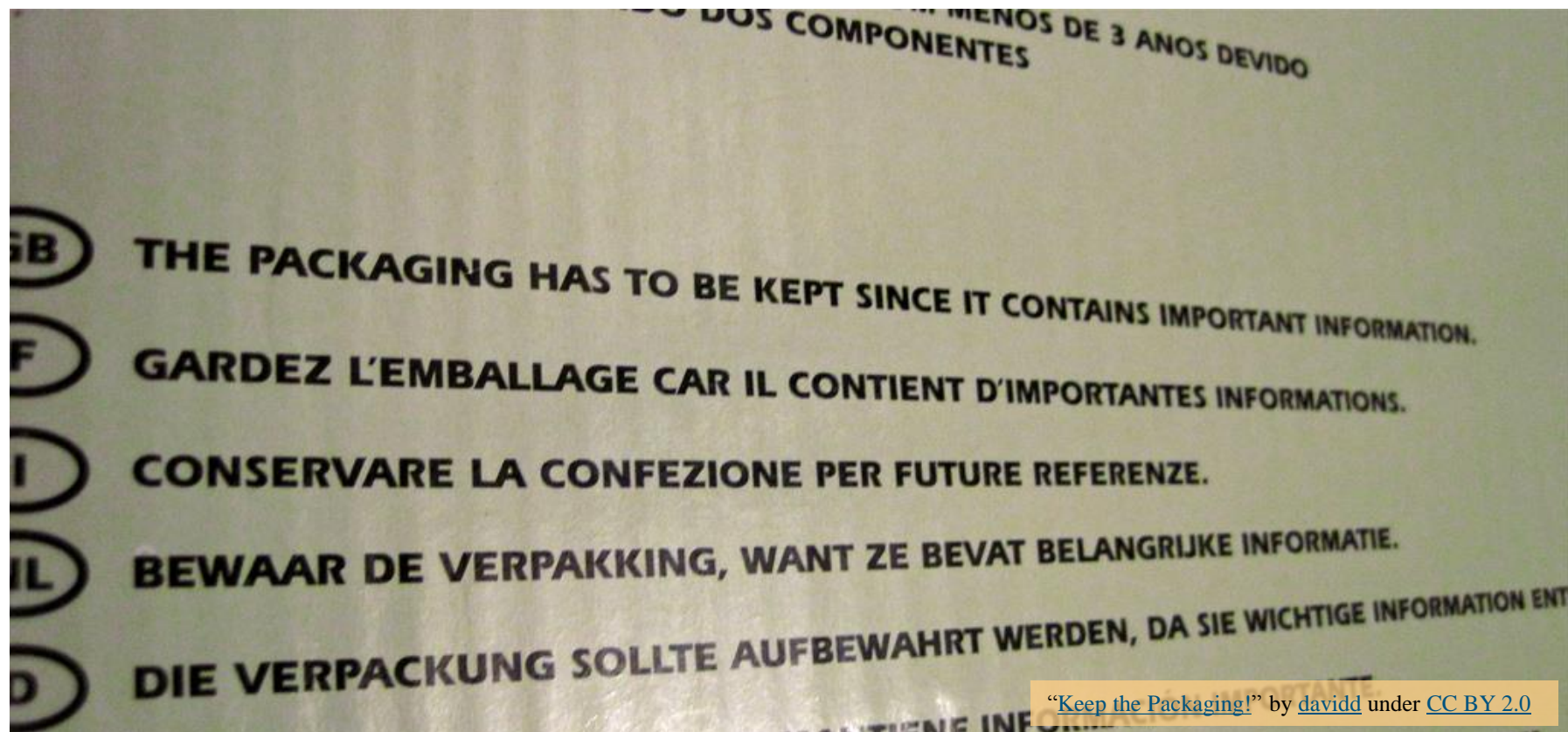
serve reusable assets from a single, common location



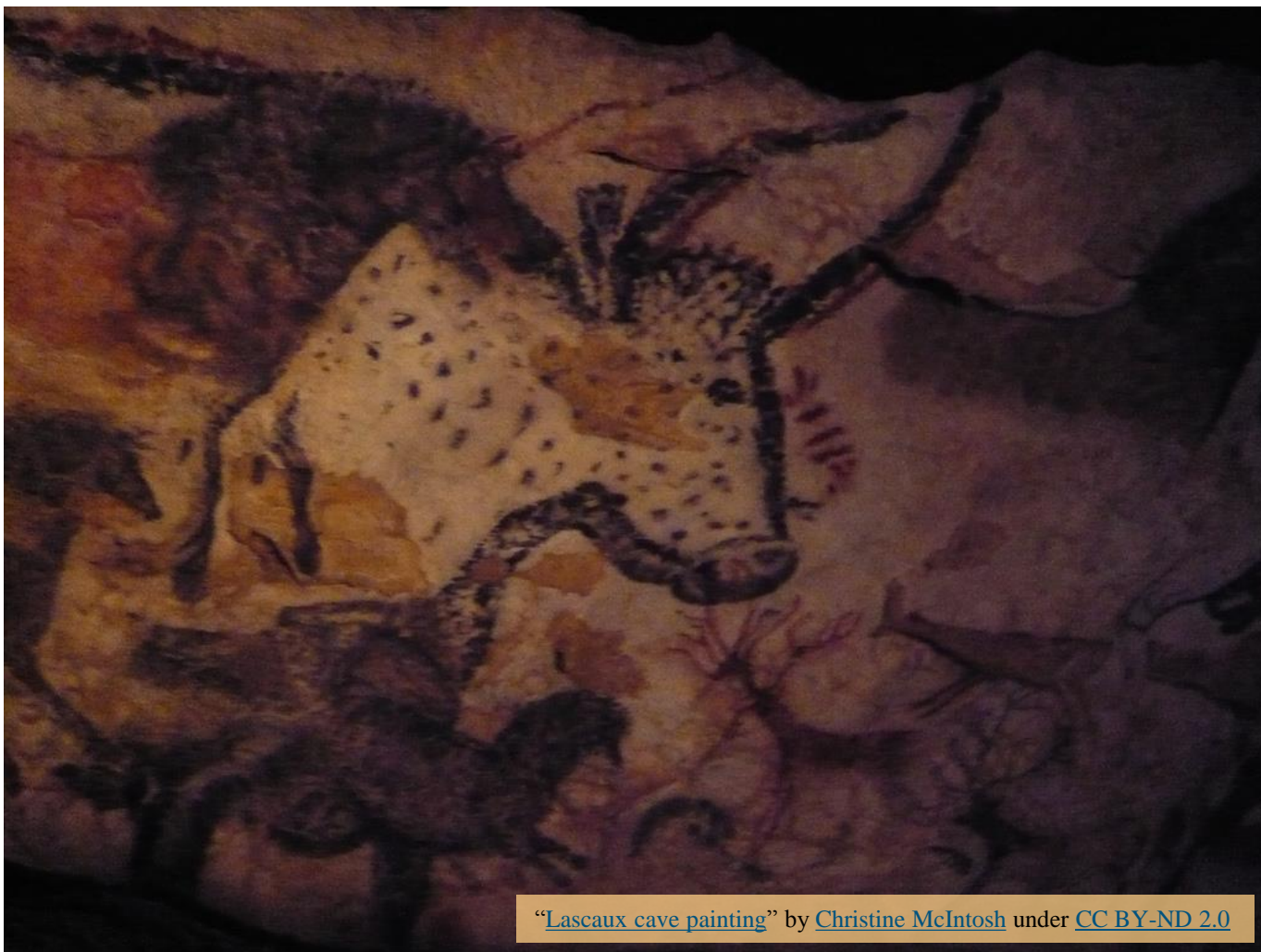
specify HTTP response headers for caching and content encoding



embed metadata, especially character encoding



use durable data formats

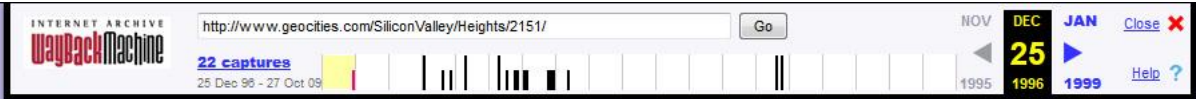


“Lascaux cave painting” by [Christine McIntosh](#) under [CC BY-ND 2.0](#)


prefer responsive design over user-agent personalization




examine your site in the Internet Archive Wayback Machine



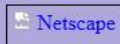


Greetings & Welcome to
A Multidimensional Perception ~/*\=
& PCGuru


[RealAudio 3.0 format](#)


 Free Speech Online
Blue Ribbon Campaign

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”

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

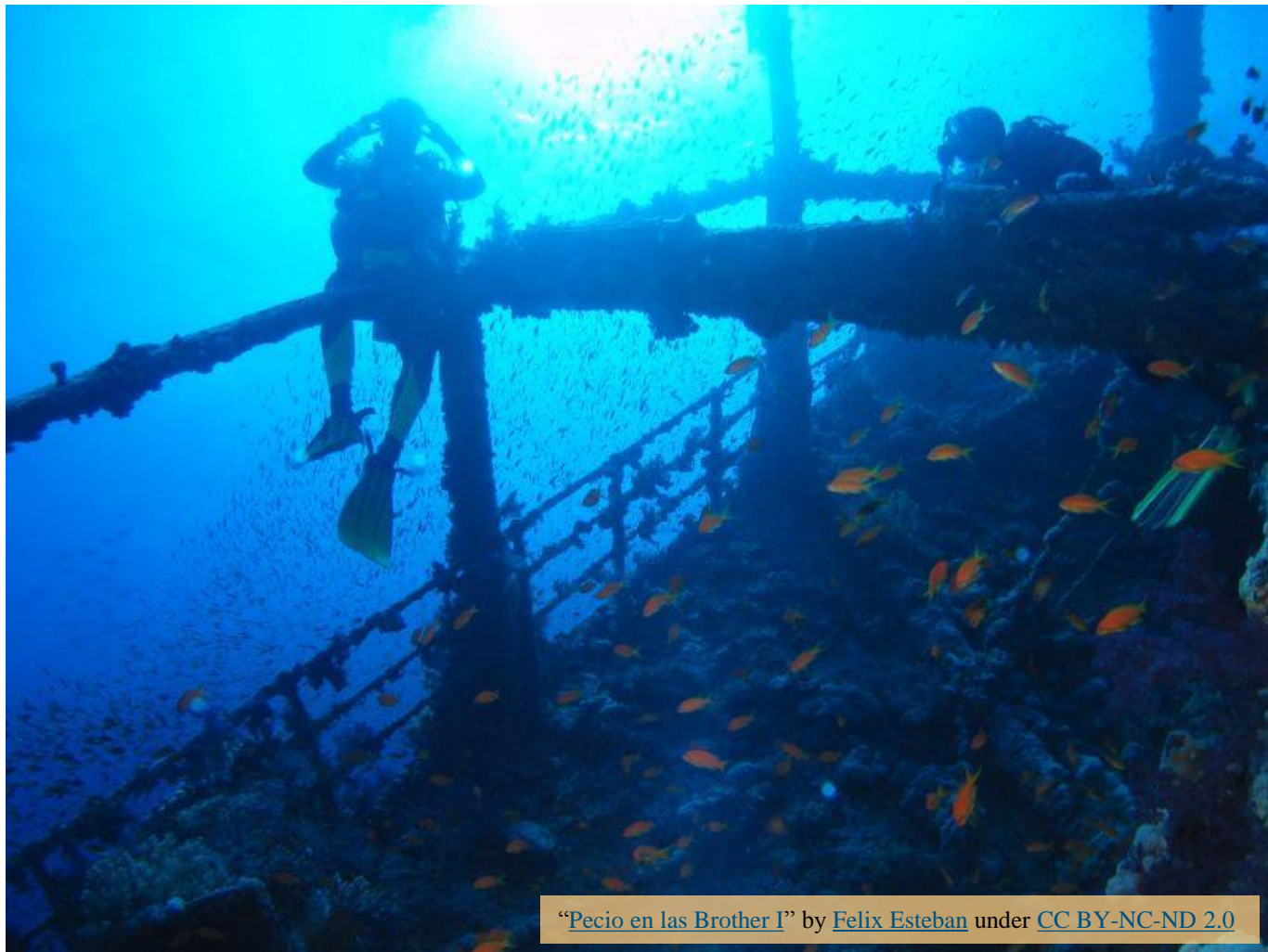
ArchiveReady was created by [Vangelis Banos](#) [🔗](#) (© 2012-2013). It is a **personal research project** and is not supported by any public or private organization.

Archive Ready



Opportunities for **COLLABORATION**

discover content to be archived



“Pecio en las Brother I” by [Felix Esteban](#) under [CC BY-NC-ND 2.0](#)

custom 404 page

```
(function() {
  var s = document.createElement('script');
  s.type = 'text/javascript';
  s.async = true;

  var page_url = window.location.href;
  page_url = page_url.replace('http://www-ilya.', 'http://'); //strip test host
  page_url = page_url.replace('http://www-ilya.', 'http://'); //strip test host
  var wb404_url = 'https://archive.org/wayback/available.php?callback=wb404_callback&url='+page_url;
  s.src = wb404_url;
  var h = document.getElementsByTagName('head').item(0) || document.documentElement;
  h.appendChild(s);

  var l = document.createElement('link');
  l.rel = 'stylesheet';
  l.type = 'text/css';
  l.href = 'https://archive.org/web/wb404.css';
  h.appendChild(l);
})();

wb404_callback = function(obj) {
  var archived_text = "Would you like to <a href='$url' onClick='wb404_record_click(this); return false;'>see an archived version of this page</a> in the Internet Archive's Wayback Machine?";
  var maybe_text = "Would you like to <a href='$url' onClick='wb404_record_click(this); return false;'>check the Internet Archive's Wayback Machine</a> for an archived version of this page?";
  var wb_image = 'https://archive.org/images/wayback404.png';

  if (!obj.archived_snapshots || !obj.archived_snapshots.closest || !obj.archived_snapshots.closest.available) {
    return false;
  }

  var url = obj.archived_snapshots.closest.url;
  var html = "<br/><div class='wb404_imagediv'><a href='"+url+"' onClick='wb404_record_click(this); return false;'><img class='wb404_image' src='"+wb_image +
  "'/></a></div>";

  if (true) {
    html += "<div class='wb404_text'>" + archived_text.replace('$url', url) + "</div><br clear='both'>";
  } else {
    //Not supporting this for now
    html += "<div class='wb404_text'>" + maybe_text.replace('$url', url) + "</div><br clear='both'>";
  }

  var wb404_div = document.getElementById('wb404');
  wb404_div.innerHTML = html;
}

wb404_record_click = function(link) {
  var img = new Image(1,1);
  img.src='https://analytics.archive.org/0.gif?wb404_click=1';
  setTimeout(function(){window.location.href = link.href;}, 100);
}
```



Would you like to [see an archived version of this page](#) in the Internet Archive's Wayback Machine?

<https://archive.org/wayback/available.php>

Internet Archive: "Free "404: File Not Found" Handler for Webmasters to Improve User Experience"

deprecate without deletion



“Banksy Rat Mural on Canal Street, Chinatown, New York City” by [caruba](#) under [CC BY-NC 2.0](#)

archive and redesign



“Layers” by [David Ingram](#) under [CC BY-NC 2.0](#)

thank you!

