



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](#)



[“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



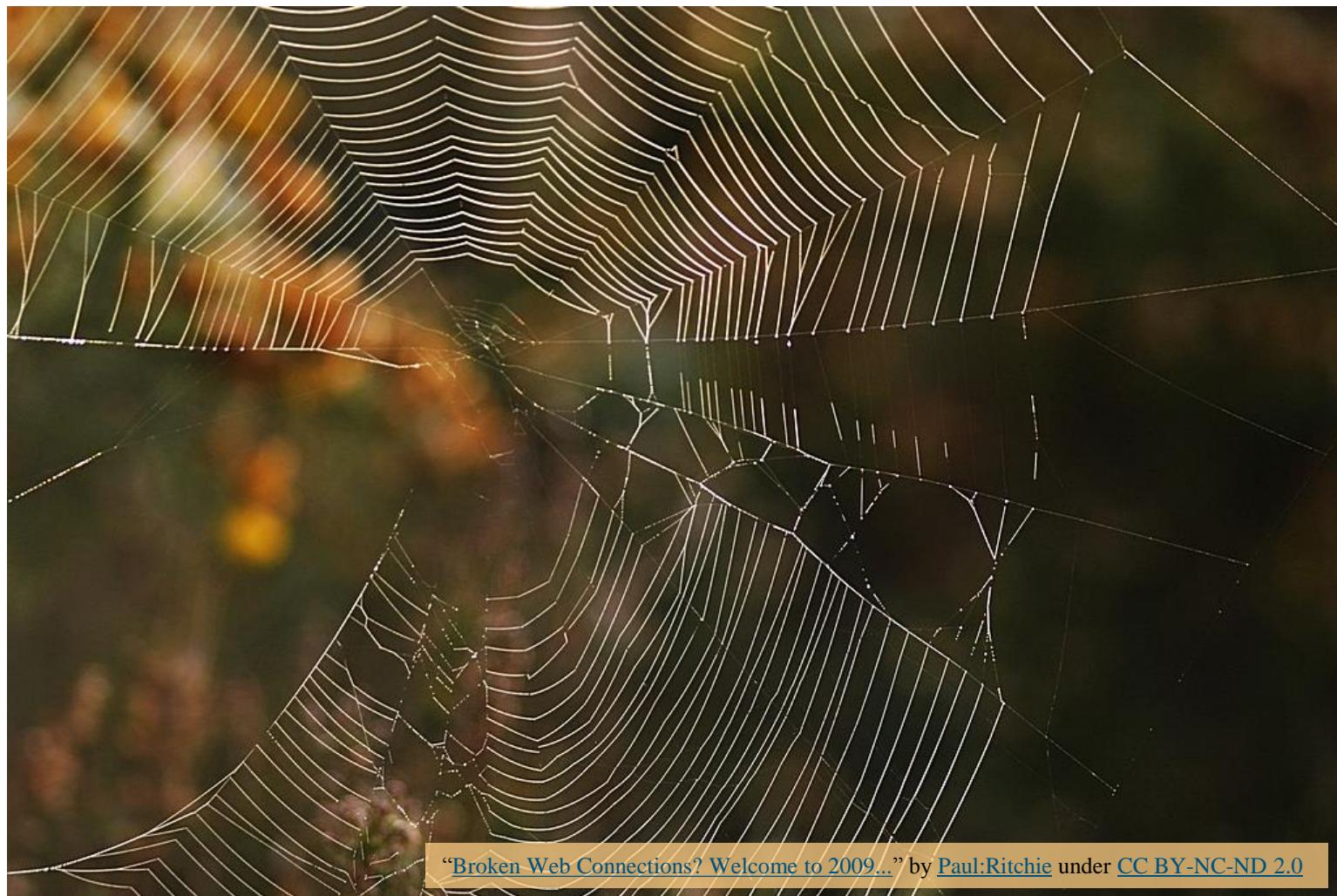
future users are users, too



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



maintain web usability



[“Broken Web Connections? Welcome to 2009...”](#) by [Paul:Ritchie](#) under [CC BY-NC-ND 2.0](#)



improve temporal web usability



<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.



JAN					FEB					MAR					APR				
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30	31	1	2	3	4	5	6	7	8	9
10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
30	31	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	1	2	3	4	5
MAY					JUN					JUL					AUG				
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30	31	1	2	3	4	5	6	7	8	9
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	1	2	3	4
26	27	28	29	30	31	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Internet Archive: "Wayback Machine"

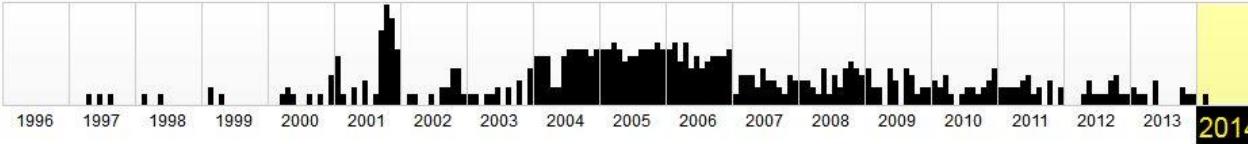


improve temporal web usability

INTERNET ARCHIVE
WayBack Machine 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.



Month	Day
JAN	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
FEB	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
MAR	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
APR	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
MAY	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
JUN	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
JUL	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
AUG	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

Internet Archive: “Wayback Machine”



recover your lost website

The screenshot shows the homepage of the Warrick website. At the top left is a blue square containing a large white 'W' and the word 'WARRICK' below it. To the right is a navigation bar with links: Home (highlighted in green), Recover a Website, Recovery Status, About, Disclaimer, and System Stats. Below the navigation is a row of five blue rectangular buttons with white text: 'RECOVER A WEBSITE' (with 'Start the recovery process' below it), 'RECOVERY STATUS' (with 'Check on recovery status' below it), 'ABOUT WARRICK' (with 'Learn more about Warrick' below it), 'DOWNLOAD' (with 'Download Warrick' below it and a 'tar.gz' file icon to its right), and 'DISCLAIMER' (with 'Please read our disclaimer' below it).

Old Dominion University
Computer Science Department
Norfolk, VA

[“Warrick”](#)



refer to earlier website versions

The image shows a printed page with several horizontal lines of text, each representing a different edit to a Wikipedia page. The edits are listed from oldest at the top to most recent at the bottom. Each edit entry includes the edit number, revision ID, date and time, user, and a brief description of the changes made.

- Edit 1 | Revision 8021313**
18:33, 1 December 2004
Stevertigo
Line 1 : Line 1
#REDIRECT [[Iraq war]]
+
#REDIRECT [[Occupation of Iraq, 2003-2004]]
- Edit 2 | Revision 8404705**
22:10, 13 December 2004
Jmabel
(previous handling of this as a redirect misses
the initial phase of the war)
Line 1 : Line 1
#REDIRECT [[Occupation of Iraq, 2003-2004]]
+
"Wikipedia's main coverage of the 'Iraq War'
that began in [[2003]] is currently divided into
two pages: [[2003 Invasion of Iraq]] and
[[Occupation of Iraq, 2003-2004]], with [[May
2003]] in between the two. This situation is due
to the fact that the occupation began in May 2003,
but the invasion happened earlier in the year."
- Edit 5 | Revision 11625864**
23:20, 28 March 2005
Stevertigo
(update to current - its inferior to refer to past
events (2 years ago) instead of current and more
relevant events)
Line 1 : Line 1
#REDIRECT [[Occupation of Iraq, 2003-2004]]
+
#REDIRECT [[Post-invasion Iraq, 2003-2005]]
- Edit 6 | Revision 12045239**
18:05, 8 April 2005
Trilobite
Line 1 : Line 1
#REDIRECT [[Post-invasion Iraq, 2003-2005]]
+
#REDIRECT [[Iraq war (disambiguation)]]
- Edit 7 | Revision 15323404**
03:15, 17 June 2005
Stevertigo
Line 1
"The Iraq War: Wikipedia Historiography" by [STML](#) under [CC BY-SA 2.0](#)



archivability is the new accessibility





institutional history

 STANFORD
UNIVERSITY 



ABOUT STANFORD

- ▶ [Stanford Facts](#)
- ▶ [History](#)
- ▶ [President's Office](#)

TEACHING & RESEARCH

- ▶ [Schools](#)
- ▶ [Hoover SLAC Centers](#)
 - ▶ [Libraries](#)
 - ▶ [Medical Center](#)

ADMISSIONS

- ▶ [Undergraduate](#)
- ▶ [Graduate](#)
- ▶ [Course Catalog, Time Schedule](#)

ALUMNI/ARTS/ATHLETICS

- ▶ [Alumni](#)
- ▶ [Events, Arts, Tickets](#)
- ▶ [Athletics](#)

NEWS & INFORMATION

- ▶ [News Service](#)
- ▶ [Calendar of Events](#)
- ▶ [Directories](#)

STUDENTS

- ▶ [Activities](#)
- ▶ [Organizations](#)
- ▶ [Residences](#)
- ▶ [ASSU Directory](#)

BUSINESS/ADMINISTRATION

- ▶ [Administration](#)
- ▶ [Business Organizations](#)
- ▶ [The Portfolio Collection](#)

TECHNOLOGY/COMPUTING

- ▶ [Info Tech](#)
- ▶ [Internet Archive Wayback Machine: "Stanford University Homepage"](#)
- ▶ [Commis](#)



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

Stanford | Office of Audit, Compliance and Privacy

Search this site... 

Home Internal Audit Services Compliance and Ethics Privacy Helpline Staff / Contact Us

Compliance and Ethics

Stanford | office of the General Counsel

Home
Contact Us
Attorney-Client Privilege
Retention of Counsel
Stanford Legal Facts
University FAQs
Hospital Legal Information
Stanford Links
External Legal Resources
Recent Legal Developments

Welcome to Stanford's Office of the General Counsel. Our mission is to facilitate compliance and ethics. In addition to legal services, we provide links to Stanford resources, post recent legal developments, and our community can comment on legal matters related to the University.

The Office of the General Counsel handles issues arising out of the University's Clinics and Clinics and Lucile Salter Packard Children's Hospital. The General Counsel, Debra L. Zumwalt, the office's director, is available for partnering with outside law firms.

The mission of the OGC is to educate our Students, faculty, and staff about operations, to solve legal problems and facilitate representation to help our clients fulfill their missions.

RECENT LEGAL DEVELOPMENT
NLRB Ruling related to Northwestern Football
The regional director of the National Labor Relations Board has ruled that Northwestern University football players

PROVIDING LEGAL SERVICES TO
 STANFORD UNIVERSITY

BYSTANDER
REPORT YOUR CONCERN
RELATE A RELATED TO:
Time / Reporting
of Interest and

My Account Feedback Staff login

Search 

STANFORD UNIVERSITY LIBRARIES

About Libraries Using the libraries Collections Research support Academic technology Ask us Search

Home Libraries Special Collections Managing university records

At a glance

Special Collections & University Archives

Managing university records

Records created by Stanford affiliates are one of the University's most valuable assets. Among other things, records support administrative decision-making and operations, demonstrate compliance, and document Stanford's institutional history. Just like other University assets, records need to be properly managed. These web pages are designed to provide basic records management guidelines to help you:

- Effectively organize and maintain records
- Improve efficiency and access to information
- Comply with legal obligations
- Ensure that historical records are captured and maintained in perpetuity

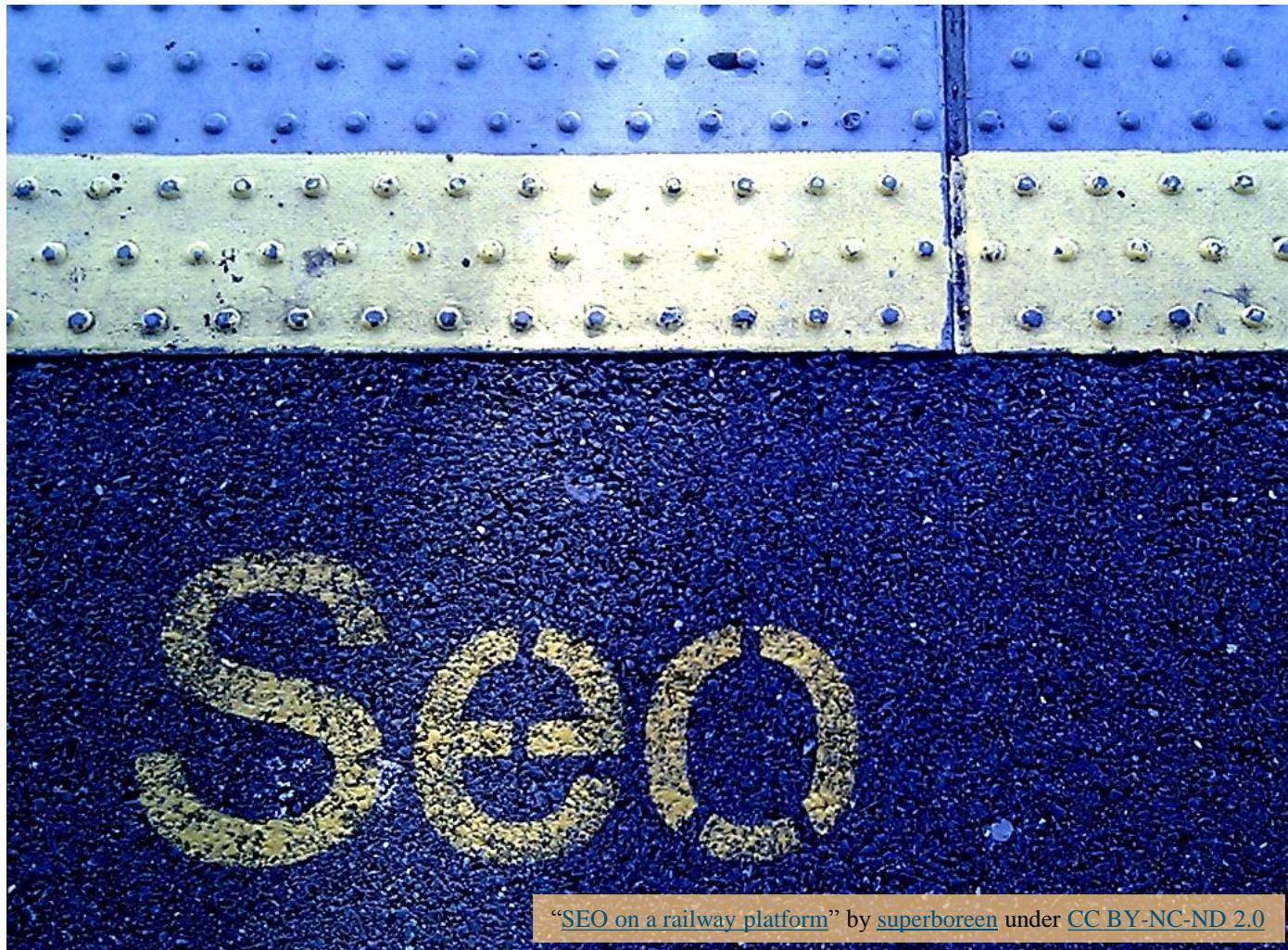
For help getting started please contact us at archivesref@stanford.edu.

Sources and references

- Beinecke Rare Book and Manuscript Library: [Authors' Guidelines for Preserving Digital Archives](#)
- Born-Digital Collections: An Inter-Institutional Model for Stewardship (AIMS): [white paper](#)
- Harvard University Archives Record Management Services
- The InterPARES Project: Creator Guidelines
- Smithsonian Institution Archives: [Electronic Records: Responsible Record-Keeping](#)
- University of Michigan: [Deep Blue Preservation and Format Support Policy](#)



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](#)



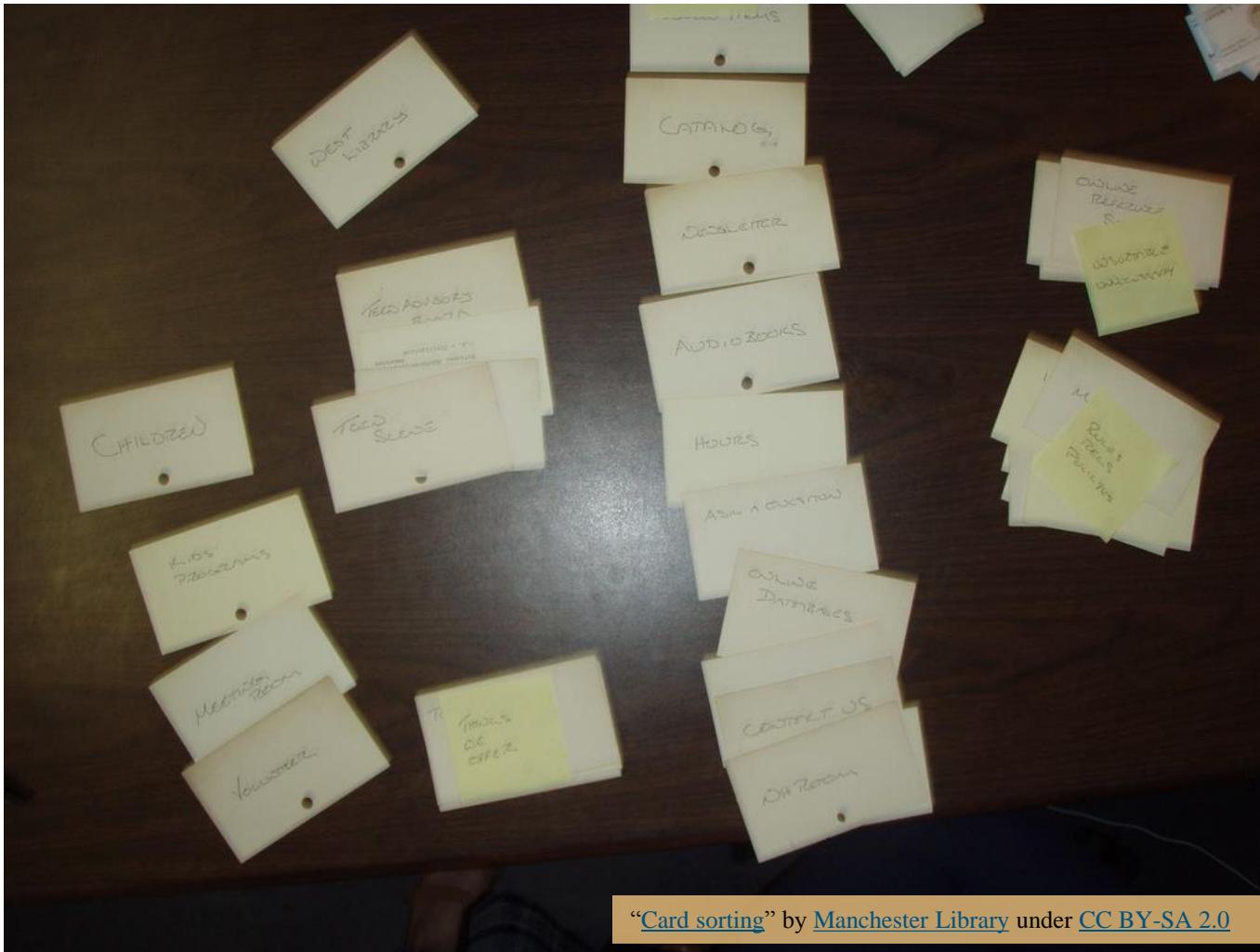
follow web standards and accessibility guidelines



[“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





use semantically-meaningful URLs

The screenshot shows a web browser window with the following details:

- URL Bar:** http://w63.w63.org
- Content Area:** m9(^Д^)フキヤ——ツ
Welcome to w63.w63.org!
- Network Tab:** Shows a single request: GET /w63.org. The status is 200 OK.
- Response Headers:** Includes 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, and Vary: Accept-Encoding.
- Request Headers:** Includes 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: www.w63.w63.org, and User-Agent: Mozilla/5.0 (Windows NT 6.1; WOW64; rv:24.0) Gecko/20100101 Firefox/24.0.
- Bottom Status Bar:** Shows the URL w63.w63.org, 0 B, and 0ms (on.lol).



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 aspilicious: 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 # Files
29 Disallow: /CHANGELOG.txt
30 Disallow: /cron.php
31 Disallow: /INSTALL.mysql.txt
32 Disallow: /INSTALL.pgsql.txt
33 Disallow: /INSTALL.sqlite.txt
34 Disallow: /install.php
35 Disallow: /INSTALL.txt
36 Disallow: /LICENSE.txt
37 Disallow: /MAINTAINERS.txt
38 Disallow: /update.php
39 Disallow: /UPGRADE.txt
40 Disallow: /xmiproxy.php
```

Disallow: /scripts/
Disallow: /themes/

“drupal/robots.txt at 7.x”

The screenshot shows the GitHub interface for the Drupal repository. The user is viewing the 'robots.txt' file under the '7.x' branch. The file contains standard directives for search engines, including 'Disallow' rules for '/scripts/' and '/themes/'. These two lines are highlighted with a red box and a red arrow points from them to a callout bubble. The callout bubble contains the text 'Disallow: /scripts/' and 'Disallow: /themes/'. Another callout bubble at the bottom right of the code area indicates the specific commit '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

Search this site...



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



minimize reliance on external assets necessary for presentation

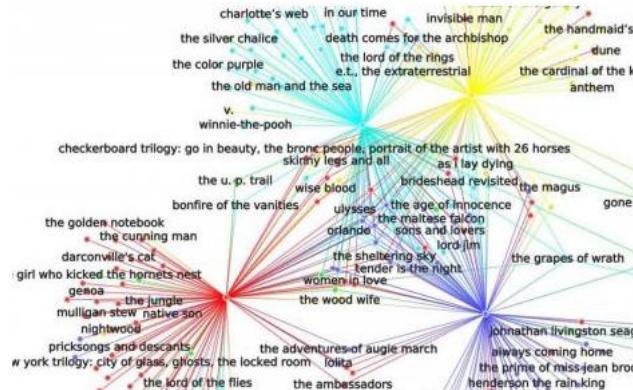
Stanford University

SUNetID Login

Stanford
Department
of English

Home About People Courses Degree Programs Department Bookshelf News
Events

Search this site...



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

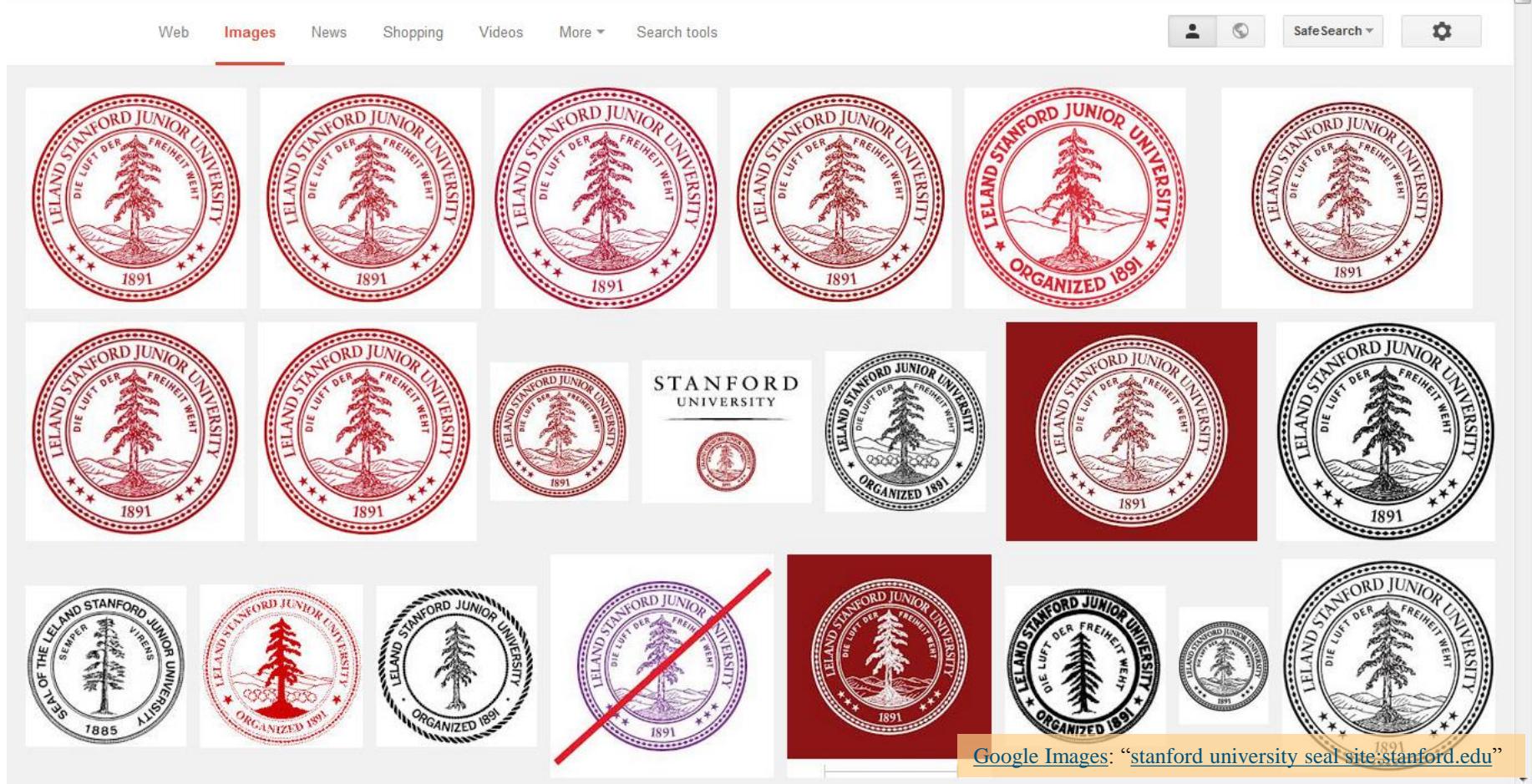
Creative Writing Program
["Stanford Department of English"](#)



serve reusable assets from a single, common location

Google +Nicholas 

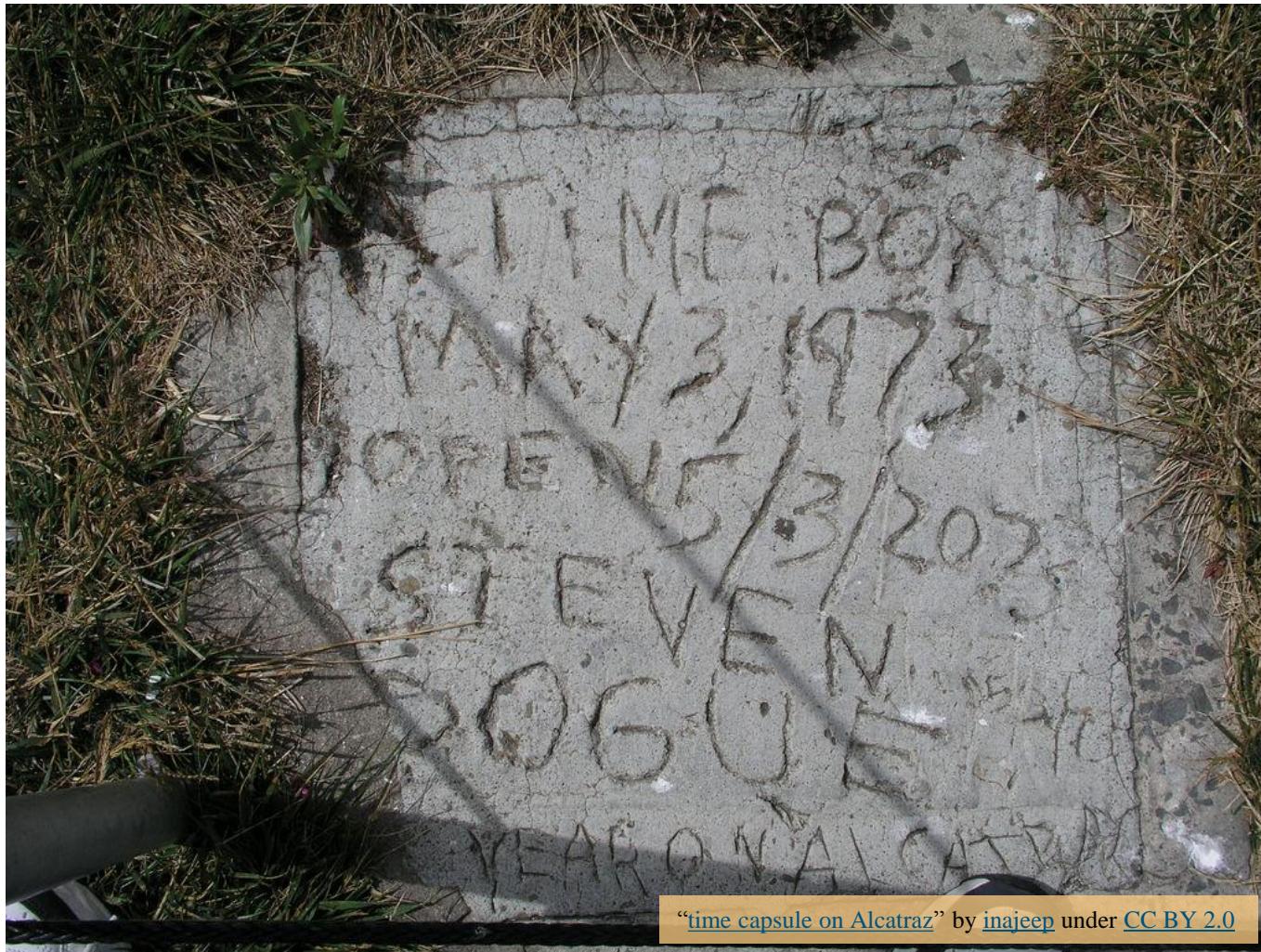
Web Images News Shopping Videos More ▾ Search tools



Google Images: "stanford university seal site:stanford.edu"

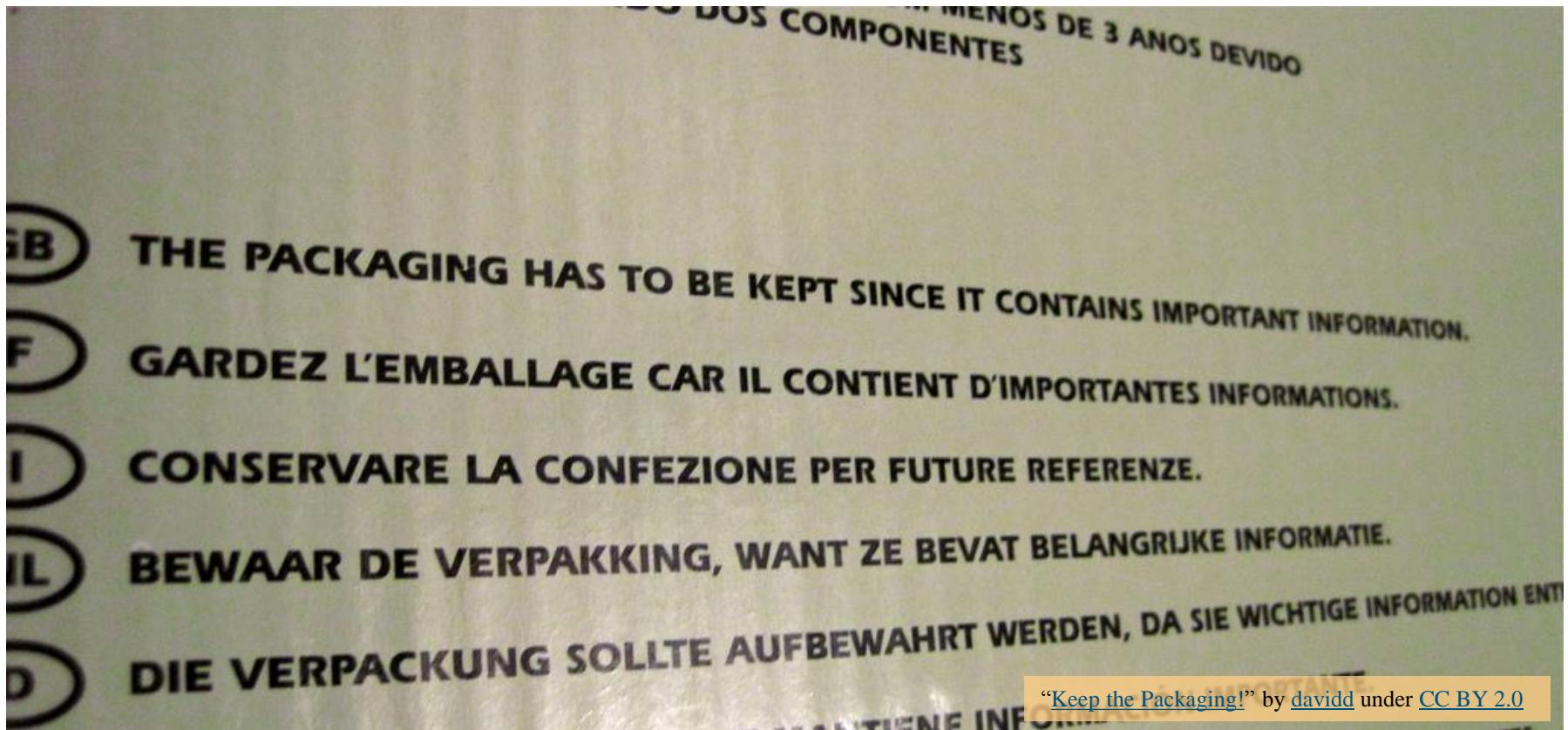


specify HTTP response headers for caching and content encoding



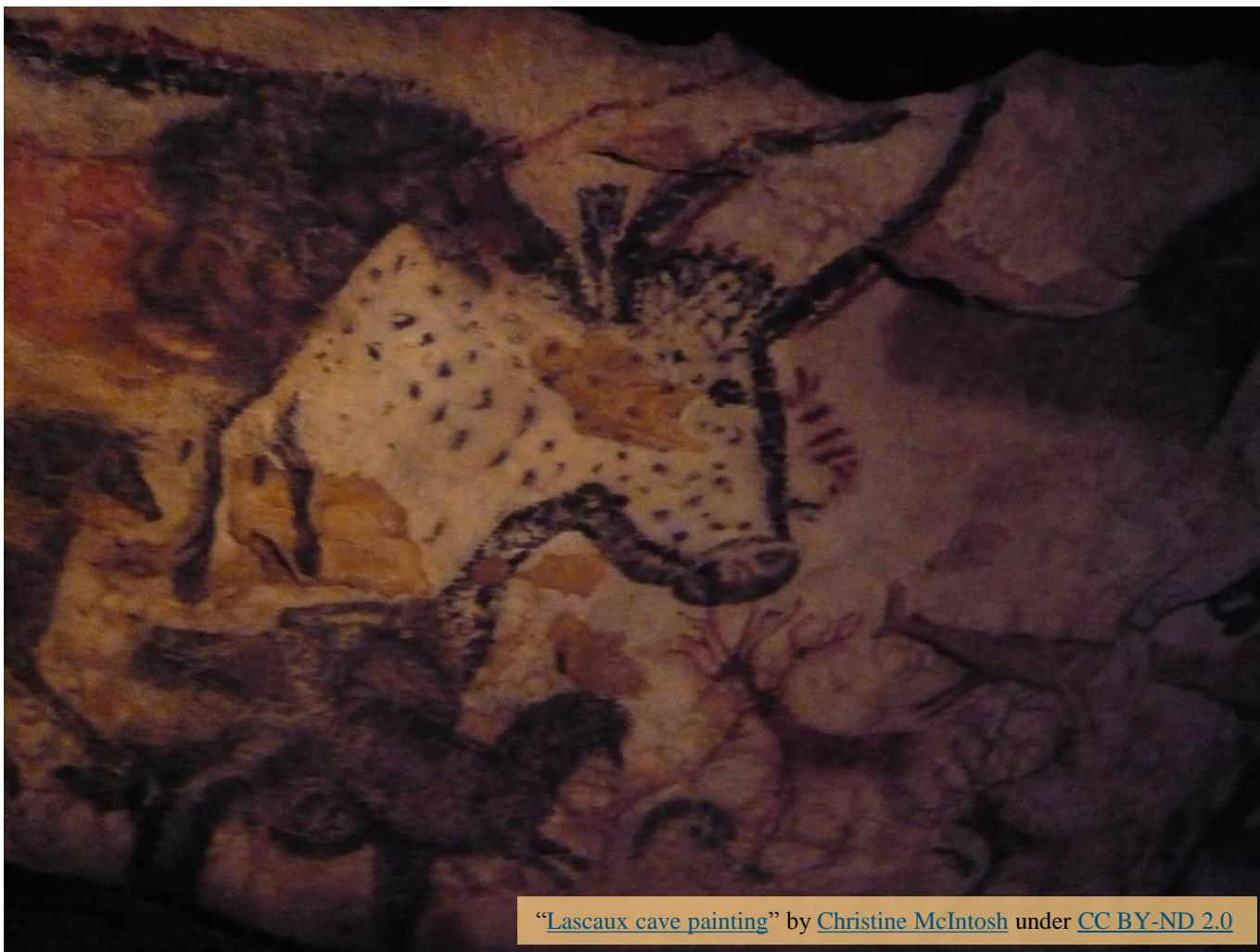


embed metadata, especially character encoding





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

INTERNET ARCHIVE
WayBack Machine

http://www.geocities.com/SiliconValley/Heights/2151/ Go NOV DEC JAN Close X
22 captures 25 Dec 96 - 27 Oct 09 1995 1996 1999 Help ?

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

Netscape Microsoft 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.
- 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

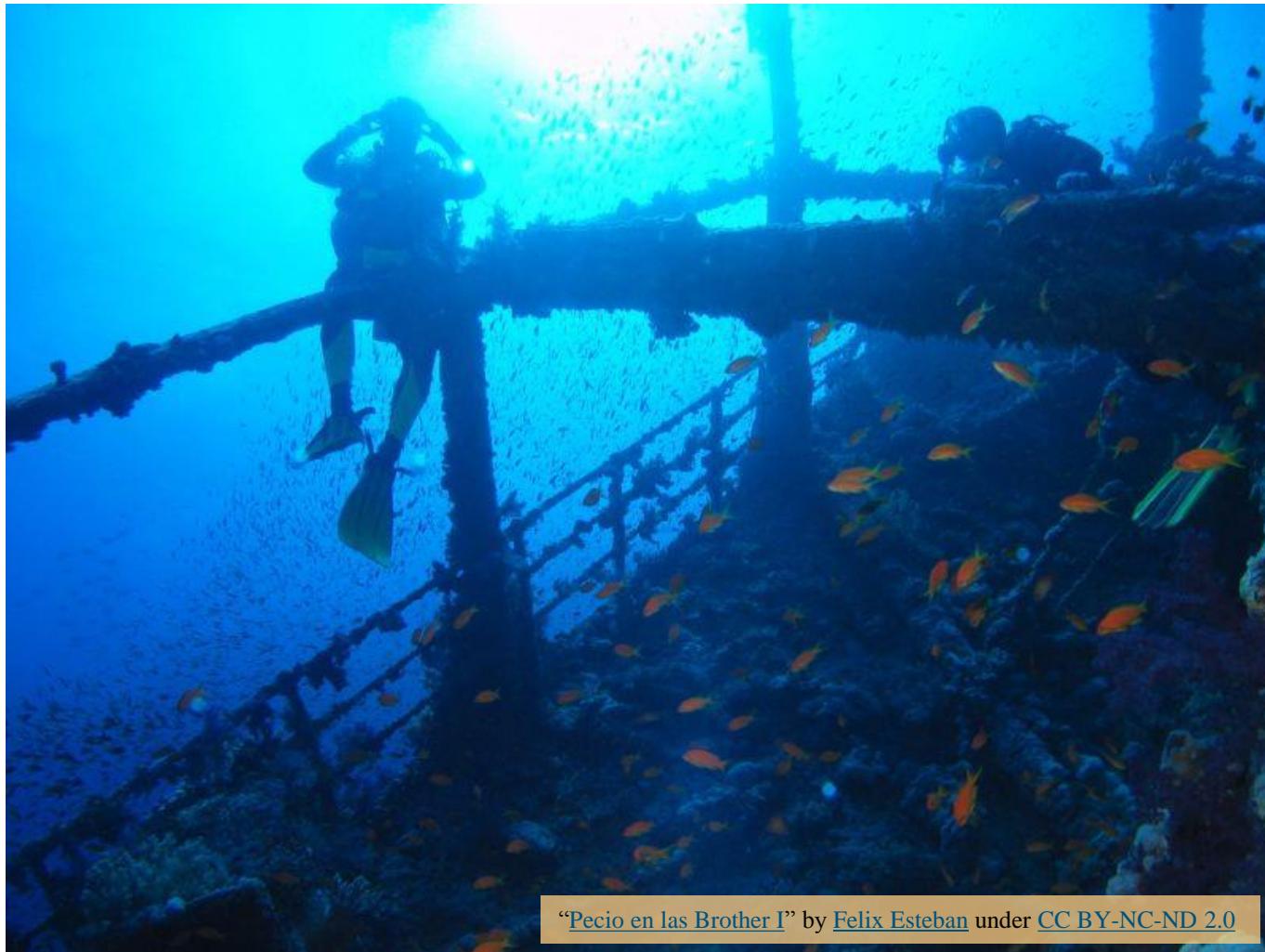
[About](#)

ArchiveReady was created by [Vangelis Banos ↗](#) (© 2012-2013). It is a personal research project and ["Archive Ready"](#)





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('://www-ilya.', '//'); //strip test host
    page_url = page_url.replace('://www-ikumar.', '//'); //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 Internet Archive: "Free "404: File Not Found" Handler for Webmasters to Improve User Experience" |
    setTimeout(function(){window.location.href = link.href}, 100);
}
```



deprecate without deletion



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



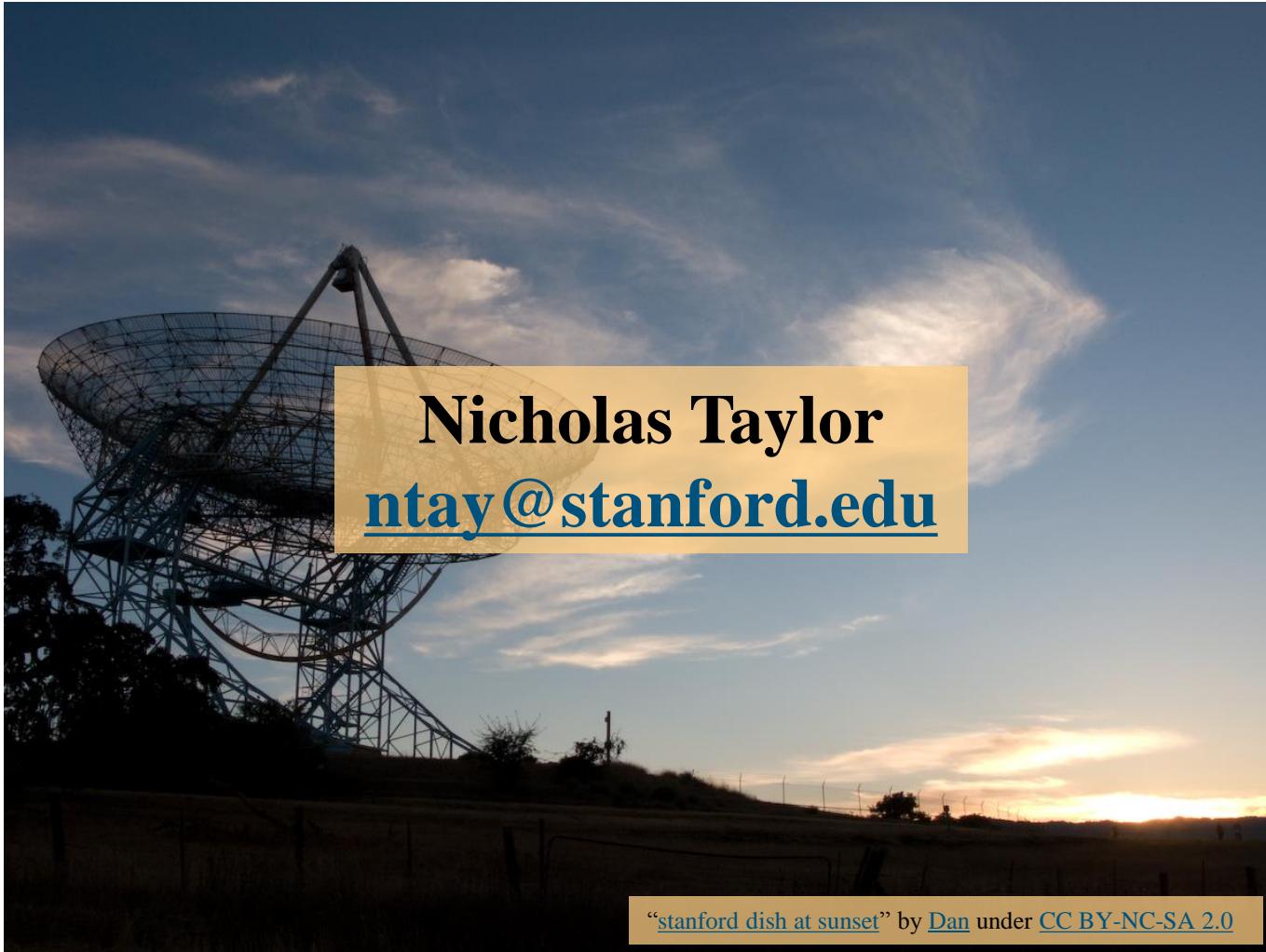
archive and redesign



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



thank you!



Nicholas Taylor
ntay@stanford.edu

“[stanford dish at sunset](#)” by [Dan](#) under [CC BY-NC-SA 2.0](#)