Outreach to Campus Webmasters for a Better Web, and Better Web Archiving

Nicholas Taylor
Web Archiving Service Manager
Stanford University Libraries

Archives 2015 Session 306 – Seeding Engagement: Web Archiving Outreach Opportunities and Strategies
August 21, 2015
webmasters don’t care about web archiving
webmasters more concerned w/ Web-mastery
look for points of alignment
decentralized Web production

- Stanford Web Services
- Stanford Communications
- tens of distributed webmasters for administrative units
- many more individual content creators
- no central coordination
- centralized services available (Drupal, AFS) but diversity of platforms used

“hectormilla.com” by Hector Milla under CC BY-NC 2.0
become a free SEO consultant
discoverable links
planned site structure
responsive design
reusable assets
caching headers

“time capsule on Alcatraz” by inajeepe under CC BY 2.0
reduce reliance on external assets
reduce reliance on external assets
partnership opportunities
Archivability guidance

Archivability

Web builders already design for accessibility, performance, search engine optimization, standards compliance, and usability. An emerging best practice is to design also for archivability. Archivability refers to the ease with which the content, structure, functionality, and front-end presentation(s) of a website can be preserved and later re-presented, using contemporary web archiving tools.

Many organizations are now engaged in preserving many different types of content from the Web. The cohort of users of the temporal web can only grow with time, underscoring the breadth of its stakeholders and the outsized impact of investments to improve the usability of archived resources. As a web builder, you can help facilitate web archiving efforts by third parties or your own organization by following the archivability recommendations presented in the sub-pages accessible through the links in the left menu.

In addition to improving the usability of the temporal web and helping to ensure the preservation of collective cultural heritage, designing for archivability will also tend to optimize your website for access by other crawlers, boost website performance, enhance usability for contemporary users, and improve the likelihood that you’ll be able to refer to and/or recover historical versions of your own web content.

Stanford University Libraries: “Archivability”
test crawls for content assessment
more complete crawls

Up next, we are working with Psych department

Ok, I'll crawl Psych

The requested test crawl is complete. You can find reports after logging in

Thanks so much; the crawl is very helpful

I'm not sure we've got all the Psych content

In Drupal db I see ~1700 nodes, but only 437 text/html in report

Thx for feedback. Are nodes behind webauth? I also noticed 'psych' subdomain was out of scope - corrected

News items seem not to have been picked up (e.g. node/3715 or node/1915)
locate deprecated but archived content

WorldWideWeb SLAC Home Page

SLAC 7 Sep 1994

Use the WorldWideWeb (WWW) service to gain access to a wide range of information at SLAC and elsewhere around the globe. Emphasized text like this is a hypertext link.

You may view WWW information through GUI or line-mode browsers. At least most SLAC pages have been tested on the MidasWWW X Window System browser.

Note that over time links may move around on a page, migrate to other pages, or be removed entirely, due to the dynamic nature of the Web. That there is no current Hot Topic illustrates this characteristic today.

SLAC Information

People and organizations:
  people at SLAC, particle physics people and institutions, ESnet white pages.
Library:
  SPIRES-HEP, Current PPF-list, Books, SLACspeak glossary, Library News, more databases.
Physics Preprint Bulletin Boards (full-text postscript):
  today, yesterday, last seven days, week before that, let me search.
Seminars:
  today, tomorrow, this week, next week, this year, let me search.
Conferences:
  this month, next month, next summer, next year, all future, let me search.
News:
  APS What's New, SLAC News,

SWAP: “WorldWideWeb SLAC Home Page”
take timely snapshots
Archive-It crawler isn’t malicious
contacts to unblock robots
custom 404 page

Internet Archive: “Free "404: File Not Found" Handler for Webmasters to Improve User Experience”
transactional web archiving

![Diagram of SiteStory web archiving process]

HTTP Request

HTTP Response

Apache Web Content Server

SiteStory Web Archive

HTTP PUT via mod_sitestory

File Storage

BDB Index

Memento HTTP Request

Memento HTTP Response

“SiteStory Web Archive”
webmasters don’t care about web archiving
webmasters *may* care about web archiving