why preserve the web?

World Wide Web

The World Wide Web (W3) is a wide area hypertext 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 of the project, Mailing lists, Policy, November's W3 news, Frequently Asked Questions.

Technical
- Details of protocols, formats, program internals etc
- Bibliography
  - Paper documentation on W3 and references.

People
- A list of some people involved in the project

History
- A summary of the history of the project

How can I help?
- If you would like to support the web.

Getting code
- Getting the code by anonymous FTP, etc.
web archivists aren’t visible stakeholders
search engine crawler ≠ archival crawler

“GoogleBots” by Flickr user ares64 under CC BY 2.0
what is a “preservable” website?
three priorities:

- **capture**: can resources be acquired by current web archiving technologies?
- **replay**: can the user’s experience of the original website be recreated from the archived resources?
- **preservation**: how can it be assured that the archived website remains coherent over time?
follow web standards and accessibility guidelines

“The nice thing about web standards is, there are so many to choose from.”

“Web Standards Fortune Cookie” by Flickr user mherzber under CC BY-SA 2.0
be careful with robots.txt exclusions

User-Agent: *
Disallow: /music?
Disallow: /widgets/radio?
Disallow: /affiliate/
Disallow: /affiliate_redirect.php
Disallow: /affiliate_sendto.php
Disallow: /affiliatelink.php
Disallow: /campaignlink.php
Disallow: /delivery.php
Disallow: /music/+noredirect/
Disallow: /harming/humans
Disallow: /ignoring/human/orders
Disallow: /harm/to/self
Allow: /
use a site map, transparent links, and contiguous navigation
maintain stable URLs and redirect when necessary
consider using a Creative Commons license

“2500 Creative Commons Licenses” by Flickr user qthomasbower under CC BY-SA 2.0
use durable data formats

“Lascaux cave painting” by Flickr user qoforchriss under CC BY-ND 2.0
embed metadata, especially the character encoding

source code of http://www.seo.com/
use archiving-friendly platform providers and CMSs
three tips

1. see how well your site validates on http://validator.w3.org/
2. see how your site looks on http://archive.org/
3. your favorite online sitemap generator is a good starting point
thank you!

Nicholas Taylor
@nullhandle