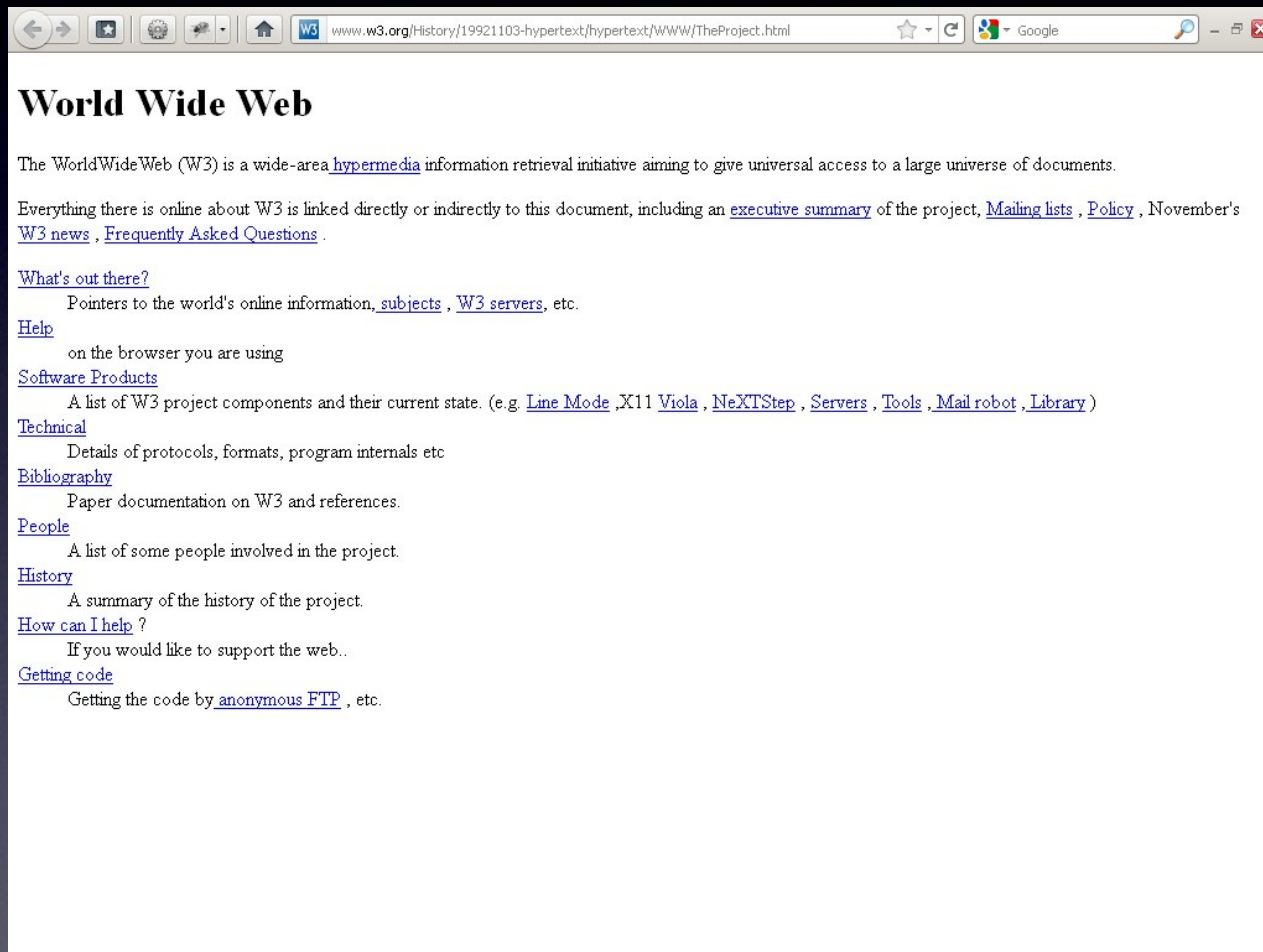


Designing Preservable Websites

Nicholas Taylor
[@nullhandle](#)

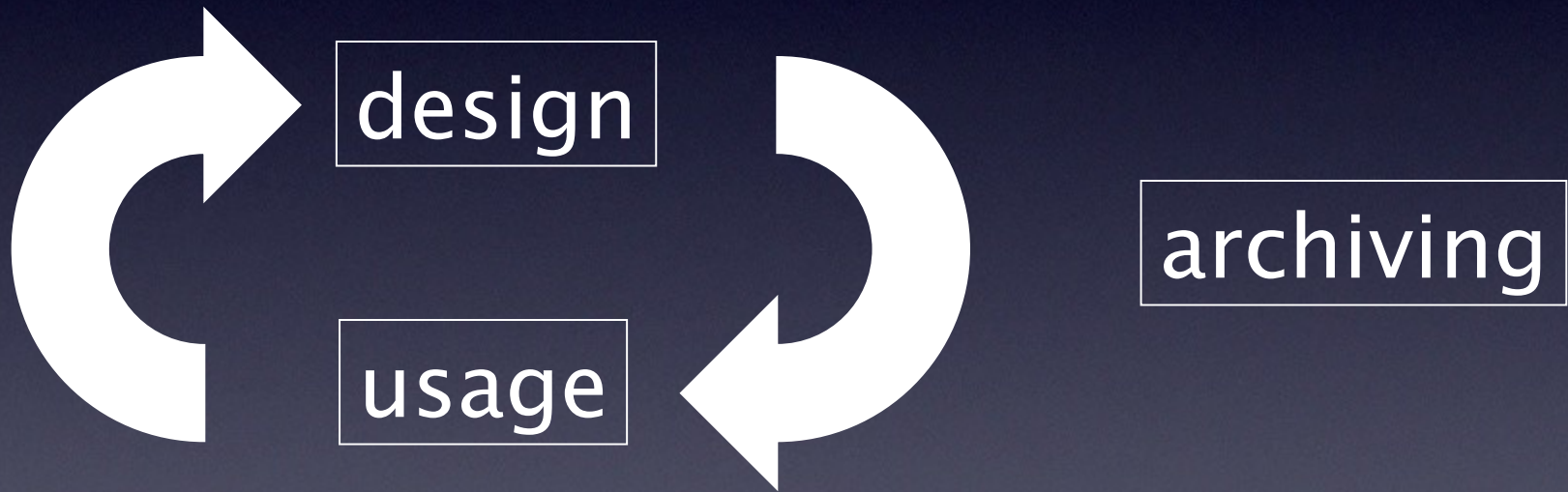
DC, VA & MD Search Engine Marketing Meetup
July 18, 2012

why preserve the web?

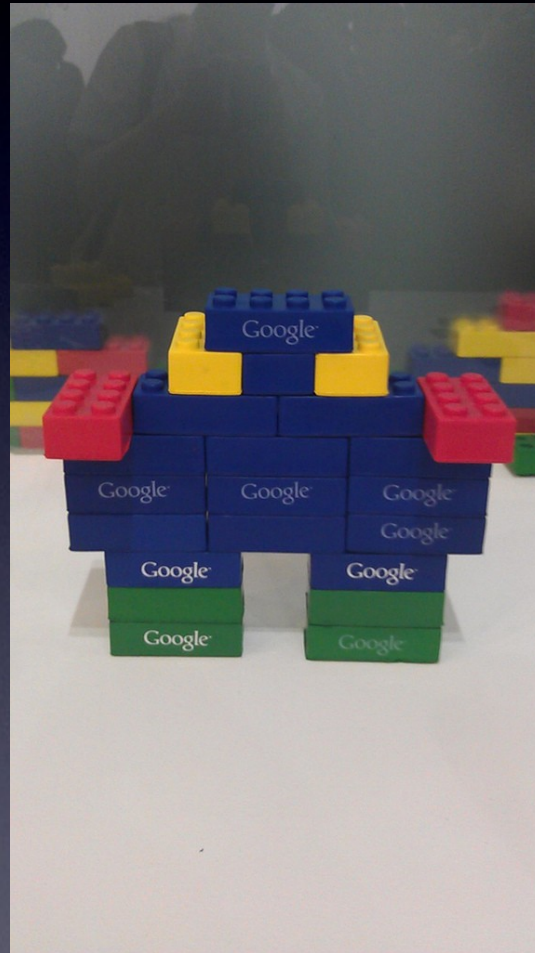


copy of the first webpage

web archivists aren't visible stakeholders



search engine crawler \neq archival crawler



"GoogleBots" by Flickr user [ares64](#) under [CC BY 2.0](#)

what is a “preservable” website?



“Fish Preserver” by Flickr user [ecstaticist](#) under CC BY-NC-SA 2.0

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

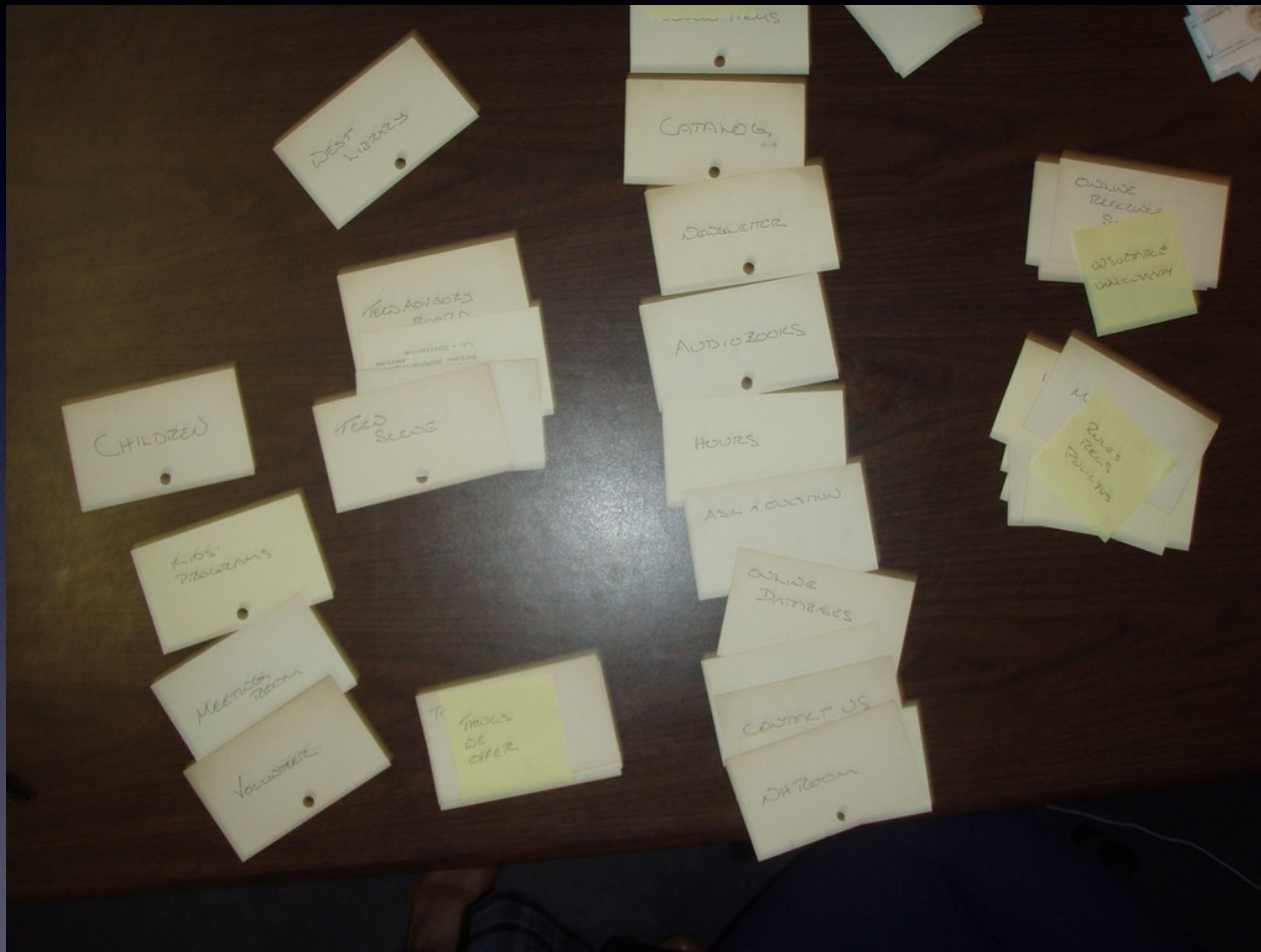


"Web Standards Fortune Cookie" by Flickr user [mherzber](#) under [CC BY-SA 2.0](#)

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



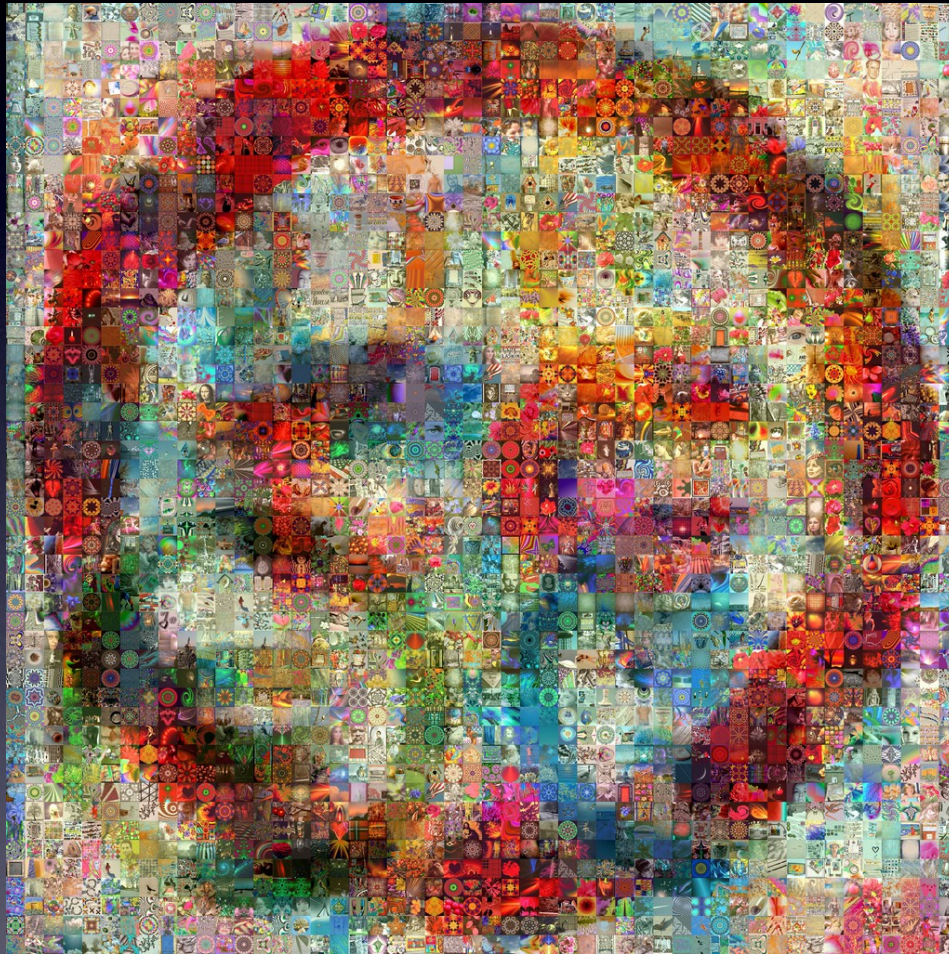
“Card sorting” by Flickr user [Manchester Library](#) under CC BY-SA 2.0

maintain stable URLs and redirect when necessary



“Improvised detour sign” by Flickr user [Jason McHuff](#) under [CC BY-SA 2.0](#)

consider using a Creative Commons license



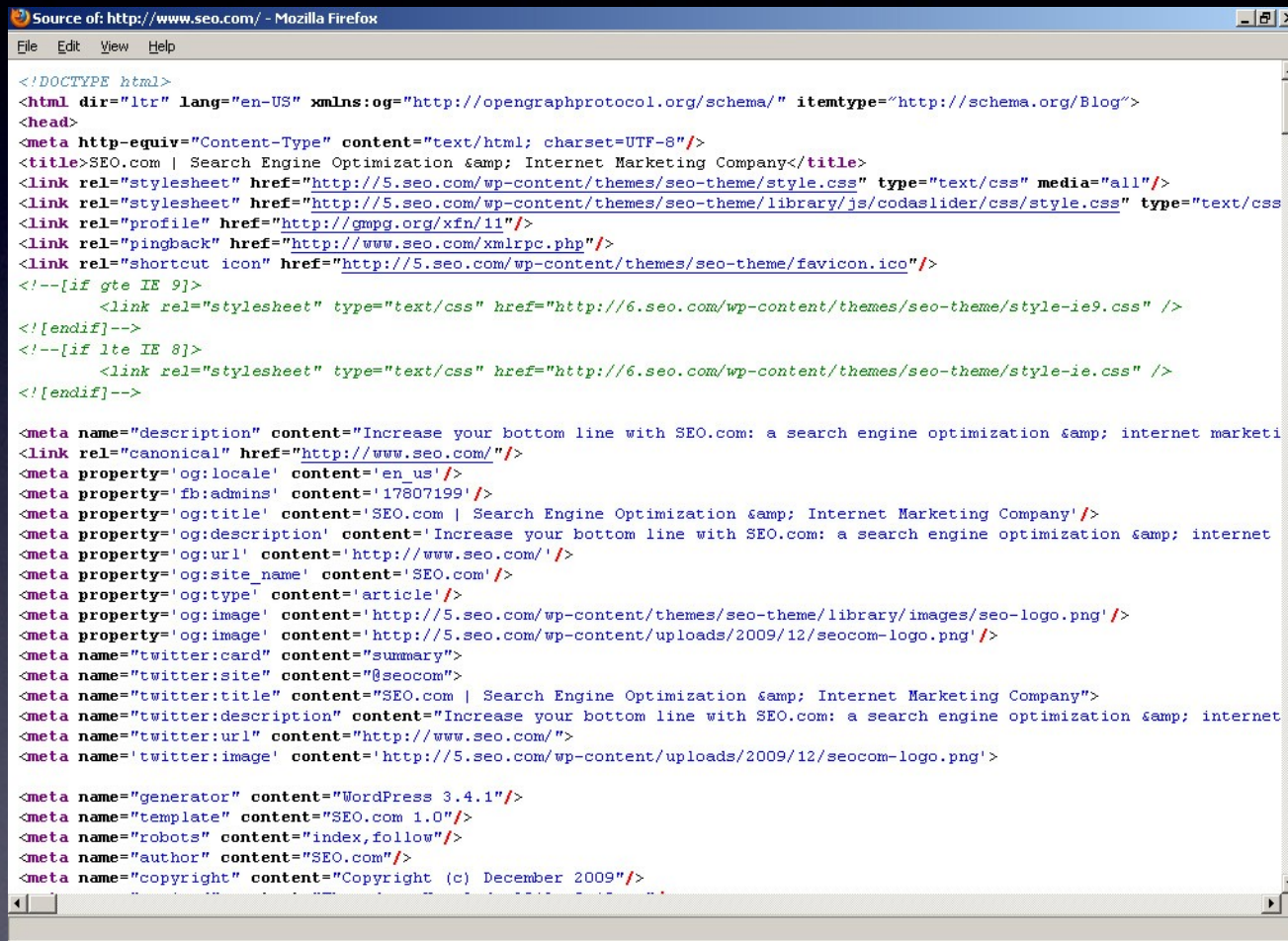
"2500 Creative Commons Licenses" by Flickr user qthomasbower under CC BY-SA 2.0

use durable data formats



“Lascaux cave painting” by Flickr user [qoforchris](#) under CC BY-ND 2.0

embed metadata, especially the character encoding



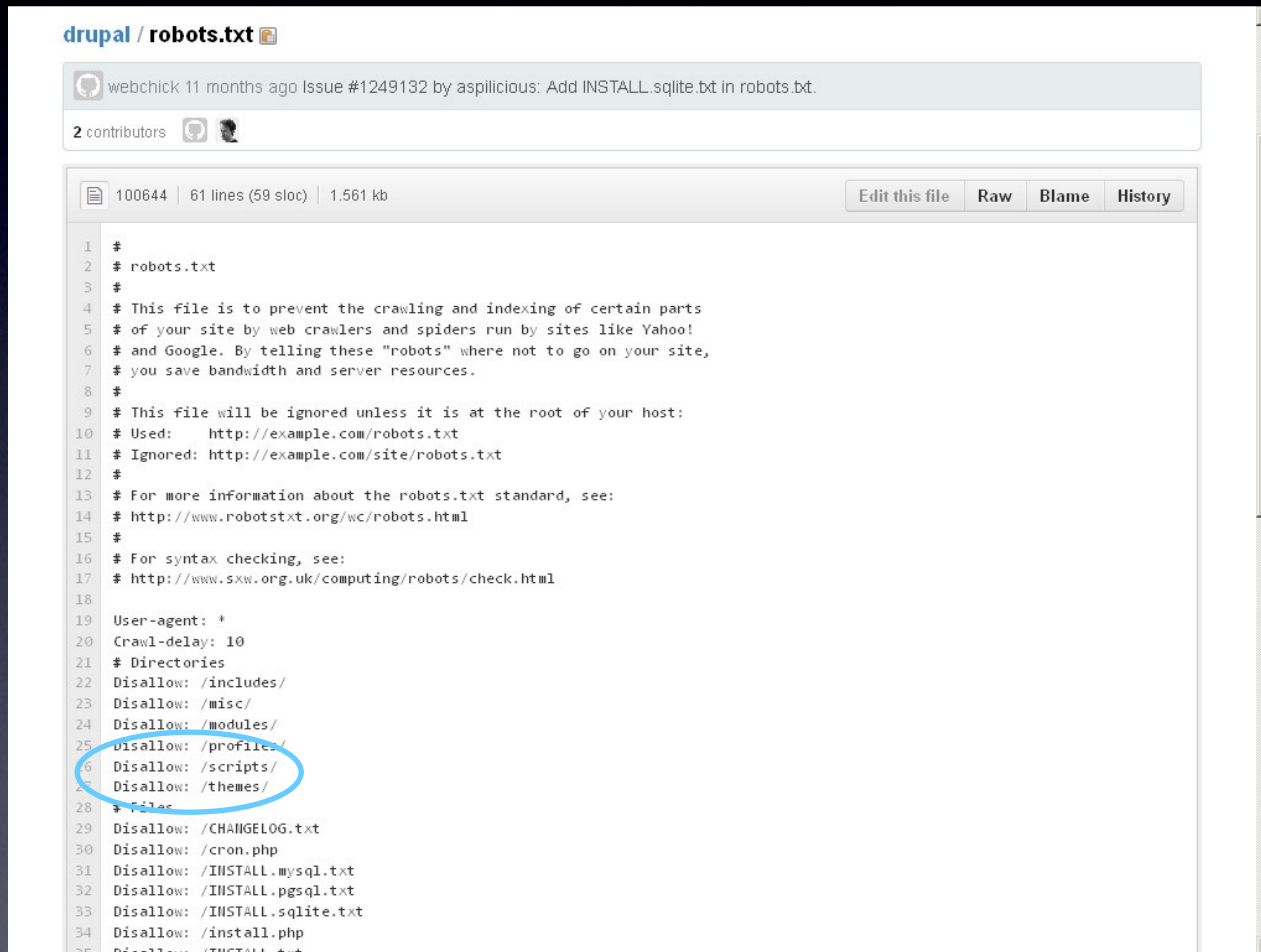
```
<!DOCTYPE html>
<html dir="ltr" lang="en-US" xmlns:og="http://opengraphprotocol.org/schema/" itemType="http://schema.org/Blog">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8"/>
<title>SEO.com | Search Engine Optimization &amp; Internet Marketing Company</title>
<link rel="stylesheet" href="http://5.seo.com/wp-content/themes/seo-theme/style.css" type="text/css" media="all"/>
<link rel="stylesheet" href="http://5.seo.com/wp-content/themes/seo-theme/library/js/codaslider/css/style.css" type="text/css" media="all"/>
<link rel="profile" href="http://gmpg.org/xfn/11"/>
<link rel="pingback" href="http://www.seo.com/xmlrpc.php"/>
<link rel="shortcut icon" href="http://5.seo.com/wp-content/themes/seo-theme/favicon.ico"/>
<!--[if gte IE 9]>
<link rel="stylesheet" type="text/css" href="http://6.seo.com/wp-content/themes/seo-theme/style-ie9.css" />
<![endif]-->
<!--[if lte IE 8]>
<link rel="stylesheet" type="text/css" href="http://6.seo.com/wp-content/themes/seo-theme/style-ie.css" />
<![endif]-->

<meta name="description" content="Increase your bottom line with SEO.com: a search engine optimization &amp; internet marketi
<link rel="canonical" href="http://www.seo.com/" />
<meta property="og:locale" content="en_us" />
<meta property="fb:admins" content="17807199" />
<meta property="og:title" content="SEO.com | Search Engine Optimization &amp; Internet Marketing Company" />
<meta property="og:description" content="Increase your bottom line with SEO.com: a search engine optimization &amp; internet
<meta property="og:url" content="http://www.seo.com/" />
<meta property="og:site_name" content="SEO.com" />
<meta property="og:type" content="article" />
<meta property="og:image" content="http://5.seo.com/wp-content/themes/seo-theme/library/images/seo-logo.png" />
<meta property="og:image" content="http://5.seo.com/wp-content/uploads/2009/12/seocom-logo.png" />
<meta name="twitter:card" content="summary">
<meta name="twitter:site" content="@seocom">
<meta name="twitter:title" content="SEO.com | Search Engine Optimization &amp; Internet Marketing Company">
<meta name="twitter:description" content="Increase your bottom line with SEO.com: a search engine optimization &amp; internet
<meta name="twitter:url" content="http://www.seo.com/">
<meta name="twitter:image" content="http://5.seo.com/wp-content/uploads/2009/12/seocom-logo.png">

<meta name="generator" content="WordPress 3.4.1"/>
<meta name="template" content="SEO.com 1.0"/>
<meta name="robots" content="index, follow"/>
<meta name="author" content="SEO.com"/>
<meta name="copyright" content="Copyright (c) December 2009"/>
```

source code of <http://www.seo.com/>

use archiving-friendly platform providers and CMSs



The screenshot shows a web browser displaying a Drupal robots.txt file. The browser's address bar shows 'drupal / robots.txt'. Below the address bar, there is a comment from 'webchick' 11 months ago, issue #1249132, suggesting to add 'INSTALL.sqlite.txt' to the robots.txt file. Below the comment, it says '2 contributors'. The main content area shows the robots.txt file with 61 lines (59 sloc) and a size of 1.561 kb. The file content is as follows:

```
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: /profiles/
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
```

The lines 25 through 27, which are 'Disallow: /profiles/', 'Disallow: /scripts/', and 'Disallow: /themes/', are circled in blue.

robots.txt for Drupal 7

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



“Highlighters” by Flickr user [KJGarbutt](#) under [CC BY-ND 2.0](#)

thank you!

Nicholas Taylor
[@nullhandle](#)