

Drupal.org Robots.txt Recommendations

This document illustrates some issues with the current robots.txt file on Drupal.org and is supplemental to the article at <http://drupalzilla.com/tutorial/seo/drupal-org-seo> – part 1 of a series consisting of an SEO analysis of Drupal.org.

The screenshot shows the Drupal.org homepage with a blue header containing navigation links: Support, Handbooks, Pages, Downloads, Contribute, Contact, and Community. A search bar is located in the top right. The main content area is divided into several sections:

- Read all about the Drupal Newsletter:** A blue box with a 'Subscribe' button and a 'Find out the new world!' link.
- Drupal.org is the official website of Drupal:** A yellow box with text about Drupal's features and a 'More!' section listing 'About Drupal', 'Screenshots', 'Features', 'Demo', 'Hosting', and 'Paid services'.
- Download:** A green box with 'Latest release: Drupal 5.2' and links for 'Contributions', 'Modules', 'Themes', and 'Translations'.
- Security updates and bugfixes available: Drupal 5.2 and 4.7.7 released:** A blue box with a 'News and announcements' tag, dated 7/30/07, and a 'Download' section with links for 'Download Drupal 5.2' and 'Download Drupal 4.7.7'.
- FroSCon approaches - August 25th and 26th:** A blue box with a 'Events' tag, dated 8/16/07, and text about the conference.
- Announcing Drupal Association Planet:** A blue box with a 'News and announcements' tag, dated 8/7/07, and text about the new association.
- Get Drupal in your local Wikipedia right by 9/8/7:** A blue box with a '12 comments' link and a 'Put Drupal in your Wikipedia' button.
- Improvements to the Drupal.org infrastructure:** A blue box with a 'News and announcements' tag, dated 7/18/07, and text about site updates.
- Drupal awarded \$5000 USD OpenID Bounty:** A blue box with a 'News and announcements' tag, dated 7/26/07, and text about the bounty.

On the right side of the page, there are several utility boxes:

- User login:** A form with 'Username' and 'Password' fields, a 'Log in' button, and links for 'Create new account' and 'Request new password'.
- New forum topics:** A list of forum topics with links to 'more'.
- Groups - Events:** A list of events with links to 'more'.

At the bottom of the page, there is a pagination bar with numbers 1 through 9, and 'next' and 'last' buttons.

```

User-agent: *
Crawl-delay: 10
# Directories
Disallow: /database/
Disallow: /includes/
Disallow: /misc/
Disallow: /modules/
Disallow: /sites/
Disallow: /themes/
Disallow: /scripts/
Disallow: /updates/
Disallow: /profiles/
# Files
Disallow: /xmlrpc.php
Disallow: /cron.php
Disallow: /update.php
Disallow: /install.php
Disallow: /INSTALL.txt
Disallow: /INSTALL.mysql.txt
Disallow: /INSTALL.pgsql.txt
Disallow: /CHANGELOG.txt
Disallow: /MAINTAINERS.txt
Disallow: /LICENSE.txt
Disallow: /UPGRADE.txt
# Paths (clean URLs)
Disallow: /admin/
Disallow: /aggregator/
Disallow: /comment/reply/
Disallow: /contact
Disallow: /logout
Disallow: /node/add/
Disallow: /search/
Disallow: /user/register
Disallow: /user/password
Disallow: /user/login
# Paths (no clean URLs)
Disallow: /?
# Extras on drupal.org
# no access for table sorting paths or any paths that have parameters.
Disallow: /*sort=

# Additional Rules

# Block user tracker pages
Disallow: /*/track$
Disallow: /*/track?page=
Allow: /project/track

```

The trailing slashes should be removed from these five lines, otherwise they will not block the pages that they are intended to block. Drupal does not create trailing slashes on URLs.

Since Drupal.org has clean URLs, this single rule can replace all the other rules for non-clean URLs.

The rule `/*sort=` was recently added, but the problem that that rule attempts to solve can also appear in the format `&sort`. This rule should be changed as shown to `/*sort=`

It also needs a *Disallow* directive, and there is no need to add a trailing asterisk.

These lines would block the user tracker pages while allowing the Tracker Project itself. Google and Yahoo both support the *Allow* directive, though MSN apparently doesn't.

Trailing Slashes

The following lines in Drupal 5's robots.txt file originally contained trailing slashes. I recommend removing the trailing slashes on the following rules because Drupal does not have trailing slashes in its URLs.

```
Disallow: /contact
Disallow: /logout
Disallow: /user/register
Disallow: /user/password
Disallow: /user/login
```

Blocking ***/user/register*** and ***/user/login*** will also block the “add comment” URLs on every node. An example of the undesirable URLs is shown in the image below.



Sorted Tables and Views

The scope of the *sort* parameter in URLs on Drupal.org can be seen by this Google query that shows 24,000 results:

<http://www.google.com/search?q=site%3Ahttp%3A%2F%2Fdrupal.org%2F+inurl%3Asort&num=100>

The following was recently added to the default robots.txt file on Drupal.org:

```
# Extras on drupal.org
# no access for table sorting paths or any paths that have parameters.
/*?sort*
```

Because the sort parameter sometimes appears as **&sort=** and not just **?sort=** I recommend changing that above rule to:

```
Disallow: /*sort=
```

(Don't forget the *Disallow* directive.)

User Tracker Pages

The following query shows over 83,000 user tracker pages indexed in Google:
<http://www.google.com/search?q=site%3Adrupal.org+inurl%3Atrack&num=100>

The following lines would block those user tracker pages:

```
# Block user tracker pages
Disallow: /*/track$
Disallow: /*/track?page=
Allow: /project/track
```

The Allow directive is supported by Google and Yahoo, but apparently not by MSN. If the above rules are implemented, MSN Live Search and other smaller search engines might be blocked from the page <http://drupal.org/project/track>. However the 83,000 low-quality tracker pages would be blocked.

The possible benefit of user tracker pages is that they may provide a way for search engine spiders to find content, though I suspect that 83,000+ pages of links are too many and I recommend adding the rules above as a test.

Future Robots.txt Rules

I would add the above changes to robots.txt to test them, and then wait for a while to see how search engines respond. I would then probably add another set of rules based on the following concepts:

Tracker Pages

The URLs in the format <http://drupal.org/tracker?page=> have created over 6000 pages of links that I believe are relatively low quality. The following Google query shows that the tracker pages are getting indexed and that they are relatively low quality:

<http://www.google.com/search?q=site%3Ahttp%3A%2F%2Fdrupal.org%2Ftracker&num=100>

When given a choice of having a site's content spidered by date or by category, I prefer to have the site spidered by category.

When search engines go to the forum posts by links in the forum sections, they are spidering the site as organized by taxonomy/category. When search engines go to the posts through the tracker pages, they are spidering the same duplicate links by date. Organization by taxonomy/category creates sections of the site on similar keyword themes. Organization by date does not give as much keyword information to search engines.

I would test blocking all tracker pages except for the first page so that search engines only spider the site through the links in the forum. The first page of tracker results would not be blocked because as soon as a page is created, search engines would be able to find it on the page <http://drupal.org/tracker>.

The following rules will block those tracker pages:

```
Disallow: /tracker?  
Disallow: /tracker/
```

The same issue exists here as with the user tracker pages. Is the benefit of providing links to the content through the tracker pages outweighing the problem of having so many low-quality pages? Because search engines can also spider the content through the forums I would add the above robots.txt rules as a test.

RSS Feeds

Search engines are also spidering a large number of feeds:

<http://www.google.com/search?q=site%3Adrupal.org+inurl%3Afeed&num=100>

The following rule would block RSS feeds:

```
Disallow: /*/feed$
```

Normally, one can add that line to a Drupal site and not block the main RSS feed which is typically found at <http://example.com/rss.xml>. However, in the case of Drupal.org, the main RSS feed is located at the following URL:

<http://drupal.org/node/feed>

That means that blocking `/*/feed$` would also block the main RSS feed.

For now I would not block the RSS feeds because it would be more beneficial to try the initial set of robots.txt changes and monitor them before adding another major robots.txt rule.

Robots.txt Standard

Some of the robots.txt rules that are mentioned in this document are not part of the robots.txt standard. However, they are supported by the three major search engines: Google, Yahoo, and MSN Live. For more information please see the following resources:

- Google: <http://www.google.com/support/webmasters/bin/answer.py?answer=40367>
- Yahoo: <http://www.ysearchblog.com/archives/000372.html>
- MSN:
http://search.msn.com.sg/docs/siteowner.aspx?t=SEARCH_WEBMASTER_REF_RestrictAccessToSite.htm – note that MSN apparently doesn't support the *Allow* directive.