

Robots.txt Tutorial

Search engines will look in your root domain for a special file named "robots.txt" (<http://www.mydomain.com/robots.txt>). The file tells the robot (spider) which files it may spider (download). This system is called, *The Robots Exclusion Standard*.

The format for the **robots.txt** file is special. It consists of records. Each record consists of two fields : a User-agent line and one or more **Disallow:** lines. The format is:

```
<Field> ":" <value>
```

The robots.txt file should be created in Unix line ender mode! Most good text editors will have a Unix mode or your FTP client *should* do the conversion for you. Do not attempt to use an HTML editor that does not specifically have a text mode to create a robots.txt file.

User-agent

The **User-agent** line specifies the robot. For example:

```
User-agent: googlebot
```

You may also use the wildcard character "*" to specify all robots:

```
User-agent: *
```

You can find user agent names in your own logs by checking for requests to robots.txt. Most major search engines have short names for their spiders.

Disallow:

The second part of a *record* consists of **Disallow:** directive lines. These lines specify files and/or directories. For example, the following line instructs spiders that it can not download email.htm:

```
Disallow: email.htm
```

You may also specify directories:

```
Disallow: /cgi-bin/
```

Which would block spiders from your cgi-bin directory.

There is a wildcard nature to the Disallow directive. The standard dictates that /bob would disallow **/bob.html** and **/bob/index.html** (both the file bob and files in the bob directory will not be indexed).

If you leave the Disallow line blank, it indicates that ALL files may be

SiteMenu:

- [Spiders](#)
- [The Players](#)
- [Meta Tags](#)
- [Tips & Tricks](#)
- [Design](#)
- [Avoidance](#)
- [Columns](#)
- [Partners](#)
- [SE News Links](#)
- [Forums](#)

- [Webmaster World](#)
- [SE Glossary](#)
- [SE Tools](#)
- [Site Search](#)
- [SEW Quarterly](#)

- [Resources](#)
- [About](#)
- [Contacts](#)
- [Advertising](#)
- [Home](#)

retrieved. At least one disallow line must be present for each User-agent directive to be correct. A completely empty Robots.txt file is the same as if it were not present.

White Space & Comments

Any line in the robots.txt that begins with # is considered to be a comment only. The standard allows for comments at the end of directive lines, but this is really bad style:

```
Disallow: bob #comment
```

Some spider will not interpret the above line correctly and instead will attempt to disallow "bob#comment". The moral is to place comments on lines by themselves.

White space at the beginning of a line is allowed, but not recommended.

```
Disallow: bob #comment
```

Examples

The following allows all robots to visit all files because the wildcard "*" specifies all robots.

```
User-agent: *  
Disallow:
```

This one keeps all robots out.

```
User-agent: *  
Disallow: /
```

The next one bars all robots from the cgi-bin and images directories:

```
User-agent: *  
Disallow: /cgi-bin/  
Disallow: /images/
```

This one bans Roverdog from all files on the server:

```
User-agent: Roverdog  
Disallow: /
```

This one bans keeps googlebot from getting at the cheese.htm file:

```
User-agent: googlebot  
Disallow: cheese.htm
```

For more complex examples, try retrieving some of the robots.txt files from the big sites like Cnn, or Looksmart.

Extensions to the Standard

Although there have been proposed standards extensions such as an **Allow** line or robot

version control, there has been no formal endorsement by the Robots exclusion standard working group.

Related

- [Robots.txt Forum](#)
- [Robots.txt Validator](#)
- [Robots.txt Tutorial Using robots.txt.](#)
- [Robots Exclusion Meta Tag Using robots metatags.](#)
- [Robots.txt : The Big Crawl](#)

We recently spidered 2million robots.txt files and found a surprising number of problems.

- [Robots Exclusion Standard rfc4.](#)
- [Root of Robots Exclusion Standard directory with some interesting files.](#)
- [Search Indexing Robots and Robots.txt article at searchtools.com.](#)



Privacy Policy (C) Copy & Copyright 1996-2002 [PHD Software Systems](#) all rights reserved