

This Free E-Book is brought to you by Natural-Aging.com.



Working with the robots.txt file

By Jagdeep.S.Pannu

Working with the robots.txt file by Jagdeep.S.Pannu

Working with the robots.txt file

What is the robots.txt file?

Working with the robots.txt file

Advantages of robots.txt

Disadvantages of the robots.txt file

Optimization of the robots.txt file

Using the robots.txt file

What is the robots.txt file?

The robots.txt file is an ASCII text file that has specific instructions for search engine robots about specific content that they are not allowed to index. These instructions are the deciding factor of how a search engine indexes your website's pages. The universal address of the robots.txt file is: www.domain.com/robots.txt . This is the first file that a robot visits. It picks up instructions for indexing the site content and follows them. This file contains two text fields. Lets study this robots.txt example :

User-agent: *

Disallow:

The User-agent field is for specifying robot name for which the access policy follows in the Disallow field. Disallow field specifies URLs which the specified robots have no access to. A robots.txt example :

Working with the robots.txt file

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

Here "*" means all robots and "/" means all URLs. This is read as, " No access for any search engine to any URL" Since all URLs are preceded by "/" so it bans access to all URLs when nothing follows after "/". If partial access has to be given, only the banned URL is specified in the Disallow field. Lets consider this robots.txt example :

```
# Research access for Googlebot.  
User-agent: Googlebot  
Disallow:
```

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

Here we see that both the fields have been repeated. Multiple commands can be given for different user agents in different lines. The above commands mean that all user agents are banned access to /concepts/new/ except Googlebot which has full access. Characters following # are ignored up to the line termination as they are considered to be comments.

Working with the robots.txt file : –

The robots.txt file is always named in all lowercase (e.g. Robots.txt or robots.Txt is incorrect)

Wildcards are not supported in both the fields. Only * can be used in the User-agent fields' command syntax because it is a special character denoting "all". Googlebot is the only robot that now supports some wildcard file extensions.

Ref: <http://www.google.com/webmasters/faq.html#12>

The robots.txt file is an exclusion file meant for search engine robot reference and not obligatory for a website to function. An empty or absent file simply means that all robots are welcome to index any part of the website.

Only one robots.txt file can be maintained per domain.

Website owners who do not have administrative rights cannot sometimes make a robots.txt file. In such situations, the Robots Meta Tag can be configured which will solve the same purpose. Here we must keep in mind that lately, questions have been raised about robot behavior regarding the Robot Meta Tag. Some robots might skip it altogether. Protocol makes it obligatory for all robots to start with the robots.txt thereby making it the default starting point for all robots.

Separate lines are required for specifying access to different user agents and Disallow field should not carry more than one command in a line in the robots.txt file. There is no limit to the number of lines though i.e. both the User-agent and Disallow fields can be repeated with different commands any number of times. Blank lines will also not work within a single record set of both the commands.

Working with the robots.txt file

Use lower–case for all robots.txt file content. Please also note that filenames on Unix systems are case sensitive. Be careful about case sensitivity when defining directory or files for Unix hosted domains. You can use this great tool to check your robots.txt from www.searchengineworld.com:

The robots.txt Validator

Please note that the full path to the robots.txt file must be entered in the field.

Advantages of the robots.txt file : –

Protocol demands that all search engine robots start with the robots.txt file. This is the default entry point for robots if the file is present. Specific instructions can be placed on this file to help index your site on the web. Major search engines will never violate the Standard for Robots Exclusion.

The robots.txt file can be used to keep out unwanted robots like email retrievers, image strippers etc.

The robots.txt file can be used to specify the directories on your server that you don't want robots to access and/or index e.g. temporary, cgi, and private/back–end directories.

An absent robots.txt file could generate a 404 error and redirect the robot to your default 404 error page. Here it was noticed after careful research that sites that do not have a robots.txt file present and had a customized 404–error page, would serve the same to the robots. The robot is bound to treat it as the robots.txt file, which can confuse it's indexing.

The robots.txt file is used to direct select robots to relevant pages to be indexed. This specially comes in handy where the site has multilingual content or where the robot is searching for only specific content.

The need for the robots.txt file was also felt to stop robots from deluging servers with rapid–fire requests or re–indexing the same files repeatedly. If you have duplicate content on your site for any reason, the same can be controlled from getting indexed. This will help you avoid any duplicate content penalties.

Disadvantages of the robots.txt file : –

Careless handling of directory and filenames can lead hackers to snoop around your site by studying the robots.txt file, as you sometimes may also list filenames and directories that have classified content. This is not a serious issue as deploying some effective security checks to the content in question can take care of it. For example if you have your traffic log on your site on a URL such as www.domain.com/stats/index.htm which you do not want robots to index, then you would have to add a command to your robots.txt file. As an example:

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

However, it is easy for a snooper to guess what you are trying to hide and simply typing the URL www.domain.com/stats in his browser would enable access to the same. This calls for one of the following remedies –

Working with the robots.txt file

Change file names:

Change the stats filename from index.htm to something different, such as stats-new.htm so that your stats URL now becomes www.domain.com/stats/stats-new.htm

Place a simple text file containing the text, "Sorry you are not authorized to view this page", and save it as index.htm in your /stats/directory.

This way the snooper cannot guess your actual filename and get to your banned content.

Use login passwords:

Password-protect the sensitive content listed in your robots.txt file.

Optimization of the robots.txt file : –

The right commands in robots.txt : Use correct commands. Most common errors include – putting the command meant for "User-agent" field in the "Disallow field" and vice-versa.

Please also note that there is no "Allow" command. Content not blocked in the "Disallow" field is considered allowed. Currently, only two fields are recognized: "The User-agent field" and the "Disallow

field". Experts are considering the addition of more robot recognizable commands to make the robots.txt file more Webmaster and robot friendly.

Bad Syntax: Do not put multiple file URLs in one Disallow line in the robots.txt file. Use a new Disallow line for every directory that you want to block access to. Incorrect Robots.txt example :

```
User-agent: *  
Disallow: /concepts/ /links/ /images/
```

Correct robots.txt example:

```
User-agent: *  
Disallow: /concepts/  
Disallow: /links/  
Disallow: /images/
```

Files and directories: If a specific file has to be disallowed, end it with the file extension and without a forward slash in the end. Study the following Robots.txt example :

For file:

```
User-agent: *  
Disallow: /hilltop.html
```

For Directory:

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

Remember if you have to block access to all files in the directory, you don't have to specify each and every file in robots.txt . You can simply block the directory as shown above. Another common error is leaving out the slashes altogether. This would leave a very different message than intended.

The right location for the robots.txt file: No robot will access a badly placed robots.txt file. Make sure that the location is `www.domain.com/robots.txt`.

Capitalization in robots.txt : Never capitalize your syntax commands. Directory and filenames are case sensitive in Unix platforms. The only capitals used per standard are: "User-agent " and "Disallow "

Correct Order for robots.txt : If you want to block access to all but one or more than one robot, then the specific ones should be mentioned first. Lets study this robots.txt example :

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

```
User-agent: googlebot  
Disallow:
```

In the above case, Googlebot would simply leave the site without indexing after reading the first command. Correct syntax is:

```
User-agent: googlebot  
Disallow:
```

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

The robots.txt file : Not having a robots.txt file at all could generate a 404 error for search engine robots, which could redirect the robot to the default 404-error page or your customized 404-error page. If this happens seamlessly, it is up to the robot to decide if the target file is a robots.txt file or an html file. Typically it would not cause many problems but you may not want to risk it. It's always a better idea to put the standard robots.txt file in the root directory, than not having it at all.

The standard robots.txt file for allowing all robots to index all pages is:

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

Using # Carefully in the robots.txt file: Adding comments after the syntax commands is not a good idea

using "#". Some robots might misinterpret the line although it is acceptable as per the robots exclusion standard. New lines are always preferred for comments.

Using the robots.txt file : –

Robots are configured to read text. Too much graphic content could render your pages invisible to the search engine. Use the robots.txt file to block irrelevant and graphic-only content.

Indiscriminate access to all files, it is believed, can dilute relevance to your site content after being indexed by robots. This could seriously affect your site's ranking with search engines. Use the robots.txt file to direct robots to content relevant to your site's theme by blocking the irrelevant files or directories.

The robots.txt file can be used for multilingual websites to direct robots to relevant content for relevant topics for different languages. It ultimately helps the search engines to present relevant results for specific languages. It also helps the search engine in its advanced search options where language is a variable.

Some robots could cause severe server loading problems by rapid firing too many requests at peak hours. This could affect your business. By excluding some robots that might be irrelevant to your site, in the robots.txt file, this problem can be taken care of. It is really not a good idea to let malevolent robots use up precious bandwidth to harvest your emails, images etc.

Use the robots.txt file to block out folders with sensitive information, text content, demo areas or content yet to be approved by your editors before it goes live.

The robots.txt file is an effective tool to address certain issues regarding website ranking. Used in conjunction with other SEO strategies, it can significantly enhance a website's presence on the net.

Related Reading : –

A Standard for Robots Exclusion.

Guide to The Robots Exclusion Protocol

W3C Recommendations

Article last updated : 11th March 2004

(c) Copyright 2004 Jagdeep.S. Pannu, SEORank

This Article is Copyright protected. If you have comments; or would like to have this article republished on your site, please contact the author here:

. We just require all due credits

carried; and text, hyperlinks and headers unaltered. This article must not be used in unsolicited mail.

Jagdeep.S.Pannu is Manager–Online Marketing, at www.SEORank.com, a leading Search Engine Optimization Services Company.

The role of the robots.txt file to improve site ranking!

By Michael Kralj

The role of the robots.txt file to improve site ranking! by Michael Kralj

Not many web master take the time to use a robots.txt file for their website. For search engine spiders that use the robots.txt to see what directories to search through, the robots.txt file can be very helpful in keeping the spiders indexing your actual pages and not other information, such as looking through your stats!

The robots.txt file is useful in keeping your spiders from accessing parts folders and files in your hosting directory that are totally unrelated to your actual web site content. You can choose to have the spiders kept out of areas that contain programming that search engines cannot parse properly, and to keep them out of the web stats portion of your site.

Many search engines cannot view dynamically generated content properly, mainly created by programming languages, such as PHP or ASP. If you have an online store programmed in your hosting account, and it is in a separate directory, you would be wise to block out the spiders from this directory so it only finds relevant information.

The robots.txt file should be placed in the directory where your main files for your hosting are located. So you would be advised to create a blank text file, and save it as robots.txt, and then upload it to your web hosting to the same directory your index.htm file is located.

Here is examples of the use of the robots.txt file:

To block out a directory in a robots.txt file, such as a subdirectory for your online store called /store/ you would do the following:

```
Disallow: /store/
```

Another example to block out your stats directory:

```
Disallow: /stats/
```

You may also want to disallow individual files that you do not want searched by the search engines. For example you dont want search.php to be parsed by the Search Engines. To do this you type in the following on its own line:

```
Disallow: /search.php
```

Following the rules outlined and creating the robots.txt file, you will keep search engine spiders out of unwanted files and directories, and letting them go through the important files to see what your web site is all about!

Michael Kralj is owner of Emenki Web Solutions and Domains at Retail. Emenki Web Solutions are web site designers and programmers based in Hamilton, Ontario, providing businesses with an

informative and strategic approach to establishing an online presence on the web. Please visit Emenki Web Solutions on the web <http://www.emenki.com> Please visit Domains at Retail on the web: <http://www.domainsatretail.com>



This Free E-Book has been brought to you by Natural-Aging.com.

[100% Effective Natural Hormone Treatment](#)
Menopause, Andropause And Other Hormone Imbalances
Impair Healthy Healing In People Over The Age Of 30!