

Robots

The robots.txt file is one of the least understood aspects of the search engine optimization world. Essentially, it is a means to tell the various search engine spiders (or robots, bots) to crawl or not to crawl specific sections of a website. The robots.txt file can also be used to prevent any indexing whatsoever or to provide certain spiders and bots with specific instructions about how to index your site.

So how do you use a robots.txt file to your advantage?

The robots.txt file is one of the least understood aspects of the search engine optimization world. Essentially, it is a means to tell the various search engine spiders (or robots, bots) to crawl or not to crawl specific sections of a website. The robots.txt file can also be used to prevent any indexing whatsoever or to provide certain spiders and bots with specific instructions about how to index your site. So, what is the need for the robots.txt file?

Since many search engine spiders look for the robots.txt file as they arrive on a site, many Search Engine Optimization (SEO) experts agree that including such a file is a good idea because it acts as an invitation to crawl and to index your website's content.

However, there are some very important instances when you may want to limit or to even exclude bots from crawling a site. Some examples of this are when there are rogue spiders that are crawling for the chief purpose of indexing your site for their own use, when there is sensitive information (e.g., unfinished projects you do not want indexed such as site redesigns or exclusive beta-tests) and in situations when site owners decide that there is no need to index portions of their site such as image files or cgi bins.

The very fact that search engines are scanning through files that surfers will never see is reason enough to put a robots file on your site. Have you looked at your site statistics recently? If your stats include a section on 'files not found', you are sure to see many entries where search engines' spiders looked for and failed to find a robots.txt file on your site.

Creating the robots.txt file

Creating a basic robots.txt file is a relatively simple process. Open notepad or your favorite text editor and follow along with the instructions below. Every robots.txt file contains records of two fields : a "User-agent" line and a "Disallow" line. The User-Agent line specifies the robot or spider that you are instructing, and the "Disallow" line provides the instructions on what can or cannot be indexed.

In the case of the User-agent, the asterisk (*) is essentially the symbol for 'all' – so allow all 'User-agents' or robots. The 'Disallow:' field informs the robot (User-agent) what to crawl or what not to crawl. Allow crawling by leaving the field blank (option #1) or

disallow all crawling by including the wildcard forward slash (option #2). If you use this disallow command while creating a website, do not forget to remove it once the site is live.

While the majority of websites welcome robots to freely index a website, there are some instances where the robots' crawling may be unnecessary or is forbidden and therefore unwelcome. If you have files on your site, for example confidential communications with clients or a new pending website design, website owners and managers can exclude certain files from all robots or from individual search engines.

Take, for instance, that you have a file called Secret-file.htm in a directory called 'CONFIDENTIAL', listing your secret to securing a number one position at Google without any effort, that you do not wish to be spidered by robots. You would then add the following lines to your robots.txt file:

If, for some reason, you choose to prohibit some robots/spiders from crawling your site, the User-agent should include the name of the specific spider indexing your site.

Alternatives to Robots.txt file:

While you do not have to go to the trouble of creating a robots.txt file, it is an older, more respected method of controlling robots and webcrawlers. Some website owners choose to use the "noindex,nofollow" attribute in their HTML meta-tags. Though not a foolproof way to eliminate robots that routinely burn up bandwidth, such an attribute does mesh with the general objectives of many websites.

Though there are hundreds if not thousands of bots and spiders crawling the Web which will take note of your robots.txt file not all of them will. As such, do not rely on such a file to protect sensitive information. For readers interested in some advanced robots.txt file techniques, visit the weblog at www.websiteservices.com and search for "Robots". You are sure to find helpful information and even several resources to use in creating your own robots.txt file.