
Web Hosting Knowledge Base

HTML Help Category

Contents

HTML Help	1
<i>How do I block visitors from my site?</i>	1
<i>Limiting what search engines can index using /robots.txt</i>	1
<i>Why do I get a red (x) where my images are supposed to be?</i>	2

HTML Help

How do I block visitors from my site?

Create a `.htaccess` file and add the following code--changing the IPs to suit your needs--each command on one line each:

```
<B>
order allow,deny
deny from 123.45.6.7
deny from 012.34.5.
allow from all
</B>
```

You can deny access based upon IP address or an IP block. The above blocks access to the site from 123.45.6.7, and from any sub domain under the IP block 012.34.5. (012.34.5.1, 012.34.5.2, 012.34.5.3, etc.)

You can also set an option for deny from all, which would of course deny everyone. You can also allow or deny by domain name rather than IP address (allow from `.friendsite.com` works, etc.)

Limiting what search engines can index using `/robots.txt`

Various search engines such as Google have what are called "spiders" or "robots" continually crawling the web indexing content for inclusion in their search engine databases. While most users view inclusion in search engine listings in a positive light and high search engine rankings can translate to big bucks for commercial sites not everyone wants every single page and file stored on their account publicly available through web searches.

This is where `/robots.txt` comes in. Most search engine robots will comply with a webmaster/site owners wishes as far as excluding content by following a robots inclusion standard which is implemented via the use of a small ASCII text file named `/robots.txt` in the root web accessible directory of a given domain.

When a compliant robot visits a given site the first thing it does is to check the top level directory for the presence of a file named "robots.txt". If found the directives within the file which tells the robot what if any content it can or cannot visit and index is read, and in most cases honored.

Creating `/robots.txt` files

To create a `/robots.txt` file simply open a plain text editor such as Windows NotePad, type or paste your directives and save the file using the file name "robots" (`robots.txt`). This file should then be uploaded to the `/public_html` directory such that it's URL will be `http://domain.com/robots.txt`

`/robots.txt` syntax

All valid `/robots.txt` files must contain at least two lines in the following format:

```
User-Agent: [robot name or * for all robots]
Disallow: [name of file or directory you do not want indexed]
```

Unless one wishes to implement different rules for specific robots the user agent line should just include an asterisk [*] which is a wildcard read as "rules apply to all robots".

Disallow lines can be used to specify specific files or folders one doesn't wish to have indexed by search engines. Each file or folder to be excluded must be listed separately on it's own line, and wildcards are not supported in Disallow directives. One can have as many or as few disallow lines as is necessary.

Example /robots.txt files

- A simple /robots.txt file which would allow all robots to access and index all content with the exception of the contents of a directory named "private" would be as follows:

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

- A /robots.txt file which would exclude all robots from indexing the content of "cgi-bin", "admin" and "stuff" directories plus a page named "private.html" would be:

```
User-agent: *  
Disallow: /cgi-bin/  
Disallow: /admin/  
Disallow: /stuff/  
Disallow: /private.html
```

- A /robots.txt file which would allow all robots to access and index all content on a given site would be:

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

- A /robots.txt file which would forbid all robots from accessing and indexing any content would be:

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

- A robots.txt file which would allow Google's spider (aka GoogleBot) to index all content with the exception of files stored under a folder named "private" and which would exclude all other robots from indexing any content would read as follows:

```
User-agent: GoogleBot  
Disallow: /private/
```

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

- A robots.txt file which would allow all robots with the exception of HotBot's (aka Inktomi Slurp) to index all content with the exception of files stored under folders named "images" and "cgi-bin" and which would exclude the HotBot spider from indexing any content would read as follows:

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

```
User-agent: Inktomi Slurp  
Disallow: /
```

More Information

For more details on /robots.txt and the Robots Exclusion Standard visit The Web Robots Pages at <http://www.robotstxt.org>

Why do I get a red (x) where my images are supposed to be?

Either the image was not uploaded and therefore does not exist on the server or your path to the image is incorrect. You may have put it in a different folder or named it using a capital but created a link to it using lowercase.

Our servers run on Linux and FreeBSD. They don't like spaces in filenames and they treat capitals and lowercase letters as completely different letters. it is best to name files in all lowercase.

You will see a red X whenever the server cannot find an image. Right click on the red X and choose Properties. This will show you the link to the image so you can track down the problem.