

---

# Tutorial: robots.txt – your guide for the search engines

Written by [Tess](#) on February 28, 2012 in [SEO basics](#) - [No comments](#)



It's amazing how one little text file can make or break your website. If you have a line or two wrong in your robots.txt file, you might actually be telling the search engine robots to **not** crawl and index your site...which means your web pages won't show up in the search engines.

Luckily it's easy to check, and easy to solve. In this blog post you'll learn about what robots.txt is, how to check your own robots.txt, and how you can improve your instructions for the search engine robots.

This blog post is part 1 of the series: **how to do a Technical SEO Audit of your website**. For a list of all blog posts in the same series, [view the blog tag "technical seo audit"](#).

Today we'll talk about the following topics:

- What is robots.txt?
- What can you use robots.txt for?
- How does it work?
- How to create a robots.txt file
- What to put in your robots.txt file
- How to check and test your robots.txt file
- Funny and creative examples of robots.txt

---

## What is robots.txt?

It's a file in the root of your website that can either **allow or restrict search engine robots** from crawling pages on your website.

Think of the search engines as a big library of all the web pages in the world. Google, Yahoo, and others, send out their spiders (also known as crawlers or robots) to find new or updated pages to add to their **index**. First thing they look for when they arrive to your website is your robots.txt file. In your robots.txt file, you show the robots which of your pages you want (or don't want) them to read (**crawl**).

Keep in mind that there is a difference between “*crawl*” and “*index*”. The search engine can *crawl* (read) a page without *indexing* (listing in the search results), and the other way around. It all depends on what instructions you put in your robots.txt and robots meta tags.

## Do you need a robots.txt file?

If the search engine spider doesn't find a robots.txt for your website, it will crawl and index all your pages (unless you've implemented robots meta tags with other instructions).

“If you want search engines to index everything in your site, you don't need a robots.txt file (not even an empty one).”

Source: [Google Webmaster Tools Help](#)

However, if you don't have a robots.txt file, your server logs will return 404 errors whenever a robot tries to access your robots.txt file.

...to prevent the “file not found” error messages in your web server log, you can create an empty file named robots.txt.

Source: [Googlebot](#)

This is not always ideal, and after reading this blog post I'm sure you can find at least one thing to improve in your own robots.txt file.

---

## What can you use robots.txt for?

Among other things, your robots.txt file can help:

- if you have pages or directories on your web site that **should not appear in the SERPs** (Search Engine Result Pages)
- if you want **duplicate pages to be ignored**, for example if your website CMS generate more than one URL for the same content
- if you don't want your website's **internal search result pages** to be indexed
- to tell search engines where your **sitemap** is
- to tell search engines **which version to index**, if you for example have both an HTML and a PDF version of the same content

Something important to keep in mind is:

“...a robots.txt file is something like putting a note “Please, do not enter” on an unlocked door – e.g. you cannot prevent thieves from coming in but the good guys will not open to door and enter.”

Source: [What is robots.txt?](#)

---

## How does it work?

Before a search engine robot crawls your website, it will first look for your robots.txt file to find out where you want them to go.

**Did you know that...** in your Google Analytics account, a visit from a robot gets added to your overall visitors stats. To see only the real (human) visitors stats, you can apply a filter to exclude the traffic from robots. But that's a topic for another blog post.

There are 3 things you should keep in mind:

1. **Robots can ignore your robots.txt.** Malware robots scanning the web for security vulnerabilities, or email address harvesters used by spammers, will not care about your instructions.
2. **The robots.txt file is public.** Anyone can see what areas of your website you don't want robots to see.
3. **Search engines can still index (but not crawl) a page you've disallowed,** if it's linked to from another website. In the search results it'll then only show the url, but usually no title or information snippet. Instead, make use of the robots meta tag for that page.

Now go to your website and **check if you have a robots.txt file.** Just add `/robots.txt` after your domain name. It will then look something like this: `http://www.yourwebsite.com/robots.txt`

If your robots.txt says this, then you're in trouble:

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

Keep on reading and you'll soon understand why.

---

## How to create a robots.txt file

If you don't have a robots.txt file, you should quickly create one before you continue reading:

1. **Create a regular text file and save it as `robots.txt`.** Remember to use all lower-case for the filename: `robots.txt` (not `Robots.TXT`).
2. **Upload it to the root directory** of your website, not a subdirectory.
3. If you've done it right, you should now be able to **see your robots.txt file at `http://www.yourwebsite.com/robots.txt`**

**Note:** if you use subdomains, you have to create a separate robots.txt file for each subdomain.

---

## What to put in your robots.txt file

Many website owners disagree about what you should and shouldn't put in the robots.txt file, and it's up to you what you think works best for your website and needs.

**WARNING:** robots.txt is **not** intended to deal with the security of your website!

It is recommended that the location of your admin area, and other private areas of your website, are **not** included in the robots.txt. Instead, you can for example use the robots meta tag to keep the

major search engines from crawling/indexing it.

If you really want to securely block robots from accessing private content, you should **look at proper security measures** for this (.htaccess and other ways).

Use robots.txt as a *guide* for the robots, but know that it's up to each robot to honor the instructions.

## What directives can you use?

First, **open some robots.txt files** and use them as references when you continue reading. Go ahead, open your competitor's robots.txt, or one for a website using the same CMS as you (just add /robots.txt after the domain). To help you, here are a few examples:

- [Google's robot.txt file](#)
- [Yoast's robot.txt file](#)
- [SmashingMagazine's robot.txt file](#)
- [Ebay's robot.txt file](#)
- [SEO beginner's robot.txt file](#)
- [WPmods' robot.txt file](#)
- [Joomla's robot.txt file](#)
- [WordPress' robot.txt file](#)
- [Drupal's robot.txt file](#)
- [Wikimedia's robot.txt file](#)

Now we'll take a look at the different lines you can have in your robots.txt file:

### User-agent:

This is the line where you define which robot you're talking to. It's like saying hello to the robot:

*"Hi all robots"*

User-agent: \*

*"Hi Google robot"*

User-agent: Googlebot

*"Hi Yahoo! robot"*

User-agent: Slurp

For Google's different website crawlers, [see this list](#). Robotstxt.org has a [Robots Database](#), but I don't know when it was last updated.

**TIPS:** You can find out which robots crawl your site by looking through your server logs, and then use the information to add user-agent specific guidelines in your robots.txt.

Not all robots/user-agents understand all directives. In the original Robots Exclusion Protocol, the *Disallow*: directive was the only official option, and later came the inclusion of the *Sitemap*: directive.

In the list below you will find some non-standard but useful directives. Google and Bing follows most of them, but unfortunately not all other robots will understand and follow them.

For each of the following directives you must have the user-agent line first. It's like saying "Hi Google", and then follow with the specific instructions for Google.

Now let's tell the robots what you want them to do...

### **Disallow:**

This tells the robots what you don't want them to crawl on your site:

*"Hi all robots, do not crawl anything on my site"*

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

*"Hi Google image robot, do not crawl my images folder (but you can crawl everything else)"*

```
User-agent: Googlebot-Image  
Disallow: /images/
```

**Note:** many website owners disallow their images directory, but this can be a good thing to allow (think Google image search). Just make sure you've named your images properly – as in, the file name should reflect what the picture is about (not picture1.jpg, picture2.jpg etc). If you want to remove your images from Google's index, [read this info from Google](#).

### **Allow:**

This tells the robots what you want them to crawl on your site.

*"Hi all robots, you can crawl everything on my site"*

```
User-agent: *  
Allow: /
```

**Note:** If these are the only lines you have in your robots.txt you could delete the file instead. If there is no robots.txt, the search engines will assume anyway that you want them to crawl everything on your website.

*"Hi all robots, I don't want you to crawl anything in the /things/ folder, except the file /things/awesomestuff.html."*

```
User-agent: *  
Disallow: /things/  
Allow: /things/awesomestuff.html
```

Remember, specific instructions overrides general instructions:

*"Hi all robots, do not crawl anything on my site...but if you're the Google robot, then I have a special instruction for you: you are allowed to crawl all pages on my site"*

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

```
Disallow: /  
User-agent: Googlebot  
Allow: /
```

Semmetrical has done a good study and write-up about this: [Google's Hidden Interpretation of Robots.txt](#)

### \* (Asterisk / wildcard)

With the \* symbol, you tell the robots to match any number of any characters. Very useful for example when you don't want your internal search result pages to be indexed:

*"Hi all robots, do not crawl my search result pages...which would be any urls containing /search.php? with something before and after it"*

```
User-agent: *  
Disallow: */search.php?*
```

Theoretically, you don't need the \* in the end, as the robots assume the url continues anyway (unless you have a \$ symbol in the end). However, [Google themselves use the \\* in the end](#), so better be safe than sorry.

*"Hi all robots, do not crawl any urls containing the word contact"*

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

This would disallow for example:

- /you-can-contact-us-here/
- /contact/form.html
- /company/contact-us.html

### \$ (Dollar sign / ends with)

The dollar sign tells the robots that it is the end of the url.

*"Hi Google robot, don't crawl any .pdf files on my website."*

```
User-agent: Google-bot  
Disallow: *.pdf$
```

*"Hi all robots, in my /help/ category I have some files that end with .php. Don't crawl any of them. But you can crawl all other things in that category."*

```
User-agent: *  
Disallow: /help/*.php$
```

### # (Hash / comments)

You can add comments after the "#" symbol, either at the start of a line or after a directive. This is useful if you want to make it clear for yourself what each section is for:

```
# Instructions for all robots
```

```
User-agent: *  
Disallow: /archives/ # disallow crawling of the archives category
```

### Sitemap:

```
Sitemap: http://www.yourwebsite.com/sitemap.xml
```

As you can see, the *Sitemap:* directive doesn't need the user-agent line. It doesn't matter where you place the *Sitemap:* line in your file, but I prefer it to be either the first or last line in the file.

You can specify more than one XML sitemap file per robots.txt file, but if you have a sitemap index file you can link to only that one and will have the same effect.

### Crawl-Delay: and Request-rate: and Visit-time:

**These directives are not commonly used**, but still worth mentioning.

If you are using any of them, please let me know in the comments below – I'd love to hear if you find them useful, and if you know which robots support these directives.

#### Crawl-Delay:

This directive asks the robot to wait a certain amount of seconds after each time it's crawled a page on your website.

*Hi Yahoo! robot, please wait 5 seconds between your requests.*

```
User-agent: Slurp  
Crawl-delay: 5
```

**Note:** Google recommends you to [set crawl speed via Google Webmaster Tools instead](#).

#### Request-rate:

Here you tell the robot how many pages you want it to crawl within a certain amount of seconds. The first number is *pages*, and the second number is *seconds*.

*Hi all robots, please only crawl 1 page per 5 seconds.*

```
User-agent: *  
Request-rate: 1/5 # load 1 page per 5 seconds
```

#### Visit-time:

It's like opening hours, i.e. when you want the robots to visit your website. This can be useful if you don't want the robots to visit your website during busy hours (when you have lots of human visitors).

```
User-agent: *  
Visit-time: 2100-0500 # only visit between 21:00 (9PM) and 05:00 (5AM) UTC (GMT)
```

**Note:** all times are set in UTC/GMT.

The above isn't widely used (as far as I know). There are other, better ways to achieve what you

The above isn't widely used (as far as I know). There are other, better ways to achieve what you want. For example: implementing LastModified, ETags, LastMod and ChangeFrequency. ...but we'll talk more about that in future blog posts.

---

## How to check and test your robots.txt file

There are a few free tools out there, but I prefer to **use Google's and Bing's Webmaster Tools for a health check** of my websites.

### Google Webmaster Tools

[Google Webmaster Tools](#) is a place where you can check your website based on information from Google. It has plenty of tools and reports, and it's completely free.

To help you **create a robots.txt file**, Google Webmaster Tools has a robots.txt generator tool (please note that it's only focused on Google's robots).

You can **check and test your existing robots.txt** via Google Webmaster Tools > Site Configuration > Crawler access.

You should also **check the problems Googlebot find when it crawls your website**. Go to Google webmaster tools > diagnostics > Web Crawl. There you will see the URLs restricted by robots.txt. You can also see your sitemap errors, HTTP errors, nofollowed URLs and URLs that time out, but we'll go into details for that in another blog post.

The Fetch as Googlebot tool in Webmaster Tools helps you understand exactly how your site appears to Googlebot. This can be very useful when troubleshooting problems with your site's content or discoverability in search results.

Source: [Googlebot](#)

### Bing Webmaster Tools

Just as Google Webmaster Tools, [Bing's Webmaster Tools](#) is free and incredibly useful for website owners.

Go to **Bing webmaster tools > crawl issues**. Beside your robots.txt problems, it also identifies HTTP status code errors, pages infected with malware and many other things you should check regularly.

Once you've got a good robots.txt file built and validated, don't just set it and forget it. Periodically audit the settings in the file, especially after you've gone through a site redesign.

Source: [Bing Webmaster Center Blog](#)

### Other robots.txt resources

Do you want to know more about robots.txt? These articles, guides and blog posts are a few of my favorites on the topic:

- [Robotstxt.org: About robots.txt](#)
  - [Google: Robots.txt Specifications](#)
  - [Google: Fetch as Googlebot](#)
  - [Google:Block or remove pages using a robots.txt file](#)
  - [Google: About Googlebot](#)
  - [Semetrical: Google's Hidden Interpretation of Robots.txt](#)
  - [WPmods: What Do Popular WordPress Websites Put In Their Robots.txt File?](#)
  - [WPmods: The Robots.txt File \(WordPress\)](#)
  - [Yoast: WordPress robots.txt Example](#)
- 

## Funny and creative examples of robots.txt

To round this off, here are some funny and creative examples of robots.txt:

- [Last.fm's robots.txt](#)
- [Explicitly.me's robots.txt](#)
- [TechChuff's robots.txt](#)
- [SEOMoz's robots.txt](#)
- [Shark SEO's robots.txt](#)
- [Malcolm Coles's robots.txt](#)
- [Vzaar's robots.txt](#)
- [Arena Flowers's robots.txt](#)
- [Rachel Reveley's robots.txt](#)

If you know of any other creative robots.txt files, please post a link to them in the comments below.

Make sure you **get notified about the next blog post:**

Receive it via [RSS](#), [Email](#), [Twitter](#) or [Facebook](#).

### About the Author



Location independent Swedish emarketing tigress based in beautiful Cape Town, South Africa...for now. Core member of Prothemer.com and OpenTranslators.org. Loves good coffee, delicious vegetarian food and baking yummie cakes. ;) Say hello to Tess on Twitter: @tessneale | @fortheloveofseo | @prothemer