What is Robots.txt?

Robots.txt are just text files also known as **"Search Engine Robots"** which functions like instructing the web robots about how to crawl pages on a website. It sets a crawling behaviour about how the web or spider bots will actually crawl the website or the different section of the website.

- The robots.txt files indicate whether certain user agents (web-crawling software) can or cannot crawl parts of a website. These crawl instructions are specified by "disallowing" or "allowing" the behaviour of certain (or all) user agents.
- The robots.txt is a part of the **Robots Exclusion Protocol (REP)** which serves as a web standard that regulates the behaviour of the robots to crawl, how to access and index content and serve that contents up to users.

Basic format:

User-agent: [user-agent name]

Disallow: [URL string not to be crawled]

- **Allowing all web crawlers access to all content**

  User-agent: *

  Disallow:

- **Blocking a specific web crawler from a specific folder**

  User-agent: Googlebot

  Disallow: /example-subfolder/

  This syntax tells only Google's crawler (user-agent name Googlebot) not to crawl any pages that contain the URL string www.example.com/example-subfolder/.

- **Blocking a specific web crawler from a specific web page**

  User-agent: Bingbot

  Disallow: /example-subfolder/blocked-page.html

  This syntax tells only Bing's crawler (user-agent name Bing) to avoid crawling the specific page at www.example.com/example-subfolder/blocked-page.

How does robots.txt work?

Search engines have two main jobs:

1. Crawling the web to discover content;
2. Indexing that content so that it can be served up to searchers who are looking for information.

Search Engine Robots crawl sites by following links from one site to another. When a search engine bot reaches a website, first it looks for the robots.txt file because it contains information about the behaviour of how to crawl the particular site. If the site does not have a robots.txt file it will proceed crawling pieces of information.

Meta Robots Tag

Meta robots tag is a term in SEO, if you want to block the bots from crawling and indexing a specific page on your website then Meta robots tags are the perfect tag to use. Your robots Meta tag should look like this and be placed in the <head> section of your website:

<meta name=”robots” content=”noindex”>

If you want to disallow a crawler from indexing the content on your page and prevent it from following any of the links, your meta robots tag would look like this:

<meta name=”robots” content=”noindex, nofollow”>

Different Meta Robot Tags commands:

- **Index** – All search engines are able to index the content on this web page.
- **Follow** – All search engines are able to crawl through the internal links on the web page.
- **Noindex** – will prevent the designated page from being included in the index.
- **Nofollow** – will prevent Google bots from following any links on the page. Note that this is different to the rel=”nofollow” link attribute.
- **Noarchive** – prevents cached versions of the page from showing in the SERPs.
- **Nosnippet** – prevents the page being cached and descriptions appearing below the page in the SERPs.
- **NOODP** – prevents the Open Directory Project description for the page replacing the description manually set for this page.
- **Noimageindex** – prevents Google indexing of the images on the page.
- **Notranslate** – prevents the page being translated in the Google SERPs.

You can use multiple commands in your meta robots tag. If you want to prevent a page on your website from being cached by all search engines and also prevent Open Directory descriptions replacing your current descriptions, you would use the following commands: noarchive and NOODP. Your meta robots tag would look like this:

<meta name="ROBOTS" content="NOARCHIVE, NOODP">

Meta Robots tag vs Robots.txt

Robot.txt files are best to use when you are likely to disallow a whole section of a site whereas Meta robots tags function more efficiently if you want to disallow single or particular files or pages of the site.

- If you want to deindex a page or directory from Google's Search Results then we suggest that you use a "Noindex" Meta tag rather than a robots.txt directive as by using this method the next time your site is crawled your page will be deindexed, meaning that you won't have to send a URL removal request. However, you can still use a robots.txt directive coupled with a Webmaster Tools page removal to accomplish this.
- Meta robots tag also ensures that your link equity is not being lost, with the use of the 'follow' command.

You could choose to use both a Meta robots tag and a robots.txt file as neither has authority over the other, but "noindex" always has authority over "index" requests.