Robots Exclusion Protocol Guide

The Robots Exclusion Protocol (REP) is a simple but powerful mechanism that webmasters and SEOs can use to instruct automated web crawlers such as search engine bots what parts of their websites not to crawl. Perhaps it is the simplicity of the robots.txt file that often causes it to be overlooked, which sometimes results in critical SEO issues. **This guide contains tips and examples to assist with the implementation and management of your robots.txt file.**

Key Points

- The robots.txt file defines the Robots Exclusion Protocol (REP) for a website. The file defines directives that exclude web robots from directories or files per website host. (Typically, these are search engine robots. However, there are other robots that adhere to the REP; see section “Web Robots” below.)
- The robots.txt file defines crawling directives, not indexing directives.
- Good web robots such as Googlebot, Bingbot, Yahoo Slurp and DuckDuckBot adhere to directives in your robots.txt file. Bad web robots may not. Do not rely on the robots.txt file to protect private or sensitive data.
- A robots.txt file is publicly accessible, so do not include any files or folders that may include business-critical information or confidential content.

  For example:
  - Website analytics folders (/webstats/, /stats/, etc.)
  - Test or development areas (/test/, /dev/)
- If a URL redirects to a URL that is blocked by a robots.txt file, the first URL will be reported as being blocked by robots.txt in Google Search Console (even if the URL is listed as allowed in the robots.txt analysis tool).
- Search engines may cache your robots.txt file. (Google may cache a robots.txt file for 24 hours). Update relevant rules in a robots.txt file 24 hours prior to adding content otherwise excluded by current REP instructions.
- To notify Google of changes to your robots.txt file and expedite its indexing and caching, use the “Submit” function in Google Search Console’s robots.txt Tester at https://www.google.com/webmasters/tools/robots-testing-tool.
- When deploying a new website from a development environment, always check the robots.txt file to ensure no key directories are excluded.
- Excluding files using robots.txt may not save (or redistribute) the crawl budget from the same crawl session. For example, if Google cannot access a number of files, it may not crawl other files in their place.
- URLs excluded by REP may still appear in a search engine index. Major search engines such as Google do not crawl or index content blocked by robots.txt, but they might find and index those URLs from other places on the web.

  For example, the search engine robot may not have revisited a website and processed the updated directives. Yet the search engine found the URL via external links and stored a reference to the URL. In this case, the engine uses information from the external sources (such as anchor text and text surrounding inbound links) to make judgments about the page. Link popularity of an excluded page may influence the page to be indexed.
- URLs excluded by robots.txt can accrue PageRank.
- This guide includes references to additional robots.txt functionality that was not part of the original specification (http://www.robotstxt.org).

Key Requirements

- File name must be lowercase (“robots.txt”) and must be publicly accessible.
- File type must be a standard file format (such as ASCII or UTF-8). File must be located at the root (i.e., highest level directory) of a website host.

  For example:
  - https://example.com/robots.txt
  - https://www.example.com/robots.txt
  - https://subdomain.example.com/robots.txt
- For websites with multiple subdomains, each subdomain should have its own robots.txt file residing at the root directory of that subdomain (for example, https://subdomain.example.com/robots.txt and https://www.example.com/robots.txt).
- A robots.txt file can be used for different protocols (such as https: and ftp:). However, each protocol needs its own robots.txt file. For example, http://example.com/robots.txt does not apply to pages under https://example.com.
- Search engines may have a robots.txt length limitation. For example, Googlebot limits robots.txt files to 500KiB; any directives beyond 500KiB are truncated.

Web Robots

A web robot (or web crawler, web spider or Internet bot) is a computer program that browses the World Wide Web in a methodical, automated manner. This process is called crawling or spidering. Many sites, in particular search engines, use crawling as a means of indexing up-to-date data.

Web robots should request a robots.txt file when it accesses a website host; however, some robots may cache the robots.txt file or ignore it altogether.

Web robots are typically used for:

- Checking links (e.g., Majestic, Ahrefs, Moz Link Explorer)
- Validating HTML code (e.g., W3C validation tool)
- Site diagnostic spidering tools (e.g., Screaming Frog, DeepCrawl, Xenu, SEOToolSet)
- Harvesting e-mail addresses (usually for spam)
- Scraping content (usually for spam)
- Translation services (e.g., Bing Translator, Google Translate).
- Downloading websites or caching websites locally for viewing later (e.g., winHTTrack)
- Creating an archive for historical purposes (e.g., Wayback Machine archive.org)
- Vertical search (specific file types, images, video, audio, torrents, file archives)

Structure of a robots.txt File

Introduction

There are a number of typical directives valid for the most common web robots (referred to in the robots.txt file as User-agents).

Typical structure:

User-agent:] - Name of the web robot
Directives] - Rules for the robot(s) specified by the User-agent

Different web robots (identified as user-agents) may interpret non-standard directives differently.

You can have many directives. Each directive must be on a separate line.

Each directive consists of an **element: instruction** pair (such as `Disallow: /webmail/`).

The elements are:

- User-agent:
- Disallow:
- Allow:
- Sitemap:
- Crawl-delay:
- Host:
- # (a comment declaration)

Each element must be in title case (i.e., starts with a capital letter followed by lowercase letters).

Each element must be followed by a colon (:) and a space before the instruction.

Each instruction shows only the path portion of the URL (for example, folders and files off the root of the URL do not include the protocol or domain in the instruction).

The instructions are matched from the left to the right, meaning that robots are blocked from anything that begins with /”pattern”.

Each instruction is case-sensitive.

For example:

```
Disallow: file.html
```

This rule does not prevent search engines from crawling File.html.

User-agent:

The User-agent specifies the web robot for which the rules that follow it apply.

The User-agent can refer to a single web robot or all user-agents (indicated by the wildcard character "*").

For example:

`User-agent: HAL`] - The following directives apply only to HAL.
`User-agent: *`] - The following directives apply to all web robots.

A list of common user-agents can be found at the websites below:

- http://www.useragentstring.com/pages/useragentstring.php
- https://www.robotstxt.org/db.html

Disallow:

- The Disallow rule specifies the folder, file or even an entire directory to exclude from web robots' access.

Examples:

1. Allow robots to spider the entire website

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

2. Disallow all robots from the entire website

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

3. Disallow all robots from "/myfolder/" and all subdirectories of "myfolder"

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

4. Disallow all robots from accessing the specific page "private-file.html"

```
User-agent: *
Disallow: /private-file.html
```

5. Disallow Googlebot from accessing files and folders beginning with "my"

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

Allow:

- The Allow rule is a non-standard rule that enables a webmaster to provide more granular access or complex rules.
- You can use the Allow directive to give robots access to a particular URL that resides in a disallowed directory. It refines previous Disallow statements.

For example:

```
Disallow: /scripts/
Allow: /scripts/page.php
```

This tells all robots that they may fetch http://example.com/scripts/page.php (and http://example.com/scripts/page.php?article=1, etc.) but not any other URL in the http://example.com/scripts/ folder.

- Allow takes precedence over Disallow as interpreted by Google, Bing and Yahoo; however, try to avoid contradictory directives as this may become unmanageable or cause unpredictable results with different robots.

Sitemap:

- The Sitemap declaration tells web robots where to find the XML sitemap or XML sitemap index file.
- Many search engines will attempt to auto-discover the XML sitemap via the Sitemap declaration in a robots.txt file.
- The Sitemap element must point to an absolute URL (for example, `Sitemap: https://www.example.com/sitemap.xml`), unlike other elements.
- A robots.txt file can have multiple Sitemap declarations.
- The Sitemap declaration can point to the standard uncompressed XML file or the compressed version.
- Sitemap auto-discovery via robots.txt does not replace sitemap submissions via Google Search Console and Bing Webmaster Tools, where you can submit your sitemaps and obtain indexation statistics.
- If your XML sitemap contains business-critical data that you do not want your competitors to see, do not use this instruction. Instead rename your XML sitemap file so that it cannot be easily guessed and submit it through Google Search Console and Bing Webmaster Tools.

Crawl-delay:

- The Crawl-delay directive requests robots to pause between subsequent page requests.
- Google does not support the Crawl-delay directive. To set a crawl delay for Googlebot, you should use the Google Search Console Site Settings to set the crawl rate by visiting https://www.google.com/webmasters/tools/settings.
- Yahoo supports Crawl-delay. The range specified by Yahoo is 0–10.
- Yahoo supports decimal numbers, but no longer references a delay in seconds. The crawl delay is a relative reduction in crawling speed.
- Bing supports Crawl-delay. The range specified by Bing is positive, whole numbers only from 1 to 20.

No Crawl-delay specified means normal crawl rate.

```
Crawl-delay: 1        ] – Slow
Crawl-delay: 10       ] - Very slow
Crawl-delay: 20       ] - Extremely slow
```

- Avoid Crawl-delay if possible or use with care, as this directive can significantly affect the timely and effective spidering of a website.

Wildcard Characters

* — The Wildcard Character

- Google, Bing, Yahoo and Ask support a limited form of “wildcards.”
- The asterisk (*) is the wildcard character that represents zero or more characters.
- It can be used to apply directives to multiple robots with one set of rules as well as to indicate zero or more unspecified characters in instructions.

For example:

The following rule disallows Googlebot from accessing any page containing “page” in the URL.

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

This rule excludes the directory /frontpage/ and all its files and subfolders from Googlebot and, thus, from Google’s index.

- The wildcard character (*) can also mean no character.

For example:

```
Disallow: /*gallery/default.aspx
```

This rule excludes /picture-gallery/default.aspx as well as /gallery/default.aspx.

$ — The End of Line Wildcard

- The $ character signifies any URL that ends with the preceding characters.

For example:

```
Disallow: /webmail/
Allow: /webmail/$
```

This rule excludes all files and subfolders of the directory but allows access to the landing page.

Combining * and $

- You can combine $ and * wildcard characters for Allow: and Disallow: directives.

For example:

```
Disallow: /*asp$
```

This rule disallows all asp files, but does not disallow files with query strings or folders, due to the $.

Excluded — /pretty-wasp
Excluded — /login.asp
Not excluded — /login.asp?forgotten-password=1

Language Encoding

- The Bing Webmaster Blog has a great guide on character encoding in a robots.txt file: https://blogs.bing.com/webmaster/2009/11/05/robots-speaking-many-languages/

Interesting robots.txt Files

Some interesting robots.txt files are:

- ➢ https://en.wikipedia.org/robots.txt (no bad bots ... pretty please)
- ➢ https://ebay.com/robots.txt
- ➢ https://www.robotstxt.org/faq/legal.html (Can REP be used in a court of law?)
- ➢ https://www.last.fm/robots.txt
- ➢ https://en.wikipedia.org/wiki/Three_Laws_of_Robotics (Isaac Asimov would be proud!)

Resources for More Information on robots.txt

For specific details regarding the interpretation of REP directives by the key search engines, see the following websites:

Google Search Console Help

- How to use the robots.txt tester
  https://support.google.com/webmasters/answer/6062598?hl=en

Google Developers Help

- How to create a robots.txt file
  https://developers.google.com/search/docs/advanced/robots/create-robots-txt
- How Google interprets the robots.txt specification
  https://developers.google.com/search/docs/advanced/robots/robots_txt

Bing Webmaster Help

- How to create a robots.txt file
  https://www.bing.com/webmasters/help/how-to-create-a-robotstxt-file-cb7c31ec

Disclaimer

Some tips documented in this reference may be experimental or unofficially supported. Always verify robots exclusion protocol directives using a robots.txt validator available at:

- Google Search Console robots.txt Tester
  https://www.google.com/webmasters/tools/robots-testing-tool

If you have additional questions, search these resources:

- Google Search Console Help
  https://support.google.com/webmasters/?hl=en#topic=9128571

- Bing Webmaster Tools
  https://www.bing.com/webmasters/about

Like this guide? Find more resources for SEOs and marketers at BruceClay.com:

- *SEO Guide* (https://www.bruceclay.com/seo/search-engine-optimization/)
  This self-paced, hands-on tutorial teaches you ethical search engine optimization step by step, with free tools included.

- *Bruce Clay Blog* (https://www.bruceclay.com/blog/)
  With in-depth articles on search engine optimization, pay-per-click ad management, SEO tools and content development, the Bruce Clay Blog has earned its place among the search industry's most respected blogs.

- *SEO Training* (https://www.seotraining.com/)
  The full Bruce Clay SEO Training is available online as part of our SEOtraining.com membership platform. As a member, you can keep learning with monthly live Q&As, where you can ask your questions directly, and many additional resources, including in-depth video mini-courses, downloadable e-books, brief AUA videos to answer specific questions, a subscription to the SEOToolSet and more.

If you're wringing your hands over an SEO project, a site redesign or just a website that isn't meeting your goals, we can help. Call us for a free services quote and consultation to find out how.

BRUCE
CLAY.

Contact us
1-866-517-1900 (toll free) • 1-805-517-1900
https://www.bruceclay.com/