

Robots Exclusion Protocol Guide

The Robots Exclusion Protocol (REP) is a very simple but powerful mechanism available to webmasters and SEOs alike. Perhaps it is the simplicity of the file that means it is often overlooked and often the cause of one or more critical SEO issues. To this end, we have attempted to pull together tricks, tips and examples to assist with the implementation and management of your robots.txt file. As many of the non-standard REP declarations supported by Google, Yahoo and Bing may change, we will be providing updates to this in the future.

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 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 (GoogleBot, Yahoo Slurp and Bing MSNbot) 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 from search engines.



The robots.txt file is publicly accessible so do not include any files or folders that may include business critical information.

For example:

- Website analytics folders (/webstats/, /stats/ etc).
- Test or development areas (/test/, /dev/).
- XML Sitemap element if your URL structure contains vital taxonomy.



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 Webmaster Tools (even if the URL is listed as allowed in the robots.txt analysis tool).



Search engines may cache your robots.txt file (For example Google may cache your robots.txt file for 24 hours). Update relevant rules in the robots.txt file 24 hours prior to adding content otherwise excluded by current REP instructions.



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.

For example:

- The search engine robot has not revisited a website and processed the updated directives.
- The search engine identified the URL via external links to the URL and stored the reference to the URL. In this case search engines will use information from these external sources such as anchor text and surrounding text of inbound links to make judgments about the page. Link popularity of an excluded page may be a factor to cause the page to be indexed.



URLs excluded by robots.txt can accrue PageRank.



This guide includes references to additional robots.txt functionality that were not part of the original specification (<http://www.robotstxt.org>).

Key requirements



File must be lower-case (for example, "robots.txt").



File must be publicly accessible.



File type must be in a standard file format (for example, ASCII or UTF-8).



File must be located at the root of a website host.

For example:

- <http://example.com/robots.txt>.
- <http://www.example.com/robots.txt>.
- <http://subdomain.example.com/robots.txt>.



File is also valid for secure versions of a domain (For example, <https://www.example.com/robots.txt>).



Search engines may have a robots.txt length limitation.

Web robots



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



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



Typical uses of Web robots to be aware of are:

- Checking links (Open Site Explorer).
- Validating HTML code (W3C validation tool).
- Checking URLs (Xenu).
- Harvesting e-mail addresses (usually for spam).
- Scraping content (usually for spam).
- Translation services (Yahoo Babblefish, Google Translate).
- Downloading websites or caching websites locally for viewing later (winHTTrack).
- Creating an archive for historical purposes (Wayback Machine archive.org).
- Vertical search (specific file types, images, video, audio, torrents, file archives).

Structure

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 (User-agents) may interpret non-standard directives differently.



Each directive must be on a separate line.



Each directive consists of an element: instruction pair. (For example, Disallow: /webmail/).



The elements are:

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



Each element must be in word case (start with a capital letter, following letters lower-case).



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



Each instruction matches on the URI (folders and files off the root of the URL, i.e., do not include the 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.

User-agent:



The User-agent specifies the Web robot for which the following rules 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.user-agents.org/>
- <http://www.useragentstring.com/>

Disallow:



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

For example,

```
# Allow robots to spider the entire website
User-agent: *
Disallow:
```

```
# Disallow all robots from the entire website
User-agent: *
Disallow: /
```

```
# Disallow all robots from "/myfolder/" and all subdirectories of "myfolder"
User-agent: *
Disallow: /myfolder/
```

```
# Disallow all robots from accessing any file beginning with "myfile.html"
User-agent: *
Disallow: /myfile.html
```

```
# Disallow "googlebot" from accessing files and folders begining with "my"
User-agent: googlebot
Disallow: /my
```

Allow:



The Allow rule is a "non-standard" rule that allows a webmaster to provide more granular access or complex rules.



Refines previous "disallow" statements.

```
# Disallow all robots from the /scripts/ folder except page.php
Disallow: /scripts/
Allow: /scripts/page.php
```

```
# Tells robots that they may fetch http://example.com/scripts/page.php,
# or http://example.com/scripts/page.php?article=1,
# but not any other URL in http://example.com/scripts/ folder.
```



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

For example:

```
User-agent: *
Disallow: /example.php
Allow: /example.php
```

Noindex:



The noindex directive is "unofficially" only supported by Google.



The Noindex directive behaves per the Disallow directive and in addition, removes all matching site URLs from Google.



Use the Noindex directive with care as behavior or support may change.

Sitemap:



The Sitemap declaration points to the XML Sitemap or XML Sitemap index file.



The sitemap element must point to an absolute URL (unlike other elements). For example, Sitemap: <http://www.example.com/sitemap.xml>.



A robots.txt file can have multiple sitemap declarations.



The sitemap declaration can point to the standard uncompressed XML file or the compressed version.



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 so that it cannot be easily guessed and submit it through Google, Yahoo! and Bing webmaster Tools.




Many search engines will attempt to auto discover the XML Sitemap via the sitemap declaration in a robots.txt file.





Sitemap auto discovery via robots.txt does not replace sitemap submissions via Google, Yahoo and Bing webmaster tools where you can submit your sitemaps and obtain indexation statistics.


Crawl-delay:


 The Crawl-delay directive requests robots to pause between subsequent page requests.


 Google does not support the crawl-delay directive.

 Yahoo supports crawl-delay. Ranges specified by Yahoo range from 0 – 10.

 Yahoo supports decimal numbers however no longer reference a delay in seconds. The crawl delay is a relative reduction in crawling speed.


 Bing supports crawl-delay. Ranges specified by Bing range from 1 - 10.

 Bing does not recommend using values higher than 10.


 Bing supports positive, whole numbers only.

For example:

No crawl-delay	-	Normal
1	-	Slow
5	-	Very slow
10	-	Extremely slow

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


* - The wildcard character

 The asterisk (*) is the wildcard character. It can apply directives to multiple robots with one set of rules or to indicate one or more characters when declaring instructions.

```
# The following rule will disallow googlebot from accessing any URL
# containing "page"
User-agent: googlebot
Disallow: /*page
```

```
# This rule excludes the following files and folders from googlebot
# and thus being indexed in Google
# beauty-pageants.php
# /myfolder/example-page.php
```


```
# /frontpage/ (and all subfolders and files in this directory)
```

 The * can also mean “no character”

For example:

```
Disallow: /*gallery/default.aspx
# Excludes /picture-gallery/default.aspx
# Also excludes /gallery/default.aspx
```


\$ - The end of line wildcard


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

For example:

```
# Exclude all sub files and folders of a directory but allow
# Access to the landing page
Disallow: /webmail/
Allow: /webmail/$
```

Combining * and \$: Examples

 Can combine \$ and * wildcard characters.

 Can be combined for allow and disallow directives.

For example:

```
# Disallow all asp files
Disallow: /*asp$
# This will not exclude files with query strings or folders due to the $
# Excluded - /pretty-wasp
# Excluded - /login.asp
# Not excluded - /login.asp?forgotton-password=1
```

Language encoding

 The Bing Toolbox has a great guide on character encoding in a robots.txt file:
<http://www.bing.com/toolbox/blogs/webmaster/archive/2009/11/05/robots-speaking-many-languages.aspx>

Interesting robots.txt files



Some interesting robots.txt files are:

- <http://en.wikipedia.org/robots.txt> (no bad bots...pretty please)
- <http://ebay.com/robots.txt> (can REP be used in a court of law? See <http://www.robotstxt.org/faq/legal.html>)
- <http://siteriver.com/robots.txt> (selectively allow image types)
- <http://www.last.fm/robots.txt> (Isaac Asimov would be proud; see http://en.wikipedia.org/wiki/Three_Laws_of_Robotics)

Resources



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

- Google Webmaster Tools Help
<http://www.google.com/support/webmasters/>
- Bing Webmaster Center Help
http://help.live.com/help.aspx?mkt=en-au&project=wl_webmasters
- Yahoo! Search Help
<http://help.yahoo.com/l/us/yahoo/search/>
- Ask.com Webmaster Help
<http://about.ask.com/en/docs/about/webmasters.shtml>

Disclaimer



Many of the features and tips documented in this reference may be experimental or unofficially supported. Always verify REP directives using a robots.txt validator available at:

- Google Webmaster Tools - <https://www.google.com/webmasters/>
- Bing Toolbox - <http://www.bing.com/toolbox/webmasters/>

Acknowledgements



Special thanks to John Mueller (@johnmu) of Google for clarifying a few GoogleBot behaviors.