The robots.txt standard – implementations and usage

Sebastian Nagel
sebastian@commoncrawl.org

OSSYM 2022 – 4th International Open Search Symposium,
Geneva, Switzerland, Oct 10–12, 2022

The robots.txt standard

- allows web masters to signal web crawlers how to best crawl their sites

- a text file `robots.txt` is deployed in the root folder of a web site (eg. `http://example.org/robots.txt`)
- readable for web crawlers (“robots”)
- contains policies how crawlers shall access the site’s content

Example robots.txt

```
                    # (this is a comment)
User-agent: badbot  # rules in block apply to "badbot" only
Disallow: /         # disallow everything

User-agent: goodbot # next block: rules for "goodbot"
Disallow: /login/   # disallow paths below /login/
                    # (everything else is implicitly allowed)

User-agent: *       # wildcard rule block applies to all other bots
Allow: /news/       # paths below /news/ are explicitly allowed
Disallow: /         # all other paths must not be visited
```

Example URLs (dis)allowed if robots.txt found at `http://example.org/robots.txt`

badbot	goodbot	mybot	URL
✗	✓	✗	`http://example.org/index.html`
✗	✓	✓	`http://example.org/news/amazing-news.html`
✗	✗	✗	`http://example.org/login/signup?lang=en-US`

Real-world robots.txt

```
User-agent: Googlebot-News
Disallow: /angebote/

User-agent: *
Disallow: /zeit/
Disallow: /templates/
Disallow: /hp_channels/
Disallow: /send/
Disallow: /suche/
Disallow: /rezepte/suche/
Disallow: */comment-thread?
Disallow: */liveblog-backend*
Disallow: /framebuilder/
Disallow: /campus/framebuilder/
Disallow: /cre-1.0/tracking/*.js$

User-agent: Baiduspider
Disallow: /

User-agent: Applebot
Allow: /
Disallow: /cre-1.0/

User-agent: GrapeshotCrawler
crawl-delay: 3

Sitemap: https://www.zeit.de/gsitemaps/index.xml
```

- `https://www.zeit.de/robots.txt`
- Googlebot-News and Applebot ev. preferred (more paths allowed)
- Baiduspider penalized
- GrapeshotCrawler [1] to wait 3 seconds between requests
- default rule set excludes templates, duplicated dynamic content or user comments
- improve quality of crawled content and search results!
- the announced sitemap provides an up-to-date list of URLs (without duplicates)

Legal status

- a technical solution to coordinate different interests between the owners of content and robots
- a convention based on consensus not a legally binding regulation

> *The robots exclusion protocol has no formal status; it is not explicitly recognised in statutes or international conventions as a binding instruction to (managers of) robots. It is also not a formal standard, i.e. a standard brought about by one of the formal standard setting institutes. It is also not dealt with in an RFC (Request For Comment), i.e. a document specifying what internet protocols should look like. The protocol is based on a consensus reached on 30 June 1994 on the robots mailing list (robots-request@nexor.co.uk), between the majority of robot authors and other people with an interest in robots.*
>
> Schellekens 2013, Are internet robots adequately regulated? [2]

Adoption and standardization efforts

1994 robots.txt protocol discussed on mailing list [5]

1996 inofficial RFC proposal [6]

- adopted by all major web search engines
- various extensions, conflicting specifications and implementations

2019 RFC draft [7, 8] and reference implementations [9]

2022 RFC 9309 [10]

Implementation details

implementation details and changes from 1994 until 2022

- fine-grained access rules with * and $ pattern markers
- practical and clear definition how to resolve competing
    - allow and disallow directives (multiple paths would match)
    - user-agent line matches
- fetching the robots.txt
    - HTTP status codes
    - size limit and caching policies
- RFC 9309 is an improvement over initial RFC proposal!
- see list of implementation details and extensions in appendix

Extensions

robots meta tag

```
<meta name="robots" content="noindex, nofollow">
```

- page-level directives, supplemental to root-level robots.txt
- noindex do not index
- nofollow do not follow links
- many more to influence how pages are presented on search result pages: nocache, nosnippet, max-snippet, ...

additional robots.txt directives

- allowed by RFC 9309, but not required to be respected
- Sitemap, Crawl-delay, ...
- not all proposed directives were adopted, eg. from [11]

  ```
  Visit-time: 0600-0845
  ```

Summary and outlook: what robots.txt is (and is not)

- a technical recommendation and convention
- not a legally binding regulation
- broadly adopted, but diverging implementations and extensions, standardized as RFC 9309 very recently

- no security feature to hide confidential information
  - no guarantee that every search engines supports robots.txt or the same set of directives [13]
- no copyright control
  - robots meta tags provide some sort of (nosnippet, nocache)

- what it wasn't meant for?
  - introduce bias and favor one search engine over others [14, 15]
  - censorship [3]

Analyzing robots.txt usage on web sites

- six years of robots.txt files archived at Common Crawl [17]
  - one crawl analyzed per year (run in August or September)
- robots.txt records of 10,000 top-ranking domains
  - harmonic centrality ranks calculated on latest CC domain-level web graphs [18]
  - select most recent robots.txt capture of domain "home site" (domain.com or www.domain.com)
  - full analysis of all robots.txt captures would be biased towards the long tail and domains with many subdomains
- missing data points because of
  - site and robots.txt not visited by crawler
  - domain not registered in years before 2022
- detailed results and code available on

  `https://github.com/sebastian-nagel/ossym2022-robotstxt-experiments`

Robots.txt usage

robots.txt crawl	found %	with rules %
2016-36	72.21	67.54
2017-34	71.56	66.99
2018-34	75.21	70.34
2019-35	75.94	71.20
2020-34	76.58	71.89
2021-39	76.77	72.32
2022-33	75.88	71.61

- 70% of top-10k domains with parseable robots.txt
- 35% resp. 38.5% were reported for 2005/2006, based on 7.5k web sites [12]

User-agents addressed

	2016	2017	2018	2019	2020	2021	2022
(any)	6754	6699	7034	7120	7189	7232	7161
*	6632	6578	6911	7012	7075	7107	7034
googlebot	430	402	461	448	463	467	453
twitterbot	251	310	363	364	408	447	448
mediapartners-google	369	353	341	335	326	317	297
ahrefsbot	154	164	187	237	287	294	297
adsbot-google	93	97	95	202	218	249	247
bingbot	173	193	211	226	236	239	241
mj12bot	127	135	165	180	214	230	224
semrushbot	36	48	95	145	189	213	220
baiduspider	190	190	204	214	222	216	213
yandex	167	170	201	209	200	211	209
ia_archiver	193	173	191	178	189	187	185
dotbot	75	86	118	129	152	173	170
googlebot-news	87	101	125	152	156	165	152
googlebot-image	128	134	133	147	160	169	148
slurp	172	177	180	167	168	161	146
msnbot	171	157	156	147	145	135	113

User-agents addressed

- rules for the wildcard * user-agent are almost always provided
- 6% of robots.txt Googlebot, the most commonly addressed "named" user-agent

Length of allow/disallow rule sets

- How many allow/disallow statements address a single user-agent (or the wildcard user-agent)?
- robots.txt rule sets can be long, eg. `https://www.etsy.com/robots.txt`

length ruleset	count
1 (disallow: /)	19800
1 (allow: /)	2049
1	1310
2	1358
2-4	969
5-9	1481
10-19	1433
20-49	1567
50-99	784
100-199	305
200-499	201
500-999	46
1000-	24

User-agent bias i

Are some user-agents (or search engines) preferred via robots.txt over others?

- in 2007, [14] counted disallowed path prefixes in 3,000 robots.txt files and found a "strong correlation between the search engine market share and the bias toward corresponding robots"

  > *Such biases may lead to a "rich get richer" situation, in which a few popular search engines ultimately dominate the Web because they have preferred access to resources that are inaccessible to others.*

- in 2008, [15] found support for this thesis by counting the number of disallowed URLs for Yahoo and Google crawlers
- in 2015, Apple announced to follow Googlebot's rules (instead of the wildcard user-agent) if there are no specific rules for Applebot [15, 19]. Neevabot also applies this policy [20]
- in 2020, [21] found further support by manually analyzing few robots.txt files

User-agent bias ii

- policies which restrict the robot access to agreed agents are known, eg. https://www.linkedin.com/robots.txt

```
...

User-agent: *
Disallow: /

# Notice: If you would like to crawl LinkedIn,
# please email whitelist-crawl@linkedin.com to apply
# for white listing.
```

- to get recent measures, we simply count which of the top-10k domains grant user-agents unlimited, partial or no access

User-agent bias iii

	addressed	allow-part	disallow-all	allow-all
twitterbot	448	6018	58	1512
mediapartners-google	297	5967	74	1547
googlebot	453	6191	38	1359
bingbot	241	6174	55	1359
adsbot-google	247	6106	75	1407
msnbot	113	6148	65	1375
googlebot-news	152	6142	74	1372
googlebot-image	148	6113	82	1393
slurp	146	6131	79	1378
applebot	46	6150	75	1363
*	7034	6145	78	1365
neevabot	2	6144	79	1365
seznambot	36	6137	93	1358
ccbot	44	6108	117	1363
yandex	209	6102	134	1352
baiduspider	213	6068	156	1364
petalbot	110	6072	166	1350
ia_archiver	185	6058	177	1353
dotbot	170	6020	222	1346
semrushbot	220	5979	272	1337
mj12bot	224	5960	278	1350
ahrefsbot	297	5938	318	1332

User-agent bias iv

- a correlation between market share and preference in robots.txt rules seems to be visible
- search engines focused on regional markets, archive and SEO crawlers are even more penalized
- although – we cannot evaluate whether partial restrictions differ between robots
- …does the policy of Applebot and Neevabot pay off?

	addressed	allow-part	disallow-all	allow-all
twitterbot	448	6018	58	1512
mediapartners-google	297	5967	74	1547
googlebot	453	6191	38	1359
neevabot (googlebot)	2	6190	39	1359
applebot (googlebot)	46	6193	43	1352
bingbot	241	6174	55	1359
…				

Questions?

References i

(web resources visited on 2022-10-07)

[1] *Oracle Data Cloud Crawler*. https://www.oracle.com/corporate/acquisitions/grapeshot/crawler.html.

[2] MHM Schellekens. "Are internet robots adequately regulated?" In: *Computer Law & Security Review* 29.6 (2013), pp. 666–675. DOI: https://doi.org/10.1016/j.clsr.2013.09.003. https://www.sciencedirect.com/science/article/pii/S0267364913001659.

[3] Greg Elmer. *Robots.txt: The politics of search engine exclusion*. 2009.

[4] Greg Elmer. "Exclusionary rules? The politics of protocols". In: *Routledge handbook of internet politics* (2008), pp. 376–383.

[5] Martijn Koster. *A Standard for Robot Exclusion*. 1995. https://www.robotstxt.org/.

[6] Martijn Koster. *A method for web robots control*. 1996. https://www.robotstxt.org/norobots-rfc.txt.

References ii

[7] Martijn Koster et al. *Robots Exclusion Protocol*. Internet-Draft draft-koster-rep-00. Work in Progress. Internet Engineering Task Force, July 2019. 10 pp. https://datatracker.ietf.org/doc/draft-koster-rep/00/.

[8] Henner Zeller, Lizzi Sassman, and Gary Illyes. *Formalizing the robots exclusion protocol specification*. 2019. https://developers.google.com/search/blog/2019/07/rep-id.

[9] *Google Robots.txt Parser and Matcher Library*. https://github.com/google/robotstxt.

[10] Martijn Koster et al. *Robots Exclusion Protocol*. RFC 9309. Sept. 2022. DOI: 10.17487/RFC9309. https://www.rfc-editor.org/info/rfc9309.

[11] Sean Conner. *An Extended Standard for Robot Exclusion*. 2002. http://www.conman.org/people/spc/robots2.html.

[12] Yang Sun, Ziming Zhuang, and C Lee Giles. “A large-scale study of robots.txt”. In: *Proceedings of the 16th international conference on World Wide Web*. 2007, pp. 1123–1124. https://dl.acm.org/doi/abs/10.1145/1242572.1242726.

References iii

[13] Sergey Kratov. “About leaks of confidential data in the process of indexing sites by search crawlers”. In: *International Andrei Ershov Memorial Conference on Perspectives of System Informatics*. Springer. 2019, pp. 199–204.

[14] Y. Sun et al. “Determining bias to search engines from robots.txt”. In: *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, WI 2007*. 2007, pp. 149–155. DOI: `10.1109/WI.2007.98`.

[15] Santanu Kolay et al. “A larger scale study of robots.txt”. In: *Proceedings of the 17th international conference on World Wide Web*. 2008, pp. 1171–1172. `https://dl.acm.org/doi/abs/10.1145/1367497.1367711`.

[16] C Lee Giles, Yang Sun, and Isaac G Councill. “Measuring the web crawler ethics”. In: *Proceedings of the 19th international conference on World wide web*. 2010, pp. 1101–1102. `https://dl.acm.org/doi/abs/10.1145/1772690.1772824`.

[17] *Data Sets Containing Robots.txt Files and Non-200 Responses – Common Crawl*. `https://commoncrawl.org/2016/09/robotstxt-and-404-redirect-data-sets/`.

References iv

[18] *Host- and Domain-Level Web Graphs May, June/July and August 2022.* https://commoncrawl.org/2022/09/host-and-domain-level-web-graphs-may-jun-aug-2022/.

[19] *About Applebot.* https://support.apple.com/en-us/HT204683.

[20] *About Neevabot.* https://neeva.com/neevabot.

[21] *Knuckleheads' Club – The evidence we've found so far.* 2020. https://knuckleheads.club/the-evidence-we-found-so-far/.

[22] Wikipedia contributors. *Robots exclusion standard.* https://en.wikipedia.org/wiki/Robots_exclusion_standard.

[23] *Googlebot.* https://developers.google.com/search/docs/crawling-indexing/googlebot.

[24] *How Google interprets the robots.txt specification.* https://developers.google.com/search/docs/crawling-indexing/robots/robots_txt.

References v

[25] *Overview of Google crawlers (user agents)*. https://developers.google.com/search/docs/crawling-indexing/overview-google-crawlers.

[26] *Robots meta tag, data-nosnippet, and X-Robots-Tag specifications*. https://developers.google.com/search/docs/crawling-indexing/robots-meta-tag.

[27] *Usage of the robots.txt file*. https://yandex.ru/support/webmaster/controlling-robot/robots-txt.html.

[28] *Robots meta tag and HTTP-header X-Robots-Tag*. https://yandex.ru/support/webmaster/controlling-robot/meta-robots.html.

[29] *Which Crawlers Does Bing Use?* https://www.bing.com/webmasters/help/which-crawlers-does-bing-use-8c184ec0.

[30] *How to Create a robots.txt File – Bing Webmaster Tools*. https://www.bing.com/webmasters/help/?topicid=cb7c31ec.

References vi

[31] *Robots meta tags – Bing Webmaster Tools*. https://www.bing.com/webmasters/help/which-robots-metatags-does-bing-support-5198d240.

[32] *SeznamBot crawler*. https://napoveda.seznam.cz/en/seznambot-crawler/.

[33] *Crawling control – Seznam*. https://napoveda.seznam.cz/en/full-text-search/crawling-control/.

[34] *About FacebookBot*. https://developers.facebook.com/docs/sharing/bot.

[35] Martijn Koster. *A Standard for Robot Exclusion*. 1996. https://www.robotstxt.org/meta.html.

[36] *Apple's Applebot Follows Googlebot's Instructions in Robots.txt Files*. 2015. http://www.thesempost.com/apples-applebot-follows-googlebots-instructions-in-robots-txt-files/.

[37] *sitemaps.org*. https://www.sitemaps.org/protocol.html.

References vii

[38] Uri Schonfeld and Narayanan Shivakumar. “Sitemaps: above and beyond the crawl of duty”. In: *Proceedings of the 18th international conference on World wide web*. 2009, pp. 991–1000.

Appendix: robots.txt extensions and implementation details i

- a quick overview over extensions and implementation details
- and how RFCs or crawlers (following the specificiation) handle these
- mentioned RFCs and crawlers
  - NoRobotsRFC [6]
  - RFC9309 [10]
  - Googlebot [23, 24, 25, 26]
  - Yandex [27, 28]
  - Bingbot [29, 30, 31]
  - Seznambot [32, 33]
  - Applebot [19]
  - FacebookBot [34]
  - Neevabot [20]

Appendix: robots.txt extensions and implementation details ii

Robots meta tag

- supplemental to the root-level robots.txt file [35]
- on page level
  - HTML meta tag

    ```
    <meta name="robots" content="noindex, nofollow">
    ```

  - HTTP response header

    ```
    HTTP/1.1 200 OK
    …
    X-Robots-Tag: noindex
    …
    ```

- robots meta directives
  - meta directives proposed in [35] and supported by most search engines

Appendix: robots.txt extensions and implementation details iii

index robots are “welcome” to index the page and include in search results

follow links on this page

noindex and nofollow do not index resp. follow

none same as noindex, nofollow

all same as index, follow

- additional meta directives addressing how results are presented on search result pages

nosnippet no preview text snippet (Googlebot, Applebot)

nocache no link to the cached page (Bingbot)

noarchive same as nocache (Googlebot, Bingbot, Yandex)

max-snippet: <n> snippet length in characters (Googlebot, Bingbot)

max-image-preview: <none|standard|large> and max-video-preview: ... (Googlebot, Bingbot)

Appendix: robots.txt extensions and implementation details iv

        - … and many more, eg. nositelinkssearchbox, notranslate, noimageindex, unavailable_after: … (Googlebot), noyasa no automatic description (Yandex)
    - the definition of all and none may include also (some) additional meta directives
- specify robots meta directives only for Googlebot [26]

  ```
  <meta name="googlebot" content="noindex">
  ```

- <span data-nosnippet> inline exclusion of content from search result snippets (Googlebot)

user-agent

- limitiations on user-agent name (“token”)
    - [-!#$%&'*+.0-9A-Z^_`a-z~]+ (NoRobotsRFC)

Appendix: robots.txt extensions and implementation details v

  - `[a-zA-Z_-]+` (RFC9309)
- match user-agent directives
  - substring match (NoRobotsRFC)
  - full user-agent token (RFC9309)
- select user-agent rule block
  - block of first matched user-agent (NoRobotsRFC)
  - merge multiple matched blocks (RFC9309)
- fall-back user-agent (if "my" user-agent token is unmatched)
  - `*` wildcard (NoRobotsRFC, RFC9309)
  - Googlebot (Applebot, Neevabot, cf. [36])
  - some crawlers specify a hierarchy of user-agent tokens used to select rules, eg. Google's image crawler first looks for Googlebot-Image then for Google [25]

Appendix: robots.txt extensions and implementation details vi

URL path matching

- `*` path pattern: zero or more characters in URL path (Googlebot, Yandex, Seznambot, RFC9309)
- `$` end of path marker: full URL path, not prefix match (Googlebot, Yandex, Seznambot, RFC9309)

  ```
  Disallow: /download/*.zip$
  ```
- `\` and `[<chars>]` (Seznambot)
- competing allow and disallow directives (multiple paths would match)
  - first match (NoRobotsRFC)
  - longest rule / pattern (Googlebot, RFC9309)

Appendix: robots.txt extensions and implementation details vii

Fetching the robots.txt

- 500 kiB size limit (RFC9309)
- caching policy: max. 24 hours (RFC9309)
- HTTP response status code
    - (NoRobotsRFC)

        404 no crawling restrictions

        401,403 access to the site completely restricted

        - temporary failures: defer visits
        - redirects: follow redirects until robots.txt found
    - (RFC9309)

        400-499 “unavailable”: no crawling restrictions

        500-599 “unreachable”: complete disallow

        - redirects: at least five consecutive redirects to be followed

Appendix: robots.txt extensions and implementation details viii

Additional robots.txt directives

- note: RFC9309 mentions additional directives (apart from user-agent, allow and disallow) but does not require crawlers to respect them
- Sitemap link to a sitemap (Googlebot, Bingbot, Yandex, Seznambot) – the sitemap protocol is specified in [37] and is widely adopted [38]
- Crawl-delay: 1.0 wait n seconds between successive requests (Yandex until 2018, Bingbot, Neevabot)
- Request-rate: 10/1m (Seznambot)

Appendix: robots.txt extensions and implementation details ix

- Clean-param URL normalization, remove URL query params (Yandex)

  ```
  Clean-param: ref&sort /forum/*.php
  ```

  `https://example.com/forum-music/showthread.php?sid=123&ref=321&sort=newest`
  normalized to `https://example.com/forum-music/showthread.php?sid=123`
- Host specify preferred domain among mirrors (Yandex, not supported anymore)