Robots.txt – Crawler Politeness in the Age of GenAI

Sebastian Nagel, Thom Vaughan

The Robots Exclusion Protocol (REP)

A text file `robots.txt` is placed in the root folder of a web site and contains access policies that specify which file paths web crawlers ("robots") are allowed to read on the site.

Access policies can be specified for individual crawlers by "user-agent" name, or by a "wildcard" rule block that catches all crawlers not addressed by a named policy.

The REP is a simple but effective way to coordinate the different interests web site owners and content creators, as well as web crawler operators

Adoption and standardization efforts

1994 Robots.txt protocol discussed on mailing list [1]
1996 Inofficial RFC proposal [2]
Adopted by all major web search engines
Conflicting specifications and implementations [3]
Various extensions: `Crawl-delay`, sitemaps [4, 5]
2019 RFC draft [6, 7] and reference implementations [8]
2022 RFC 9309 [9]
2024 RFC drafts extending robots.txt
"REP Extension for URI Level Control" [10], standardizes page-level "robots meta tags" [11]
"User Agent Purpose Extension" [12]
2025 "AI Preferences" vocabulary [13]

Legal Status

Despite its widespread adoption, REP is a technical standard and a convention based on consensus not a legally binding regulation: "it is not explicitly recognised in statutes or international conventions as a binding instruction" [15]

Nevertheless, it is argued that "under certain circumstances, violations of restrictions outlined in robots.txt can lead to legal liabilities." [16]

Recently, with the raise of generative AI (GenAI), there's been a renewed interest in the REP as a way to coordinate between the interests of content owners and the operators of crawlers collecting web data which is used to train AI models.

In this context, it is stated that "ignoring a robots.txt opt-out could negatively impact a fair use assessment." [19]

Similarily, in an assessment of AI training datasets: "The lowest score of 1 point is assigned when data is acquired through circumventing Robots.txt or other problematic methods. Robots.txt represents a website administrator's explicit refusal of crawling, and circumventing it may lead to both moral criticism and legal disputes." [20]

Robots.txt Bias

An early example of bias introduced by the REP is the White House website robots.txt exclusion controversy, when the Bush administration was reported to be blocking search engines from indexing key Iraq war-related documents on the White House website. [21, 22]

In 2007, the authors of [23] counted the 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". The study concludes: "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."

Support for this thesis was found also by [24] in the number of disallowed URLs for the Yahoo and Google web crawlers.

To circumvent the competitive disadvantage introduced by the REP, Apple announced in 2015 it would follow `Googlebot`'s rules (instead of the wildcard user-agent) if there are no specific rules for `Applebot` [25, 26]. `Neevabot` also applied this policy [27].

In the context of GenAI, several authors [28, 29, 16] found that since 2023 a growing portion of websites block GenAI web crawlers via robots.txt and other measures. This leads to a decrease in the availability of training data. The authors of [30, 31] even see the openness of the Web at risk.

About Common Crawl

Common Crawl is a non-profit organization which regularly crawls a significant sample of the web and makes the data accessible free of charge to everyone interested in running machine-scale analysis on web data.

At present, we crawl every month up to 3.0 billion web pages. The data is hosted in the Amazon cloud as part of the AWS Open Data program.

Contact: sebastian@commoncrawl.org
https://commoncrawl.org/ thom@commoncrawl.org

Example 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/

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`
- visited 2022-08-20
- the `*` or "wildcard" user-agent defines rules for any other user-agent not addressed directly
- the "wildcard" rule set excludes templates, dynamic content or user comments; it improves the quality of crawled content and search results
- Googlebot-News and Applebot are preferred (more paths allowed)
- Baiduspider is penalized
- GrapeshotCrawler to wait 3 seconds between requests
- the announced sitemap provides an up-to-date list of crawlable URLs

Experiment: Robots.txt Usage and Bias

The experiment described below analyzes eight years of archived robots.txt files and looks at robots.txt usage and how access policies change over time. The research is an extension of prior work done in 2022. [35, 36] The source code of the experiment is available on Github.

The Data – Sampling Archive Robots.txt Files

- Robots.txt files are archived by Common Crawl's web crawler (`CCBot`) since 2016 [37]
- 2 million top-k sites are selected by combining Tranco lists [38] of the years 2020, 2022 and 2024
- The combined Tranco site ranks are used to define multiple stratified samples (top-1k, 5k, 10k, 100k, 1M, 2M)
- URL index (columnar format, [39]) used to lookup robots.txt URLs and redirect locations
- Download the found WARC records from the web archives and parse the robots.txt files

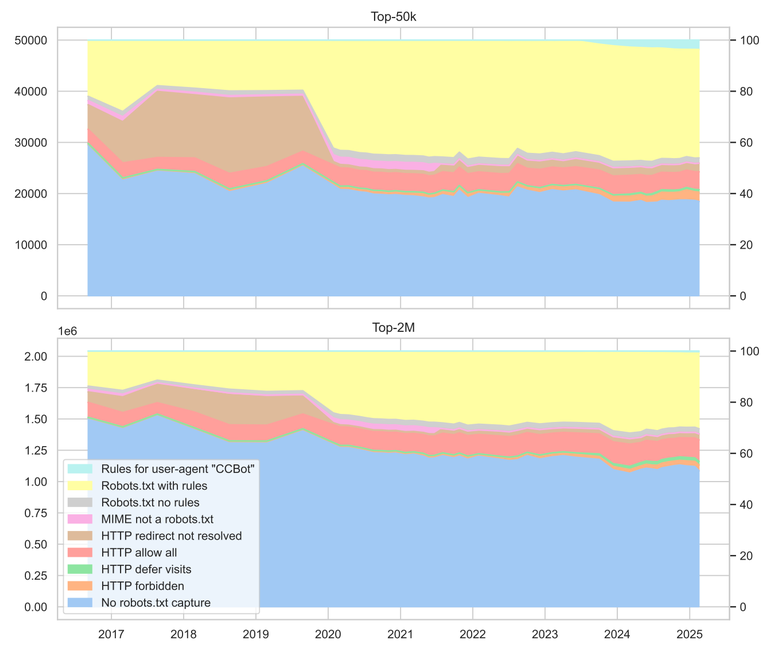

Figure 1. Status of robots.txt captures of top-50k and all (2M) web sites

In the top-50k sample, a valid robots.txt is found for more than 40% of the web sites. In the full sample of 2 million web sites, the number is lower, about 30%.

There are several reasons why there is no robots.txt capture:

- The site wasn't visited in a given monthly crawl. Consequently, there is no robots.txt capture.
- No HTTP connection could be established to fetch the robots.txt.
- The robots.txt wasn't successfully fetched, the server responded with a HTTP status code other than "200 Ok":
  - HTTP 403 "Forbidden" (or equiv.): CCBot will abstain from crawling the site. Note: HTTP error codes are handled differently in the RFC draft [2] and the final RFC 9309 [9]. CCBot follows the more polite RFC draft.
  - HTTP 503 "Server Error" and 429 "Too many requests" (and equiv.): CCBot will defer visits of the site for now.
  - HTTP 404 "Not Found": the site has no robots.txt – implicitely, crawling is "allowed".
  - HTTP 302 "Moved" (or equiv.): redirects are followed.
    Note: since November 2019 the redirect target location is stored in the URL index which allows to easily follow the redirects in this experiment. Originally the crawler followed one redirect, since November 2023 it follows five levels of redirection as specified by RFC 9309 [9].
- The robots is not archived:
  - If it is not a text file, i.e., a HTML page or any other MIME type.
  - If the request is redirected, and the URL path is not `/robots.txt` and is not allowed by the robots.txt on the target site.

  The strict robots.txt archiving policies are imperative to prevent an attacker is able to place secret or sensitive content into the robots.txt archives by redirecting a robots.txt to arbitrary files on a different site.

Two data artifacts visible in Figure 1 can be explained by the collection and sampling methods:

- More redirects are left unresolved before 2019 because redirects were not indexed (although followed).
- Better coverage 2020 – 2024; the Tranco lists are from this period.

But there's one small effect worth noting: in recent years, the number of robots.txt requests that respond with HTTP status codes indicating that crawling the site is not desired has increased. These are shown as "HTTP forbidden" and "HTTP defer visits" in figure 1.

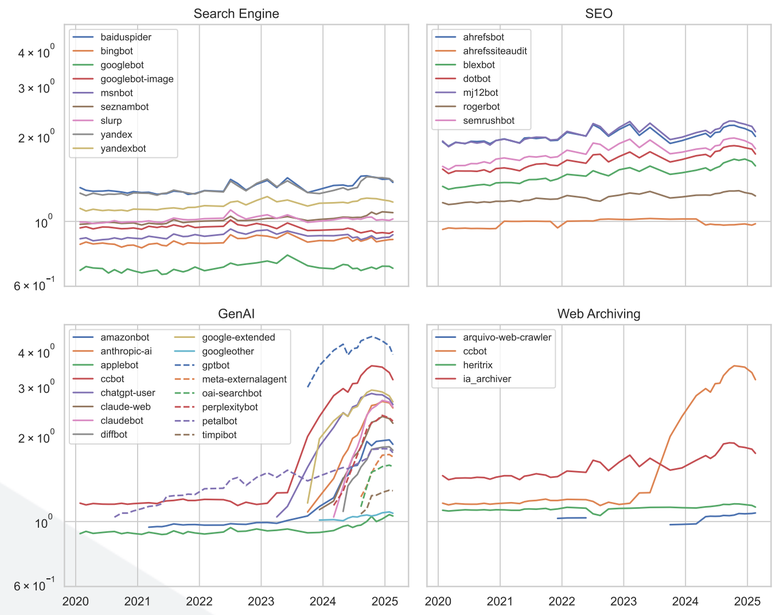

Figure 2. Log ratio of disallowed sites per user-agent, in comparison to the "wildcard" user-agent, top-50k sample. Values lower than 1.0 indicate that a user-agent is preferred, values higher than 1.0 indicate that the user-agent is penalized.

Robots.txt User-Agent Bias

Which user-agents are preferred or penalized? – Figure 2 answers this question by comparing the number of entirely disallowed sites for a given user-agent with that for the wildcard user-agent:

- Search Engine crawlers: two companies (Google and Microsoft) are preferred and are less disallowed than the wildcard user-agent. Baidu and Yandex are penalized, while Yahoo Slurp and Seznam are not, or only to a little extent.
- SEO crawlers are (mostly) penalized with marginal changes over time.
- GenAI crawlers are increasingly blocked since 2023.
- Following the blocking pattern, CCBot is perceived by webmasters as a GenAI crawler, not an archive crawler.

The analysis is based on the top-50k sample. Choosing a different top-k stratum shifts the scale, but generally paints a similar picture of preferred or penalized user-agents.

The amount of sites disallowed for CCBot

There is only a small fraction of sites which address the user-agent "`CCBot`" directly. Since 2023, this share has grown, especially for higher ranking sites, see the top-50k stratum in Figure 1. If "`CCBot`" is addressed directly, it is most likely by the rule "`Disallow: /`" which disallows crawling entirely.

While this trend is visible on all strata in Figure 3, it is more "dramatic" for the higher ranking sites. Here, the percentage of disallowed sites has increased by 10% or more. For the top-1M/2M strata, "the long tail", there is still a significant increase from 3% to 7% disallowed sites.

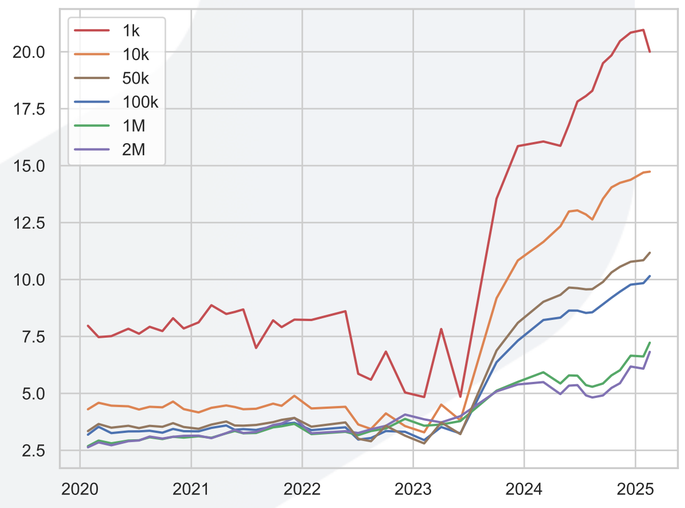

Figure 3. Percentage of top-k sites entirely disallowed for the "`CCBot`" user-agent. Sites without robots.txt capture are ignored.

AI Training Opt-In / Opt-Out Protocols

Several AI training opt-in/out protocols and initiatives exist:

"Text & Data Mining Reservation Protocol (TDMRep)" [41]
- Opt-out mechanism as defined by DSM EU Directive 2019/790
- HTTP headers, HTML metadata and `/.well-known/tdmrep.json`

DECORAIT — A decentralized opt-in/opt-out registry [42]

"IAB Workshop on AI-CONTROL" [43] (multiple papers)

One lesson from the REP is that simplicity and consensus guarantee high adoption rates. Will other protocols than the REP gain traction or will they meet the same fate as the ACAP [45] protocol, introduced in 2007?

Links / Resources

Bibliography, source code and data:
`https://github.com/commoncrawl/robotstxt-experiments`

IIPC Web Archiving Conference 2025, 8 – 10 April 2025, Oslo, Norway

References

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

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

[3] 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.

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

[5] 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.

[6] 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/.

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

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

[9] Martijn Koster et al. *Robots Exclusion Protocol*. Tech. rep. 9309. Sept. 2022. 12 pp. 10.17487/RFC9309. https://www.rfc-editor.org/info/rfc9309.

[10] Gary Illyes. *Robots Exclusion Protocol Extension for URI Level Control*. Internet-Draft draft-illyes-repext-02. Work in Progress. Internet Engineering Task Force, Oct. 2024. 6 pp. https://datatracker.ietf.org/doc/draft-illyes-repext/02/.

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

[12] Gary Illyes. *Robots Exclusion Protocol User Agent Purpose Extension*. Internet-Draft draft-illyes-rep-purpose-00. Work in Progress. Internet Engineering Task Force, Oct. 2024. 4 pp. https://datatracker.ietf.org/doc/draft-illyes-rep-purpose/00/.

[13] Thom Vaughan. *Vocabulary for Expressing Content Preferences for AI Training*. Internet-Draft draft-vaughan-aipref-vocab-00. Work in Progress. Internet Engineering Task Force, Jan. 2025. 13 pp. https://datatracker.ietf.org/doc/draft-vaughan-aipref-vocab/00/.

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

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

[16] Chien-yi Chang and Xin He. *The Liabilities of Robots.txt*. 2025. arXiv: 2503.06035 [cs.CY]. https://arxiv.org/abs/2503.06035.

[17] Avv. Gino Fontana. "Web scraping: Jurisprudence and legal doctrines". In: *The Journal of World Intellectual Property* (2024), pp. 1–16. https://doi.org/10.1111/jwip.12331. https://onlinelibrary.wiley.com/doi/pdf/10.1111/jwip.12331. https://onlinelibrary.wiley.com/doi/abs/10.1111/jwip.12331.

[18] Geoffrey Xiao. "Bad Bots: Regulating the Scraping of Public Personal Information". In: *Harvard Journal of Law & Technology (Harvard JOLT)* 34 (2020), p. 701. https://heinonline.org/HOL/Page?handle=hein.journals/hjlt34&id=713&div=&collection=.

[19] Peter Henderson et al. *Foundation Models and Fair Use*. 2023. arXiv: 2303.15715 [cs.CY]. https://arxiv.org/abs/2303.15715.

[20] Jaekyeom Kim et al. "Do Not Trust Licenses You See–Dataset Compliance Requires Massive-Scale AI-Powered Lifecycle Tracing". In: *arXiv preprint arXiv:2503.02784* (2025). https://arxiv.org/abs/2503.02784.

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

[22] Greg Elmer. "The spam book: On viruses, porn and other anomalies from the dark side of digital culture". In: ed. by Jussi Parikka and Tony D. Sampson. Creskill, New Jersey: Hampton Press, 2009. Chap. Robots.txt: The politics of search engine exclusion, pp. 217–227.

[23] 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. 10.1109/WI.2007.98.

[24] 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.

[25] *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/.

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

[27] *About Neevabot*. https://web.archive.org/web/20210519031312/https://neeva.com/neevabot.

[28] Shayne Longpre et al. *Consent in Crisis: The Rapid Decline of the AI Data Commons*. 2024. arXiv: 2407.14933 [cs.CL]. https://arxiv.org/abs/2407.14933.

[29] Enze Liu et al. *Somesite I Used To Crawl: Awareness, Agency and Efficacy in Protecting Content Creators From AI Crawlers*. 2024. arXiv: 2411.15091 [cs.HC]. https://arxiv.org/abs/2411.15091.

[30] Melany Amarikwa. *Internet Openness at Risk: Generative AI's Impact on Data Scraping*. en. SSRN Scholarly Paper. Rochester, NY, Feb. 2024. 10.2139/ssrn.4723713. https://papers.ssrn.com/abstract=4723713 (visited on 03/20/2025).

[31] Shayne Longpre. *AI crawler wars threaten to make the web more closed for everyone*. en. 2025. https://www.technologyreview.com/2025/02/11/1111518/ai-crawler-wars-closed-web/.

[32] 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.

[33] Wei Li, Jian Liao, and Jianping Zeng. "Efficiency Analysis on Robots Exclusion Protocol Based on Game Theory". In: *2019 IEEE 13th International Conference on Anti-counterfeiting, Security, and Identification (ASID)*. IEEE. 2019, pp. 1–5. https://ieeexplore.ieee.org/abstract/document/8925189.

[34] Hanlin Chen, Hongmei He, and Andrew Starr. "An overview of web robots detection techniques". In: *2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)*. IEEE. 2020, pp. 1–6. https://ieeexplore.ieee.org/abstract/document/9138856.

[35] Sebastian Nagel. "The robots.txt standard – Implementations and Usage". In: *Open Search Symposium 2022, 10-12 October 2022, CERN, Geneva, Switzerland*. 2022. https://indico.cern.ch/event/1149330/contributions/5074600/.

[36] *Experiments and metrics about robots.txt captures*. 2022. https://github.com/sebastian-nagel/ossym2022-robotstxt-experiments.

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

[38] Victor Le Pochat et al. "Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation". In: *Proceedings of the 26th Annual Network and Distributed System Security Symposium*. NDSS 2019. Feb. 2019. 10.14722/ndss.2019.23386. https://tranco-list.eu/.

[39] *Index to WARC Files and URLs in Columnar Format*. 2018. https://commoncrawl.org/2018/03/index-to-warc-files-and-urls-in-columnar-format/.

[40] 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.

[41] W3C. *TDM Reservation Protocol (TDMRep)*. 2024. https://www.w3.org/community/reports/tdmrep/CG-FINAL-tdmrep-20240510/.

[42] Kar Balan et al. "DECORAIT - DECentralized Opt-in/out Registry for AI Training". In: *Proceedings of the 20th ACM SIGGRAPH European Conference on Visual Media Production*. CVMP '23. London, United Kingdom: Association for Computing Machinery, 2023. ISBN: 9798400704260. 10.1145/3626495.3626506. https://doi.org/10.1145/3626495.3626506.

[43] *IAB Workshop on AI-CONTROL (aicontrolws)*. https://datatracker.ietf.org/group/aicontrolws/about/.

[44] *AI Preferences (aipref)*. https://datatracker.ietf.org/group/aipref/about/.

[45] Wikipedia contributors. *Automated Content Access Protocol — Wikipedia, The Free Encyclopedia*. 2024. https://en.wikipedia.org/w/index.php?title=Automated_Content_Access_Protocol.