Barkhausen AI

DATA INSIGHT · BA-DI-2

Crawl-delay: a directive written more often than it is read

Crawl-delay is a non-standard robots.txt field that asks a crawler to wait between requests.

“Every leap begins as a crackle.”

DOCUMENT	BA-DI-2
KIND	Data insight
DATE	2026
SUBJECTS	Crawlers & access; Engines & documentation; Market behavior
LICENSE	CC-BY-4.0
AUTHOR	Barkhausen AI

ABSTRACT

Crawl-delay is a non-standard robots.txt field that asks a crawler to wait between requests. It is not part of RFC 9309, and the operators of several major crawlers treat it differently: Google's documentation lists the fields it supports and states that crawl-delay is not among them; Bing documents that its crawler honors it; and Yandex documents that it stopped taking the directive into account in 2018, directing operators to a crawl-rate control instead. Against that mixed and partly negative support, the 2026 crawler-access census finds crawl-delay written into 228 of 1,381 parsed robots files — 16.5%, spread across all four sectors measured. This note pairs the support documentation with the census count to make one observation: a directive's presence in robots.txt is a separate fact from whether any crawler acts on it, and the two need not track each other. That gap is a base-rate reminder for the newer fields now appearing in robots.txt, whose eventual honoring their presence today does not establish.

CONTENTS

`Crawl-delay` is a robots.txt line that asks a crawler to pause a stated number of seconds between requests. It predates the standardization of robots.txt and was never folded into it: RFC 9309, the 2022 specification, defines the `User-agent`, `Allow`, and `Disallow` fields and no throttling field. `Crawl-delay` is therefore a convention that each crawler operator chooses to honor or ignore, and the record shows they choose differently. This note sets the operators' own support statements next to how often the directive is actually written, and observes that the two are only loosely connected.

1 Who acts on it

The three operators with public documentation on the point do not agree. Google's robots.txt documentation enumerates the fields it supports — `user-agent`, `allow`, `disallow`, and `sitemap` — and states in the same place that "other fields such as `crawl-delay` aren't supported" [1]. A `Crawl-delay` line addressed to Googlebot is, by that documentation, read past. Bing takes the opposite position: its webmaster documentation describes the directive as one it acts on, telling operators to treat "the value listed after the colon as a relative amount of throttling down you want to apply to MSNBot from its default crawl rate" [2]. Yandex occupies a third position, and a temporal one: its documentation states that "from February 22, 2018, Yandex doesn't take into account the Crawl-delay directive," and points operators to a crawl-rate setting in its webmaster console instead [3]. So among three documented crawlers, one never supported the field, one supports it, and one supported it and then stopped — a directive can lose its audience without being removed from a single robots.txt file.

2 How often it is written

The 2026 crawler-access census scanned the 1,381 robots files it could parse for a `Crawl-delay` line. It found one in 228 of them: 16.5% of parsed files carry the directive. The share is not concentrated in one kind of site — the count is 54 files among universities, 61 among news domains, 69 among e-commerce domains, and 44 among government domains — so this is a broadly distributed authoring habit rather than a sector artifact. Recounting the parsed corpus directly reproduces the census figure; a single e-commerce file that uses bare carriage returns as line endings shifts the e-commerce count between 69 and 70 depending on newline handling, which leaves the headline at 16.5%, or 16.6% on the alternative count.

Read against the previous section, the number sits oddly. A directive that a major crawler documents as unsupported, and that another documents having abandoned in 2018, is nonetheless present in one parsed robots file out of six. The census does not observe whether any of these files' `Crawl-delay` values changed a crawler's behavior — it observes only that the line is there.

3 Presence and support are separate facts

The observation is deliberately modest, and it is not that writing `Crawl-delay` is a mistake: for a crawler that honors it, the line does what it says, and a cautious operator may reasonably keep it for the crawlers that still read it. The point is narrower. The presence of a directive in a robots file and the honoring of that directive by a crawler are two different facts, established by two different kinds of evidence — a line scan for the first, an operator's documentation or measured behavior for the second — and they do not have to move together. Directives accumulate in robots.txt: once written, a line tends to stay, whether or not the crawlers it addresses ever read it, and whether or not they still do.

That has a forward-looking use. Robots.txt is again acquiring new, non-standard fields — signals addressed to AI training and retrieval that no RFC yet covers, several of which the same census tracks in its signal inventory. `Crawl-delay` is a worked base rate for how such fields behave once introduced: a

period in which a directive is widely written is not evidence that it is widely honored, and adoption measured by presence can run well ahead of, or behind, adoption measured by effect. A study that wants to claim a new directive "is being adopted" has to say which of the two it counted. The distinction, and the crawler-access dataset behind these counts, are the durable part of this note.

4 Limitations

The census measures presence, not effect: it records that a `Crawl-delay` line exists in a file, not that any crawler slowed down or that the file's author expected it to. The support picture is drawn from three operators that publish on the point and is current as of the access date; a crawler's stated position can change, as Yandex's did, and many crawlers publish nothing at all, so "supported" and "unsupported" are documented facts about specific operators rather than a complete accounting. The 16.5% figure is over parsed files only — files that could not be fetched or parsed are excluded — and a line scan counts a directive's presence without validating its syntax or value, so a malformed `Crawl-delay` line counts the same as a well-formed one.

REFERENCES

1. Google, Search Central / crawling documentation. *How Google interprets the robots.txt specification* (2026). https://developers.google.com/crawling/docs/robots-txt/robots-txt-spec Accessed 2026-07-10. [archived]
2. Microsoft Bing, Bing Webmaster Blog. *Crawl delay and the Bing crawler, MSNBot* (2009). https://blogs.bing.com/webmaster/August-2009/Crawl-delay-and-the-Bing-crawler,-MSNBot Accessed 2026-07-10. [archived]
3. Yandex, Webmaster documentation. *The Crawl-delay directive*. https://yandex.com/support/webmaster/en/robot-workings/crawl-delay.html Accessed 2026-07-10. [archived]

HOW TO CITE

Barkhausen AI (2026). Crawl-delay: a directive written more often than it is read. https://barkhausen.ai/notes/crawl-delay-and-crawler-support/