Engineer

robots.txt

Datei, die Crawlern sagt, welche Pfade sie meiden sollen

Was es ist

Jeder Host im Web kann unter dem festen Pfad `/robots.txt` eine einfache Textdatei mit Bitten an die Programme veröffentlichen, die ihn durchsuchen. Martijn Koster legte die Konvention 1994 fest. Im September 2022 veröffentlichte die IETF sie als RFC 9309, das Robots Exclusion Protocol, mit genauer Sprache für die Regeln und Vorgaben zum Umgang mit Fehlern und Caching.

Die Datei ist eine Liste von Gruppen. Eine Gruppe beginnt mit einer oder mehreren `user-agent` -Zeilen. Jede benennt einen Crawler anhand seiner Produktkennung, etwa `ExampleBot` , oder verwendet `*` für alle Crawler. Darunter folgen `allow` - und `disallow` -Zeilen mit jeweils einem Pfad. Ein Crawler sucht die Gruppe mit seiner eigenen Kennung ohne Beachtung der Groß- und Kleinschreibung und folgt ihr. Benennen ihn mehrere Gruppen, werden deren Regeln zusammengeführt. Nur ein Crawler ohne eigene Gruppe greift auf die Gruppe `*` zurück. Findet er keine von beiden, hat er keine Regeln zu befolgen.

Innerhalb einer Gruppe gilt eine Regel, wenn ihr Pfad auf den Anfang der Adresse passt. Die genaueste Übereinstimmung gewinnt, also die mit den meisten Oktetten. Bei Gleichstand zwischen `allow` und `disallow` gewinnt `allow` . Eine Adresse, zu der keine Regel passt, ist erlaubt. Zwei Zeichen erweitern einen Pfad: `*` steht für eine beliebige Zeichenfolge, `$` verankert das Ende der Adresse. Ein `#` beginnt einen Kommentar, und der Pfad der Datei selbst ist immer erlaubt.

Die RFC legt auch den Abruf fest. Die Datei liegt unter diesem kleingeschriebenen Pfad an der Wurzel des Dienstes. Deshalb antwortet jeder Host für sich, und eine Datei auf einer Subdomain sagt nichts über eine andere

aus. Sie ist einfacher UTF-8-Text. Ein Crawler sollte beim Abruf mindestens fünf aufeinanderfolgenden Weiterleitungen folgen können und eine Cache-Kopie nicht länger als 24 Stunden verwenden. Die ausgewertete Datenmenge darf er begrenzen, aber niemals auf weniger als 500 KiB.

Zeilen außerhalb des Protokolls werden geduldet. Die RFC nennt sie weitere Datensätze, führt `Sitemap` als Beispiel an und überlässt es dem Crawler, ob er sie auswertet. Es gilt eine Bedingung: Eine solche Zeile darf niemals eine Gruppe beenden.

Das Konzept für eine Entscheidung nutzen

Nutzen Sie „Was ist robots.txt?“ für eine konkrete Entscheidung. Wir können helfen, das Konzept mit dem System in Verbindung zu bringen, das Sie entwickeln oder einkaufen. Eine Website oder App, die die KI-Assistenten Ihrer Kunden finden, lesen und bedienen können — bei jeder Veröffentlichung geprüft statt einmal versprochen.

Unsere Leistungen entdecken Kontakt aufnehmen →

Die Annahmen hinter dem Begriff prüfen

Nutzen Sie die Erklärung und die verlinkten Arbeiten, um Fragen für Ihr Projekt zu formulieren. Wir können die Annahmen gemeinsam mit Ihnen prüfen, bevor ein Werkzeug ausgewählt wird.
Hugo-Website mehrsprachig in 3 Sprachen aufgebaut — 108 Vorlagendateien für drei Sprachen und vier Formate

Hugo-Website mehrsprachig in 3 Sprachen aufgebaut Kontakt aufnehmen →

Warum es wichtig ist

Die RFC benennt die Grenze in der Einleitung: „Diese Regeln sind keine Form der Zugriffsautorisierung.“ Die Datei ist eine Bitte. Ein Crawler, der sie beachtet, bleibt draußen. Einer, der sie nicht beachtet, liest alles, was der Server herausgibt. Der Sicherheitsteil ergänzt, dass ein aufgeführter Pfad öffentlich und auffindbar wird. Alles, was privat bleiben muss, braucht deshalb eine echte Zugriffskontrolle, etwa HTTP-Authentifizierung.

Googles Dokumentation grenzt den Zweck weiter ein. Die Datei werde „hauptsächlich verwendet, um eine Überlastung Ihrer Website durch Anfragen zu vermeiden“, und sei „kein Mechanismus, um eine Webseite aus Google fernzuhalten“. Eine gesperrte Adresse kann weiterhin indexiert werden, wenn andere Seiten darauf verlinken. Sie erscheint dann ohne Beschreibung in den Ergebnissen, weil der Crawler die Seite nie gelesen hat. Um eine Seite aus den Ergebnissen fernzuhalten, verweist Google auf eine `noindex` -Regel oder ein Passwort. Eine `noindex` -Regel funktioniert nur auf einer Seite, die der Crawler abrufen darf.

Die Datei erledigt gut, wofür sie geschrieben wurde: Crawler von Teilen einer Website fernhalten, die den Server mit Anfragen belasten und einer Suchmaschine wenig bringen, etwa unwichtige oder fast identische Seiten. Ihre Reichweite ist zugleich ihr Risiko. Eine Zeile unter `*` gilt für jede Adresse und jeden Crawler, der die Datei beachtet. Eine Änderung kann bis zu einem Tag brauchen, um gesehen zu werden. Deshalb schließt ein `Disallow: /` in der falschen Gruppe eine ganze Website.

Wie es in der Praxis aussieht

Die Datei dieser Website ist lang und besteht fast nur aus Kommentaren. Der Teil, nach dem ein Crawler handelt, umfasst vier Zeilen:

```
User-agent: *
Content-Signal: search=yes, ai-input=yes, ai-train=yes
Allow: /

Sitemap: https://engineer.company/sitemap.xml
```

Eine Gruppe deckt alle Crawler ab, und `Allow: /` passt auf jede Adresse. Eine Gruppe ohne Regeln würde genau dasselbe erlauben, weil eine Adresse ohne passende Regel erlaubt ist. Die Zeile steht dennoch ausdrücklich da, weil die Offenheit eine Entscheidung ist. Der Kommentar darüber nennt die Richtlinie „Decided rather than inherited“, also bewusst entschieden statt geerbt. Die Seite darüber, was KI-Agenten von einer Website benötigen, beschreibt die Crawl-Richtlinie als eine der bewussten Entscheidungen dieser Website.

`Content-Signal` ist kein in RFC 9309 definierter Datensatz. Er stammt aus dem Content-Signals-Vokabular und stimmt hier der Erfassung für Suche, für Antworten zum Inferenzzeitpunkt und für Training zu. Nach der RFC ist es einer jener weiteren Datensätze. Ein Crawler, der die Konvention kennt, darf ihn lesen. Ein älterer überspringt die Zeile, wie der eigene Kommentar der Datei sagt. Keiner darf sie als Ende der Gruppe behandeln. Die Zeile `Sitemap:` ist ein weiterer solcher Datensatz und verweist auf den Sitemap-Index der Website. Die Credits-Seite fasst die Datei ebenso zusammen: Sie „erlaubt alles, Such-Crawlern wie KI-Systemen“, und nennt die Sitemap.

Auch was die Datei auslässt, ist bewusst entschieden. Unter den fünf Empfehlungen, die diese Website ablehnte, nennt die Seite zur KI-Bereitschaft Regeln pro Agent. Sie behält eine gemeinsame Robots-Regel für alle Agenten,

„weil eine für einen Crawler geschriebene Regel für ihn unbemerkt die allgemeine ersetzt". Hätte die Datei einem KI-Crawler eine eigene Gruppe gegeben, würde dieser nur jene Gruppe lesen. Ein später ergänztes `Disallow` unter `*` würde ihn nie erreichen.

Der Rest sind Kommentare für Menschen, die die Datei öffnen: die Gründe für die Richtlinie und dann die Adressen von `llms.txt`, `agents.txt`, `humans.txt`, `security.txt` und weiteren maschinenlesbaren Dateien. Sie bleiben Kommentare, „because robots.txt has no field for them", also weil robots.txt dafür kein Feld hat. Damit gehören die Kommentare zum ausgelieferten Inhalt. Eine `#`-Zeile aus der Vorlage geht Byte für Byte an jeden Crawler. Zweimal enthielt eine solche Zeile eine Notiz für die Wartung, die auf ein Planungsdokument in einem nicht öffentlichen Repository zeigte. Eine Prüfung der gebauten Website schlägt inzwischen fehl, wenn eine ausgelieferte Datei einen Pfad ins Repository enthält. Wartungsnotizen stehen in der Vorlage in Vorlagenkommentaren, die der Generator entfernt.

Wo es schiefgeht

Eine allgemeine Regel gilt für einen Crawler nicht mehr. Jemand ergänzt eine Gruppe für genau einen Bot, vielleicht um ihn von einem Verzeichnis fernzuhalten. Von da an ignoriert der Bot alles unter `*`, einschließlich jedes dort eingetragenen `Disallow`. RFC 9309 schickt einen Crawler nur dann zur Gruppe `*`, wenn ihn keine eigene Gruppe nennt. Googles Dokumentation sagt, dass spezifische Gruppen und die Gruppe `*` „nicht kombiniert werden". Wiederholen Sie jede allgemeine Regel in jeder benannten Gruppe, oder bleiben Sie bei einer einzigen Gruppe wie diese Website.

Eine gesperrte Seite steht trotzdem in den Suchergebnissen. Die Adresse erscheint ohne Beschreibung, vielleicht mit den Wörtern, die andere Websites in ihren Links verwendet haben. Google hat die Seite nie abgerufen und kennt daher nur ihre Existenz und die Bezeichnungen anderer. Ein `Disallow` steuert das Crawling, nicht die Indexierung. Entfernen Sie die Regel und liefern Sie `noindex` auf der Seite aus, oder schützen Sie die Seite mit einem Passwort. Ein `noindex`, das hinter einem `Disallow` verborgen ist, wird nie gelesen.

Die Datei verrät, was sie verbergen sollte. Eine Zeile wie `Disallow: /admin/` sagt jedem Leser, ob höflicher Crawler oder nicht, genau, wo er suchen muss. Die Datei ist absichtlich öffentlich. Der Sicherheitsteil der RFC sagt, dass aufgeführte

Pfade auffindbar werden, und fordert stattdessen Zugriffskontrolle. Schützen Sie private Inhalte durch Authentifizierung und lassen Sie sie ganz aus der Datei heraus.

Ein Serverfehler stoppt das Crawling der ganzen Website. Die Seiten laden, aber der Abruf von `/robots.txt` liefert einen 5xx-Status oder läuft in einen Timeout. Ein RFC-konformer Crawler muss das als vollständige Sperre behandeln. Google stoppt das Crawling für die ersten 12 Stunden und arbeitet dann bis zu 30 Tage mit der letzten gültigen Kopie. Der umgekehrte Fehler ist leiser: Eine andere 4xx-Antwort als 429, einschließlich 401 und 403, behandelt Google so, als gäbe es keine Datei. Damit ist jede Adresse offen. Liefern Sie die Datei als statischen Text mit Status 200 aus und überwachen Sie `/robots.txt` selbst, nicht nur die Startseite.

Weiterführende Lektüre

- RFC 9309: Robots Exclusion Protocol — Der Standard selbst: Gruppen, Abgleich, Fehlerbehandlung, Caching und der Sicherheitshinweis.
- Einführung in robots.txt — Wozu Google die Datei vorsieht und warum sie keine Seite aus den Suchergebnissen verbirgt.
- Wie Google die robots.txt-Spezifikation auslegt — Statuscodes, Caching, ignorierte Felder und Gruppenauswahl.
- Suchindexierung mit noindex verhindern — Die Regel, die eine Seite tatsächlich aus den Ergebnissen hält, und warum diese Seite dafür crawlbar sein muss.

Glossar

Für Ihre tatsächlichen Rahmenbedingungen auswählen

Beschreiben Sie den Arbeitsablauf, die vorhandenen Schnittstellen und die Nutzer, die die Veränderung benötigen. Wir können eine prüfbare Lieferung definieren und die Übergabe so gestalten, dass Ihr Team die gelieferten Bestandteile warten kann.

Preise ansehen Kontakt aufnehmen →

Die Begriffe praktisch nutzbar machen

Fragen Sie, wie „Was ist robots.txt?“ auf Ihre Organisation zutrifft. Ein konkretes Beispiel aus Ihrem bestehenden System reicht aus, um das Gespräch zu beginnen.

Unsere Leistungen entdecken

welcome@engineer.company +45 50 10 61 84 Kopenhagen, Dänemark

Nehmen Sie diese Seite mit

Kontaktkarte

Mit Sorgfalt und Liebe gestaltet von D.

CVR 45969703 · Tagensvej 105C, 3. th, 2200 København N

Copyright © 2025 – heute · Engineer ApS · Alle Rechte vorbehalten.

Gedruckt von https://engineer.company/de/glossary/robots-txt/