Engineer

robots.txt

filen som forteller søkeroboter hvor de ikke skal gå

Hva det er

Hver vert på weben kan publisere en ren tekstfil på den faste stien `/robots.txt`, med anmodninger rettet til programmene som gjennomsøker den. Martijn Koster definerte konvensjonen i 1994, og i september 2022 publiserte IETF den som RFC 9309, Robots Exclusion Protocol, med presist språk for reglene og instruksjoner for feil og caching.

Filen er en liste med grupper. En gruppe åpner med én eller flere `user-agent`-linjer, som hver navngir en søkerobot ved produkttokenet dens (product token), f.eks. `ExampleBot`, eller `*` for alle søkeroboter. Under dem kommer `allow`- og `disallow`-linjer, hver med en sti. En søkerobot ser etter gruppen som navngir dens eget token, uten hensyn til store og små bokstaver, og følger den gruppen; når flere grupper navngir den, slås reglene deres sammen til én. Bare en søkerobot uten egen gruppe faller tilbake på `*`-gruppen, og en som ikke finner noen av dem, har ingen regler å følge.

Innenfor en gruppe gjelder en regel når stien matcher begynnelsen av adressen, og det mest spesifikke treffet vinner, altså det med flest oktetter. Står `allow` og `disallow` likt, vinner `allow`, og en adresse som ikke treffes av noen regel, er tillatt. To tegn gjør en sti bredere: `*` står for en hvilken som helst rekke tegn, og `$` låser slutten av adressen. En `#` innleder en kommentar, og filens egen sti er alltid tillatt.

RFC-en fastsetter også hvordan filen hentes. Den ligger på akkurat den stien, med små bokstaver, øverst i tjenesten, så hver vert svarer for seg selv, og en fil på ett underdomene sier ingenting om et annet. Den er ren tekst i UTF-8. En søkerobot bør følge minst fem omdirigeringer for å nå den, bør ikke bruke en lagret kopi i mer enn 24 timer, og kan begrense hvor mye den parser, men aldri til under 500 KiB.

Linjer utenfor protokollen tåles. RFC-en kaller dem andre poster (other records), gir `Sitemap` som eksempel og lar en søkerobot tolke dem eller la være, på én betingelse: en slik linje må aldri avslutte en gruppe.

Bruk begrepet i en beslutning

Bruk «Hva er robots.txt?» i en konkret beslutning. Vi kan knytte begrepet til systemet dere bygger eller kjøper. Et nettsted eller en app som AI-assistentene kundene dine bruker, kan finne, lese og handle ut fra, kontrollert ved hver utgivelse i stedet for lovet én gang.

Se tjenestene våre Ta kontakt →

Undersøk antakelsene bak ordet

Bruk forklaringen og det tilknyttede arbeidet til å formulere spørsmål for prosjektet. Vi kan undersøke antakelsene sammen før et verktøy velges. Bygde et flerspråklig Hugo-nettsted på 3 språk — 108 malfiler som betjener tre språk og fire formater

Bygde et flerspråklig Hugo-nettsted på 3 språk Ta kontakt →

Hvorfor det er viktig

RFC-en trekker grensen allerede i innledningen: «These rules are not a form of access authorization.» Reglene gir eller nekter altså ikke tilgang. Filen er en anmodning. En søkerobot som respekterer den, holder seg unna, og en som ikke gjør det, leser alt serveren gir den. Sikkerhetsavsnittet legger til at en sti som står i filen, blir offentlig og lett å finne, så alt som må forbli privat, trenger ekte tilgangskontroll, f.eks. autentisering over HTTP.

Dokumentasjonen til Google snevrer formålet inn ytterligere. Den beskriver filen som et middel først og fremst for å unngå at nettstedet overbelastes av forespørsler, og sier rett ut at den ikke er en mekanisme for å holde en nettside utenfor Google. En adresse under `disallow` kan likevel bli indeksert når andre sider lenker til den, og vises da i resultatene uten beskrivelse, fordi søkeroboten aldri leste siden. For å holde en side utenfor resultatene viser Google til en `noindex` -regel eller et passord, og en `noindex` -regel virker bare på en side søkeroboten har lov til å hente.

Det filen gjør godt, er jobben den ble skrevet for: å styre søkeroboter bort fra deler av et nettsted som ville koste serveren forespørsler og gi en søkemotor lite tilbake, f.eks. uviktige eller nesten like sider. Rekkevidden er også risikoen. Én linje under `*` gjelder hver adresse og hver søkerobot som respekterer filen, og en endring kan ta opptil et døgn før den blir sett, så `Disallow: /` skrevet i feil gruppe stenger et helt nettsted.

Slik ser det ut i praksis

Filen til dette nettstedet er lang, og nesten alt i den er kommentarer. Delen en søkerobot handler etter, er fire linjer:

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

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

Én gruppe dekker alle søkeroboter, og `Allow: /` matcher hver adresse. En gruppe uten regler ville tillatt nøyaktig det samme, siden en adresse som ikke treffes av noen regel, er tillatt, men linjen er skrevet ut fordi åpenheten er et valg. Kommentaren over den kaller policyen «Decided rather than inherited», bestemt og ikke arvet, og siden om å være klar for AI-agenter beskriver policyen for gjennomsøking som en av nettstedets bevisste beslutninger.

`Content-Signal` er ikke en post som RFC 9309 definerer. Den kommer fra vokabularet Content Signals, og her samtykker den til innsamling for søk, for å svare på et spørsmål i det øyeblikket det stilles (inference time), og for trening. Etter RFC-en er den en av de andre postene: en søkerobot som kjenner konvensjonen, kan lese den, en som er eldre enn den, hopper over linjen, slik filens egen kommentar sier, og ingen av dem kan behandle den som slutten på gruppen. Linjen `Sitemap:` er en annen post av samme slag og peker på sitemap-indeksen til nettstedet. Kolofonen oppsummerer filen på samme måte: den tillater alt, «for søkeroboter og KI-systemer likt», og navngir sitemapet.

Det filen utelater, ble også bestemt. Blant de fem anbefalingene dette nettstedet takket nei til, nevner AI-siden regler per agent. Nettstedet beholder én robots-regel for alle agenter, fordi en regel skrevet for én søkerobot i det stille erstatter den generelle regelen for nettopp den. Hadde filen gitt én søkerobot for AI en egen gruppe, ville den søkeroboten lest bare den gruppen, og en `Disallow` lagt til senere under `*` ville aldri nådd den.

Resten er kommentarer, skrevet for den som åpner filen: begrunnelsen for policyen, deretter adressene til `llms.txt`, `agents.txt`, `humans.txt`, `security.txt` og andre maskinlesbare filer, holdt som kommentarer fordi robots.txt ikke har noe felt for dem, som filen selv sier. Det gjør kommentarene til en del av det som leveres. En `#`-linje i malen serveres byte for byte til hver

søkerobot, og to ganger har en slik linje båret en merknad ment for en vedlikeholder, med en henvisning til et planleggingsdokument i et repo som ikke er offentlig. En sjekk over det bygde nettstedet feiler nå når en servert fil inneholder en sti inn i repoet, og merknader for vedlikeholdere står i malkommentarer, som generatoren fjerner.

Hvor det går galt

En regel ment for alle slutter å gjelde én søkerobot. Noen legger til en gruppe for én enkelt robot, kanskje for å holde den ute av én mappe, og fra da av overser den roboten alt under `*`, også hver `Disallow` som står der. RFC 9309 sender en søkerobot til `*`-gruppen bare når ingen gruppe navngir den, og dokumentasjonen til Google sier at spesifikke grupper og `*`-gruppen ikke kombineres. Gjenta hver generelle regel inne i hver navngitte gruppe, eller hold deg til én gruppe, slik dette nettstedet gjør.

En blokkert side dukker likevel opp i søkeresultatene. Adressen vises uten beskrivelse, kanskje med ordene andre nettsteder brukte da de lenket til den. Google hentet aldri siden, så den vet bare at adressen finnes og hva andre kaller den. En `Disallow` styrer gjennomsøking, ikke indeksering. Fjern regelen og send `noindex` på siden, eller legg siden bak et passord; en `noindex` som skjules bak en `Disallow`, blir aldri lest.

Filen annonserer det den skulle skjule. En linje som `Disallow: /admin/` forteller hver leser, høflig søkerobot eller ikke, nøyaktig hvor den skal lete, og filen er offentlig med vilje. Sikkerhetsavsnittet i RFC-en sier at stier som listes, blir lette å finne, og ber om tilgangskontroll i stedet. Legg privat materiale bak autentisering, og hold det helt ute av filen.

En serverfeil stopper gjennomsøkingen av et helt nettsted. Sidene lastes, men forespørselen etter `/robots.txt` returnerer en 5xx-status eller får tidsavbrudd, og en søkerobot som følger RFC-en, må behandle det som en fullstendig `disallow`. Google stopper gjennomsøkingen de første 12 timene og arbeider deretter ut fra den siste gode kopien sin i opptil 30 dager. Den motsatte feilen er stillere: Google behandler et 4xx-svar annet enn 429, også 401 og 403, som om filen ikke fantes, så hver adresse er åpen. Lever filen som statisk tekst med status 200, og overvåk `/robots.txt` selv, ikke bare forsiden.

Les mer

- RFC 9309: Robots Exclusion Protocol — selve standarden: grupper, matching, feilhåndtering, caching og sikkerhetsmerknaden.
- Introduction to robots.txt — hva Google sier filen er til, og hvorfor den ikke skjuler en side fra resultatene.
- How Google interprets the robots.txt specification — statuskoder, caching, feltene Google overser, og hvordan den velger en gruppe.
- Block Search indexing with noindex — regelen som faktisk holder en side ute av resultatene, og hvorfor den krever at siden kan gjennomsøkes.

Ordliste

Velg ut fra deres faktiske rammer

Beskriv arbeidsflyten, eksisterende grensesnitt og brukerne som trenger endringen. Vi kan definere en testbar leveranse og en overlevering som gjør teamet i stand til å vedlikeholde løsningen.

Se hva det koster Ta kontakt →

Gjør begrepene praktiske

Spør hvordan «Hva er robots.txt?» gjelder i organisasjonen deres. Et konkret eksempel fra dagens system er nok til å starte samtalen.

Se tjenestene våre

welcome@engineer.company +45 50 10 61 84 København, Danmark

Ta med deg denne siden

Visittkort

Laget med kjærlighet av D.

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

Opphavsrett © 2025 – nå · Engineer ApS · Alle rettigheter forbeholdt.

Skrevet ut fra https://engineer.company/nb/glossary/robots-txt/