Engineer

robots.txt

filen som talar om för sökrobotar vart de inte ska gå

Vad det är

Varje värd på webben kan på den fasta sökvägen `/robots.txt` publicera en ren textfil med önskemål riktade till de program som genomsöker den. Martijn Koster definierade konventionen 1994, och i september 2022 publicerade IETF den som RFC 9309, Robots Exclusion Protocol, med ett precist språk för reglerna och instruktioner för fel och cachning.

Filen är en lista med grupper. En grupp inleds med en eller flera rader `user-agent`, som var och en namnger en sökrobot med dess produkttoken, t.ex. `ExampleBot`, eller `*` för alla sökrobotar. Under dem kommer rader med `allow` och `disallow`, som var och en innehåller en sökväg. En sökrobot letar efter den grupp som namnger dess egen token, matchad utan hänsyn till versaler och gemener, och följer den gruppen. När flera grupper namnger den slås deras regler ihop till en. Bara en sökrobot som saknar en egen grupp faller tillbaka på gruppen `*`, och en som hittar ingendera har inga regler att följa.

Inom en grupp gäller en regel när dess sökväg matchar början av adressen, och den mest specifika träffen vinner, alltså den med flest oktetter. Blir det oavgjort mellan `allow` och `disallow` vinner `allow`, och en adress som ingen regel matchar är tillåten. Två tecken vidgar en sökväg: `*` står för en godtycklig följd av tecken och `$` låser adressens slut. Ett `#` inleder en kommentar, och filens egen sökväg är alltid tillåten.

RFC:n fastställer också hur filen hämtas. Den ligger på just den sökvägen, med gemener, högst upp i tjänsten, så varje värd svarar för sig själv och en fil på en underdomän säger ingenting om en annan. Den är ren text i UTF-8. En sökrobot bör följa minst fem omdirigeringar för att nå den, bör inte använda en cachad kopia i mer än 24 timmar och får begränsa hur mycket den tolkar, dock aldrig under 500 KiB.

Rader utanför protokollet tolereras. RFC:n kallar dem andra poster, other records, ger `Sitemap` som exempel och låter en sökrobot tolka dem eller låta bli, på ett villkor: en sådan rad får aldrig avsluta en grupp.

Använd begreppet i ett beslut

Använd ”Vad är robots.txt?” i ett konkret beslut. Vi kan knyta begreppet till systemet ni bygger eller köper. En webbplats eller app som de AI-assistenter dina kunder använder kan hitta, läsa och agera på, kontrollerad vid varje release i stället för utlovad en gång.

Se våra tjänster Hör av er →

Granska antagandena bakom ordet

Använd förklaringen och det länkade arbetet för att formulera frågor till projektet. Vi kan undersöka antagandena tillsammans innan ett verktyg väljs. Byggde en flerspråkig Hugo-webbplats på 3 språk — 108 mallfiler som täcker tre språk och fyra format

Byggde en flerspråkig Hugo-webbplats på 3 språk Hör av er →

Varför det spelar roll

RFC:n anger gränsen redan i inledningen: ”These rules are not a form of access authorization.” Reglerna är alltså inte en form av behörighet. Filen är en begäran. En sökrobot som respekterar den håller sig borta, och en som inte gör det läser vad servern än lämnar ut. Säkerhetsavsnittet tillägger att en sökväg som listas i filen blir offentlig och upptäckbar, så allt som måste förbli privat behöver en verklig åtkomstkontroll, t.ex. autentisering över HTTP.

Googles dokumentation snävar in syftet ytterligare. Den beskriver filen som ”used mainly to avoid overloading your site with requests”, främst till för att inte överbelasta webbplatsen med anrop, och säger att ”it is not a mechanism for keeping a web page out of Google”. En adress som inte får genomsökas kan ändå indexeras när andra sidor länkar till den, och visas då i sökresultaten utan beskrivning, eftersom sökroboten aldrig läste sidan. För att hålla en sida borta ur sökresultaten pekar Google på en `noindex` -regel eller ett lösenord, och en `noindex` -regel fungerar bara på en sida som sökroboten får hämta.

Det filen gör bra är det jobb den skrevs för: att styra sökrobotar bort från delar av en webbplats som skulle kosta servern anrop och ge en sökmotor lite tillbaka, som oviktiga eller nästan identiska sidor. Dess räckvidd är också dess risk. En enda rad under `*` gäller för varje adress och varje sökrobot som respekterar filen, och en ändring kan ta upp till ett dygn att märkas, så ett `Disallow: /` som skrivs i fel grupp stänger en hel webbplats.

Så ser det ut i praktiken

Den här webbplatsens fil är lång, och nästan allt i den är kommentarer. Den del som en sökrobot agerar på är fyra rader:

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

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

En grupp täcker alla sökrobotar, och `Allow: /` matchar varje adress. En grupp utan regler skulle tillåta exakt detsamma, eftersom en adress som ingen regel matchar är tillåten, men raden skrivs ut eftersom öppenheten är ett val. Kommentaren ovanför kallar policyn ”Decided rather than inherited”, beslutad snarare än ärvd, och sidan om att vara redo för AI-agenter beskriver webbplatsens policy för genomsökning som ett av dess medvetna beslut.

`Content-Signal` är inte en post som RFC 9309 definierar. Den kommer från vokabulären Content Signals, och här ger den samtycke till insamling för sökning, för att besvara en frågeställning vid inferens och för träning. Enligt RFC:n är den en av de andra posterna: en sökrobot som känner till konventionen får läsa den, en som är äldre än konventionen hoppar över raden, som filens egen kommentar säger, och ingen av dem får behandla den som gruppens slut. Raden `Sitemap:` är en annan post av samma slag och pekar på webbplatsens index över webbplatskartor. Kolofonen sammanfattar filen på samma sätt: den ”tillåter allt, för både sökrobotar och AI-system” och anger webbplatskartan.

Det som filen utelämnar är också ett beslut. Bland de fem rekommendationer som den här webbplatsen valde bort listar sidan om AI-beredskap regler per agent, och den behåller en robots-regel för alla agenter, ”eftersom en regel skriven för en sökrobot i tysthet ersätter den allmänna för just den sökroboten”. Om filen hade gett en AI-sökrobot en egen grupp skulle den sökroboten bara läsa den gruppen, och ett `Disallow` som lades till senare under `*` skulle aldrig nå den.

Resten är kommentarer, skrivna för den som öppnar filen: resonemanget bakom policyn, sedan adresserna till `llms.txt`, `agents.txt`, `humans.txt`, `security.txt` och andra maskinläsbara filer, som står som kommentarer ”because robots.txt has no field for them”, eftersom robots.txt saknar fält för dem. Det gör

kommentarerna till en del av det som levereras. En `#`-rad i mallen serveras byte för byte till varje sökrobot, och två gånger bar en sådan rad en anteckning avsedd för en underhållare, med en hänvisning till ett planeringsdokument i ett repo som inte är offentligt. En kontroll över den byggda webbplatsen fallerar nu när en serverad fil innehåller en sökväg in i repot, och anteckningar till underhållare i mallen ligger i mallkommentarer, som webbplatsgeneratorn tar bort.

Var det går fel

En regel avsedd för alla slutar gälla för en sökrobot. Någon lägger till en grupp för en enda bot, kanske för att hålla den borta från en mapp, och från och med då ignorerar den boten allt under `*`, inklusive varje `Disallow` som står där. RFC 9309 skickar en sökrobot till gruppen `*` bara när ingen grupp namnger den, och Googles dokumentation säger att specifika grupper och gruppen `*` ”are not combined”, inte slås ihop. Upprepa varje allmän regel i varje namngiven grupp, eller behåll en enda grupp, som den här webbplatsen gör.

En blockerad sida syns ändå i sökresultaten. Adressen dyker upp utan beskrivning, kanske med de ord som andra webbplatser använde när de länkade till den. Google hämtade aldrig sidan, så Google vet bara att adressen finns och vad andra kallar den. Ett `Disallow` styr genomsökningen, inte indexeringen. Ta bort regeln och servera `noindex` på sidan, eller lägg sidan bakom ett lösenord. Ett `noindex` som döljs bakom ett `Disallow` blir aldrig läst.

Filen gör reklam för det den skulle dölja. En rad som `Disallow: /admin/` talar om för varje läsare, artig sökrobot eller inte, exakt var den ska leta, och filen är offentlig per konstruktion. RFC:ns säkerhetsavsnitt säger att listade sökvägar blir upptäckbara och hänvisar till åtkomstkontroll i stället. Lägg privat material bakom autentisering och låt det helt utebli ur filen.

Ett serverfel stoppar genomsökningen av en hel webbplats. Sidorna laddas, men begäran om `/robots.txt` returnerar en 5xx-status eller får en timeout, och en sökrobot som följer RFC:n måste behandla det som ett fullständigt förbud. Google slutar genomsöka under de första 12 timmarna och arbetar sedan utifrån sin senaste fungerande kopia i upp till 30 dagar. Det motsatta misstaget är tystare: Google behandlar ett 4xx-svar utom 429, inklusive 401 och 403, som om det inte fanns någon fil, så varje adress är öppen. Servera filen som statisk text med status 200, och bevaka själva `/robots.txt`, inte bara startsidan.

Läs mer

- RFC 9309: Robots Exclusion Protocol — själva standarden: grupper, matchning, felhantering, cachning och säkerhetsavsnittet.
- Introduction to robots.txt — vad Google säger att filen är till för och varför den inte döljer en sida i sökresultaten.
- How Google interprets the robots.txt specification — statuskoder, cachning, de fält som Google ignorerar och hur Google väljer en grupp.
- Block Search indexing with noindex — regeln som verkligen håller en sida borta ur sökresultaten och varför sidan måste gå att genomsöka.

Ordlista

Välj utifrån era faktiska förutsättningar

Beskriv arbetsflödet, befintliga gränssnitt och användarna som behöver förändringen. Vi kan definiera en testbar leverans och en överlämning som gör att teamet kan underhålla lösningen.

Se vad det kostar Hör av er →

Gör begreppen praktiska

Fråga hur ”Vad är robots.txt?” gäller i er organisation. Ett konkret exempel från ert nuvarande system räcker för att börja samtalet.

Se våra tjänster

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

Ta med dig den här sidan

Visitkort

Omsorgsfullt skapad med kärlek av D.

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

Upphovsrätt © 2025 – nu · Engineer ApS · Alla rättigheter förbehållna.

Utskrivet från https://engineer.company/sv/glossary/robots-txt/