Engineer

robots.txt

archivo que indica a los rastreadores dónde no entrar

Qué es

Cada host de la web puede publicar, en la ruta fija `/robots.txt`, un archivo de texto sin formato con peticiones dirigidas a los programas que lo rastrean. Martijn Koster definió la convención en 1994 y, en septiembre de 2022, el IETF la publicó como RFC 9309, el Protocolo de Exclusión de Robots, añadiendo un lenguaje preciso para las reglas y las instrucciones sobre errores y caché.

El archivo es una lista de grupos. Un grupo empieza con una o varias líneas `user-agent`, cada una de las cuales identifica un rastreador por su token de producto, como `ExampleBot`, o con `*` para todos los rastreadores. Debajo aparecen líneas `allow` y `disallow`, cada una con una ruta. El rastreador busca el grupo que nombra su propio token, sin distinguir mayúsculas y minúsculas, y obedece ese grupo; si varios grupos lo nombran, sus reglas se combinan en un único grupo. Solo un rastreador sin grupo propio recurre al grupo `*`, y el que no encuentra ninguno de los dos no tiene reglas que seguir.

Dentro de un grupo, una regla se aplica si su ruta coincide con el principio de la dirección, y gana la coincidencia más específica: la que tiene más octetos. Un empate entre `allow` y `disallow` se resuelve a favor de `allow`, y una dirección que no coincida con ninguna regla está permitida. Dos caracteres amplían una ruta: `*` representa cualquier secuencia de caracteres y `$` fija el final de la dirección. Un `#` inicia un comentario, y la ruta del propio archivo siempre está permitida.

La RFC también establece cómo se obtiene el archivo. Está en esa ruta en minúsculas, en la raíz del servicio, de modo que cada host responde por sí mismo y un archivo de un subdominio no dice nada de otro. Es texto sin formato en UTF-

8. Un rastreador debería seguir al menos cinco redirecciones para encontrarlo, no debería utilizar una copia en caché durante más de 24 horas y puede limitar cuánto analiza, pero nunca por debajo de 500 KiB.

Se toleran las líneas ajenas al protocolo. La RFC las llama otros registros, pone `Sitemap` como ejemplo y permite al rastreador interpretarlas o no, con una condición: una línea de ese tipo nunca debe terminar un grupo.

Utilice el concepto en una decisión

Aplique «¿Qué es robots.txt?» a una decisión real. Podemos ayudarle a relacionar el concepto con el sistema que está construyendo o comprando. Un sitio o una aplicación que los asistentes de IA de sus clientes puedan encontrar, leer y utilizar, con comprobaciones en cada versión en lugar de una promesa inicial.

Explore nuestros servicios Póngase en contacto →

Compruebe los supuestos detrás del término

Utilice la explicación y los trabajos enlazados para formular preguntas sobre su proyecto. Podemos examinar los supuestos con usted antes de elegir una herramienta. Sitio Hugo multilingüe en 3 idiomas — 108 archivos de plantillas para tres idiomas y cuatro formatos

Sitio Hugo multilingüe en 3 idiomas Póngase en contacto →

Por qué importa

La RFC declara el límite en su introducción: «Estas reglas no constituyen una forma de autorización de acceso». El archivo es una petición. Un rastreador que la respeta se mantiene fuera; uno que no, lee lo que le entregue el servidor. La sección de seguridad añade que enumerar una ruta en el archivo la hace pública y localizable, por lo que cualquier contenido que deba seguir siendo privado necesita un control real, como la autenticación HTTP.

La documentación de Google acota aún más su finalidad. Describe el archivo como «utilizado principalmente para evitar sobrecargar su sitio con peticiones» y dice que «no es un mecanismo para mantener una página web fuera de Google». Una dirección prohibida puede seguir indexándose si otras páginas enlazan a ella, y aparecerá en los resultados sin descripción porque el rastreador nunca leyó la página. Para excluir una página de los resultados, Google remite a una regla `noindex` o a una contraseña; una regla `noindex` solo funciona en una página que el rastreador tiene permiso para recuperar.

Lo que el archivo hace bien es el trabajo para el que se escribió: apartar a los rastreadores de partes de un sitio que consumirían peticiones del servidor y aportarían poco a un buscador, como páginas poco importantes o casi idénticas. Su alcance también es su riesgo. Una línea bajo `*` se aplica a cada dirección

y a cada rastreador que respeta el archivo, y un cambio puede tardar hasta un día en verse. Por eso, escribir `Disallow: /` en el grupo equivocado cierra un sitio entero.

Cómo se aplica en la práctica

El archivo de este sitio es largo y casi todo son comentarios. La parte sobre la que actúa un rastreador son cuatro líneas:

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

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

Un grupo abarca a todos los rastreadores, y `Allow: /` coincide con todas las direcciones. Un grupo sin reglas permitiría exactamente lo mismo, porque una dirección que no coincide con ninguna regla está permitida, pero se escribe la línea porque esa apertura es una elección. El comentario anterior llama a la política «Decided rather than inherited», es decir, decidida en vez de heredada, y la página sobre lo que necesitan los agentes de IA de un sitio describe la política de rastreo como una de las decisiones deliberadas del sitio.

`Content-Signal` no es un registro definido por la RFC 9309. Procede del vocabulario Content Signals y aquí autoriza la recopilación para búsquedas, para responder a preguntas durante la inferencia y para entrenamiento. Según la RFC, es uno de esos otros registros: un rastreador que conoce la convención puede leerlo; uno anterior a ella se salta la línea, como dice el comentario del propio archivo; y ninguno puede tratarla como el final del grupo. La línea `Sitemap:` es otro registro de ese tipo y apunta al índice de mapas del sitio. La página de créditos resume el archivo de la misma manera: «permite todo, tanto a los rastreadores de búsqueda como a los sistemas de IA», y menciona el mapa del sitio.

Lo que el archivo omite también se decidió. Entre las cinco recomendaciones que este sitio rechazó, la página sobre preparación para IA enumera las reglas por agente y conserva una regla de robots común a todos «porque una regla escrita para un rastreador sustituye en silencio la general para ese rastreador». Si el archivo hubiera dado a un rastreador de IA un grupo propio, ese rastreador solo leería ese grupo, y nunca recibiría un `Disallow` añadido después bajo `*`.

El resto son comentarios para quien abra el archivo: las razones de la política y después las direcciones de `llms.txt`, `agents.txt`, `humans.txt`, `security.txt` y otros archivos legibles por máquinas, mantenidas como comentarios «because robots.txt has no field for them», es decir, porque robots.txt no tiene un campo para ellas. Por tanto, los comentarios forman parte de lo que se publica. Una línea `#` de la plantilla se sirve byte por byte a cada rastreador, y, en dos ocasiones, una de esas líneas contenía una nota para quien mantenía el sitio que apuntaba a un documento de planificación en un repositorio no público. Ahora una comprobación del sitio generado falla si un archivo servido contiene una ruta hacia el repositorio, y las notas de mantenimiento de la plantilla están en comentarios de plantilla que el generador elimina.

Dónde surgen los problemas

Una regla para todos deja de aplicarse a un rastreador. Alguien añade un grupo para un único bot, quizá para apartarlo de una carpeta, y desde entonces ese bot ignora todo lo situado bajo `*`, incluido cualquier `Disallow`. La RFC 9309 solo dirige un rastreador al grupo `*` cuando ningún grupo lo nombra, y la documentación de Google dice que los grupos específicos y el grupo `*` «no se combinan». Repita cada regla general dentro de cada grupo con nombre o mantenga un solo grupo, como hace este sitio.

Una página bloqueada sigue apareciendo en los resultados de búsqueda. La dirección aparece sin descripción, quizá con las palabras que otros sitios utilizaron al enlazarla. Google nunca recuperó la página, por lo que solo sabe que la dirección existe y cómo la llaman otras personas. Un `Disallow` controla el rastreo, no la indexación. Quite la regla y sirva `noindex` en la página, o protéjala con contraseña; un `noindex` oculto tras un `Disallow` nunca se lee.

El archivo anuncia lo que pretendía ocultar. Una línea como `Disallow: /admin/` indica a cualquier lector, sea o no un rastreador respetuoso, exactamente dónde buscar, y el archivo es público por diseño. La sección de seguridad de la RFC dice que enumerar las rutas las hace localizables y pide control de acceso en su lugar. Proteja el material privado mediante autenticación y no lo incluya en absoluto en el archivo.

Un error del servidor detiene el rastreo de todo un sitio. Las páginas se cargan, pero la petición de `/robots.txt` devuelve un estado 5xx o agota el tiempo de espera, y un rastreador que siga la RFC debe tratarlo como una prohibición completa. Google deja de rastrear durante las primeras 12 horas y después utiliza

su última copia válida durante un máximo de 30 días. El error contrario es más silencioso: Google trata una respuesta 4xx distinta de 429, incluidas 401 y 403, como si no existiera archivo, de modo que todas las direcciones quedan abiertas. Sirva el archivo como texto estático con estado 200 y monitorice el propio `/robots.txt`, no solo la página de inicio.

Lecturas adicionales

- RFC 9309: Protocolo de Exclusión de Robots — El propio estándar: grupos, coincidencias, tratamiento de errores, caché y la nota de seguridad.
- Introducción a robots.txt — Para qué dice Google que sirve el archivo y por qué no oculta una página de los resultados.
- Cómo interpreta Google la especificación de robots.txt — Códigos de estado, caché, campos que Google ignora y cómo elige un grupo.
- Bloquear la indexación con noindex — La regla que sí excluye una página de los resultados y por qué la página debe poder rastrearse.



Glosario

Elija según sus restricciones reales

Describa el flujo de trabajo, las interfaces existentes y los usuarios que necesitan el cambio. Podemos definir una entrega comprobable y un traspaso que permita a su equipo mantener lo entregado.

Vea cuánto cuesta Póngase en contacto →

Haga práctica la terminología

Pregunte cómo se aplica «¿Qué es robots.txt?» a su organización. Basta un ejemplo concreto de su sistema actual para iniciar la conversación.

Explore nuestros servicios

welcome@engineer.company +45 50 10 61 84 Copenhague, Dinamarca

Llévese esta página

Tarjeta de contacto

Creado con esmero y cariño por D.

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

Derechos de autor © 2025 – actualidad · Engineer ApS · Todos los derechos reservados.

Impreso desde https://engineer.company/es/glossary/robots-txt/