# ROBOTS.TXT

El archivo robots.txt es un archivo de texto sin formato, que sirve para dar instrucciones a los rastreadores de Google e informarles qué áreas de una página web se pueden rastrear y cuáles no.

Se sitúa en el directorio raíz (root) en el servidor web. Cuando un rastreador llega al sitio, lo que este hace en primer lugar es “leer” este archivo.

Crear este tipo de archivos robots.txt tiene sentido para mejorar la indexación web y el crawl budget (presupuesto/estadísticas de rastreo), aunque **no es obligatorio**.

Podemos excluir URLs individuales y directorios completos del rastreo. Incluso podemos dar instrucciones al robots.txt para que ni una sola página de todo el dominio sea visitada por el robot.

### Allow o disallow

En principio hay “solo” dos tipos de información: “allow” y “disallow” (permitir o no permitir).

Por regla general, Google puede rastrear **todos** los archivos de un dominio. Si por lo contrario, queremos excluir algunos directorios o URL del rastreo, añadimos “disallow” por delante y le asignamos un nombre al directorio. Con indicar el enlace relativo es suficiente:

Ejemplo: Disallow: /admin

De este modo, estaríamos excluyendo del rastreo el directorio de admin de una instalación web.

Hay otros directorios a los que tampoco es necesario que el robot acceda, por lo que se pueden incluir en más líneas hacia abajo.

Disallow: /wp-admin/

Disallow:/xmlrpc.php

Con el comando allow pasa justo, al contrario. Hay un directorio que no tiene que ser rastreado, pero puede haber una URL que sí debe ser rastreada, aunque esté dentro de ese directorio.

Aquí se puede crear una nueva línea que comienza con Allow para crear una excepción. De este modo, estaríamos diciéndole al motor de búsquedas “quiero que ignores todas las URLs del directorio excepto esta”

### Disallow según tipo de archivo

Mediante el uso de /\* también se pueden excluir ciertos tipos de URL del rastreo.

Ejemplos:

* Excluir todos los archivos GIF: Disallow: /\*.gif$
* Bloquear el acceso a todas las direcciones URL que usen el signo de interrogación (?) : Disallow: /\*?
* Bloquear acceso a urls que comiencen con la palabra “privado”: Disallow: /privado\*/

### Instrucciones para robots específicos

Los rastreadores y los robots se identifican con un nombre específico en el servidor. Estos pueden determinar para qué rastreador deben aplicarse las normas a través de un comando en el robots.txt.

Con un asterisco \* indicamos que todos los rastreadores deben seguir esa regla.

Google utiliza varios agentes de usuario (user agents) para rastrear la web, puedes revisarlos en la página de soporte de Google: <https://support.google.com/webmasters/answer/1061943?hl=es>

El agente de usuario “Googlebot” es especialmente importante. Si deseamos bloquear con robots.txt una página solo para este agente y no para otros rastreadores, debemos colocar el siguiente prefijo delante de la regla disallow: User-agent: Googlebot

De este modo, se puede definir exactamente qué regla ha de seguirse para cada rastreador.

Aquí tenemos un ejemplo de robots.txt

User-agent: Bingbot

Disallow: /fuentes/dtd/

User-agent: \*

Disallow: /fotos/  
Disallow: /temp/  
Disallow: /example.html

La primera parte se refiere a las reglas que debe seguir el rastreador de Bing y la segunda parte sirve para el resto de los rastreadores. (Listado de rastreadores: <https://www.keycdn.com/blog/web-crawlers>)

### Incluir robots.txt en el site map

El mapa del sitio es un documento que enumera la diferente urls del dominio.

Google también recomienda señalar la ubicación del mapa del sitio en el archivo robots.txt, ya que es el primer archivo que leen los rastreadores.

Incluye simplemente una última línea con el mapa del sitio:

Sitemap: <https://www.myexample.com/sitemap.xml>

El motor de búsqueda encontrará directamente el mapa del sitio y lo tendrá en cuenta en el tratamiento posterior de los datos, para comprender la estructura de las páginas.

### INFLUENCIA DE ROBOTS

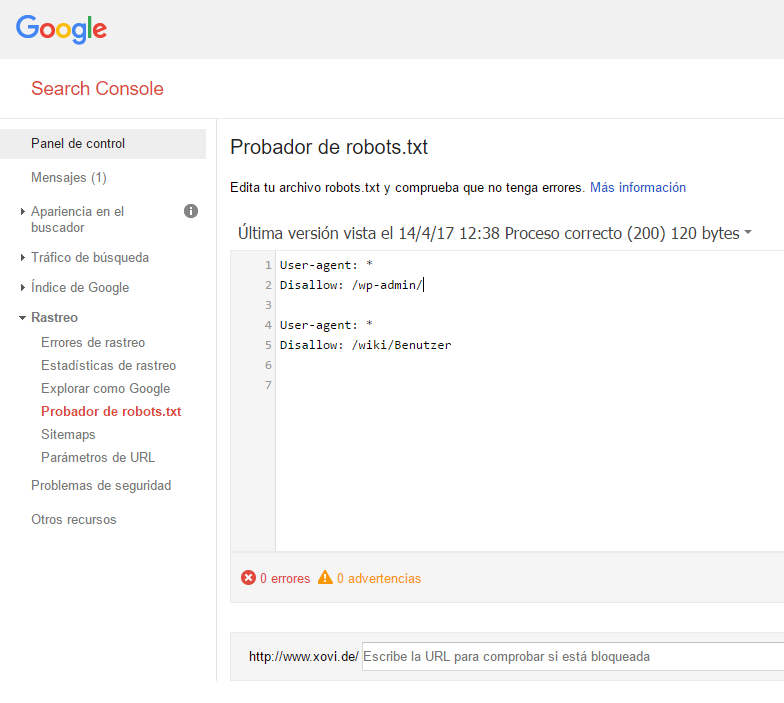
El robots.txt NO es vinculante para los rastreadores y puede ser ignorado en ocasiones. Aunque por lo general, siguen estas reglas.

* El robots.txt no impide la indexación. El rastreador puede llegar a la url desde vínculos externos o enlaces internos y añadirlos al índice.
* Si bloqueamos una página que ya estaba incluida en el índice, Google la seguirá mostrando en el resultado.
* Para asegurarnos de que Google no indexa la url, debemos permitir que vuelva a rastrear la página y acto seguido indicarle que no la indexe. ¿Cómo lo hacemos? Para ello, es preferible utilizar Meta Robots. (<https://developers.google.com/search/docs/advanced/crawling/special-tags?hl=es>)

### Prueba el archivo robots.txt con la Google Search Console

Importancia de la Google Search Console para los webmasters.

Google muestra aquí si hay problemas para leer el archivo robots.txt.

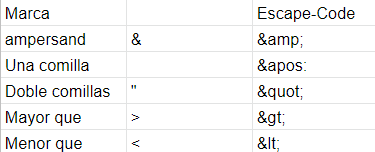


Así se prueba un archivo robots.txt en Google Search Console [https://support.google.com/webmasters/answer/6062598?hl=en](https://support.google.com/webmasters/answer/6062598?hl=en%20%20)

### Reglas Mapa del sitio

Los sitemaps tienen algunas reglas a considerar:

* Debemos usar URLs consistentes. Google rastrea las URLs tal y como las especificamos en el mapa del sitio. Por lo tanto, no debemos ingresar rutas relativas o para los dominios www la versión sin www.
* No debemos incluir ID de sesión de las URL en el mapa del sitio
* Para páginas alternativas en diferentes idiomas o regiones, usaremos un mapa del sitio con hreflag o la etiqueta hreflang para hacer referencia a estas URL alternativas.
* Si los Sitemaps grandes se dividen en varios Sitemaps más pequeños, para evitar la sobrecarga del servidor, Google debería recuperar nuestro Sitemap con mayor frecuencia. Nuestro archivo de mapa del sitio no puede contener más de 50,000 URL, y no puede tener más de 50 MB en estado no comprimido.
* Debemos usar la sintaxis de Sitemap adicional para tipos de medios adicionales, como imágenes, noticias y videos, para mejorar la indexación de estos contenidos difíciles de analizar.
* Si tenemos diferentes direcciones URL para las versiones para dispositivos móviles y de escritorio del sitio web, solo debemos referirnos a una versión. Si aún deseamos consultar ambas URL, debemos marcar las URLs del escritorio y la versión móvil con una sugerencia.
* Debemos guardar el archivo de sitemap en formato UTF-8. Para todos los valores de datos (incluidas las URLs), usar códigos de escape de entidad para los caracteres en la tabla a continuación. Nuestro sitemap solo puede contener caracteres ASCII, de lo contrario se producirán errores.



### Ubicación y sugerencia en robots.txt

El sitemap creado siempre debe estar en el directorio raíz del dominio. Si se trata de un mapa del sitio XML, debe ser accesible en www.dominio.com/sitemap.xml, porque así Google y los otros rastreadores web lo encontrarán más rápido.

Debemos apoyar a los motores de búsqueda especificando la ruta apropiada a nuestro mapa del sitio en el archivo robots.txt:

**Sitemap:** **http://www.domain.de/sitemap.xml**

Los rastreadores conocen la ubicación del sitemap, incluso si no debería estar en el directorio raíz por cualquier motivo.

### Enviar sitemap

Si la web es nueva no debemos esperar hasta que el bot de Google llegue. Idealmente, esto suele tardar unos días, pero podría tardar varias semanas.Deberemos solicitar a Google de forma proactiva que rastree la página de manera oportuna.

En Google Search Console, tenemos estas dos opciones:

* Utilizar URL Checker
* Enviar sitemap

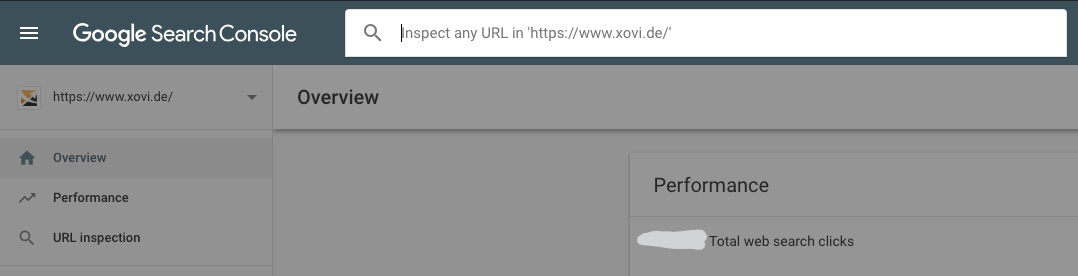
Hay que crear una cuenta en Google Search Console y demostrar la propiedad de tu sitio web.

La consola de búsqueda ofrece una variedad de herramientas e informes. Estos ayudan a medir el rendimiento del sitio web y posiblemente eliminan los errores existentes y, por lo tanto, mejoran la visibilidad y, por lo tanto, su clasificación en las SERPs (páginas de resultados de búsqueda).

#### Utilizar URL Checker

La herramienta de revisión URL es ideal si se han creado páginas de destino individuales o modificado las existentes.

* Inicia sesión en Search Console
* Selecciona la propiedad apropiada.
* Haz clic en “**URL Check**” en la barra de navegación izquierda.
* Ingresa la URL de la página de destino optimizada en la ranura de búsqueda y confirma con Enter.
* Los datos ahora se recuperarán en vivo desde el índice de Google.
* Si recibes el mensaje “**La URL está en Google**“, no tienes que hacer nada.
* Si no, haz clic en “Aplicar para la indexación”. Ahora hay una consulta en vivo sobre la indexabilidad de la URL. Esto puede tomar uno o dos minutos.
* Posteriormente aparecerá una ventana emergente con el comentario “**Se ha solicitado la indexación**” .
* Completa estos pasos para cada una de tus URL optimizadas.
* Si tienes muchas URL, puede ser más fácil volver a rastrear toda la página, es decir, volver a enviar el mapa del sitio.



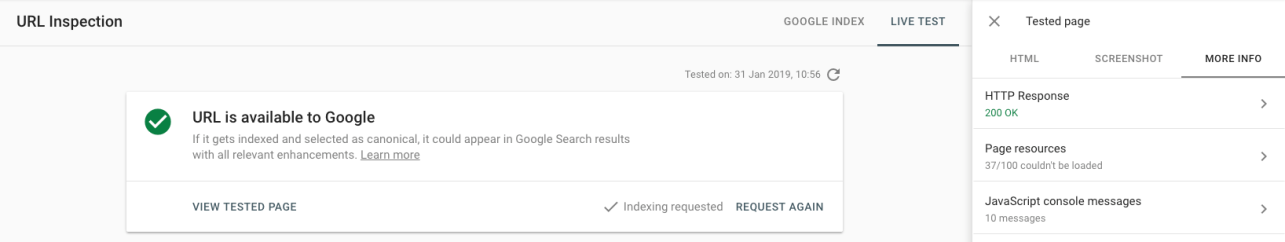
#### URL-probar

Ampliación de la herramienta de comprobación de URL

La herramienta de verificación de URL ha implementado recientemente nuevas funciones. Así que ahora puedes encontrar este botón justo arriba: boton-test-live-url

Si haces clic en él, el rastreador de Google probará en vivo la URL que ingresaste. Esto puede tomar uno o dos minutos. Luego la navegación correcta cambia de la siguiente manera:





Al hacer clic en “VER PÁGINA PRUEBA”, obtendrás esta información ampliada para su URL en tres pestañas:

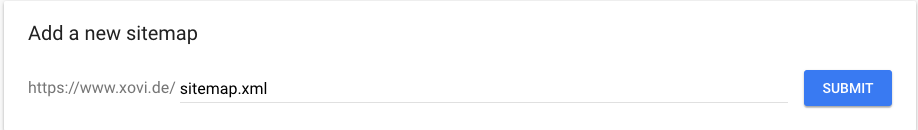
* El código fuente HTML cargado
* Una captura de pantalla de tu URL
* Más información: código de estado HTTP, recursos de la página, mensajes de consola de JavaScript

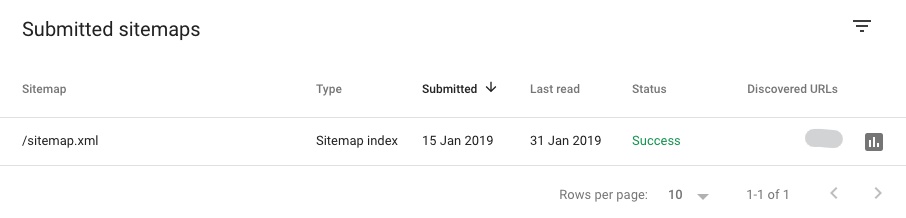
Esto facilita la resolución de problemas, ya que podremos ver de un vistazo cómo Google representa nuestra URL.

#### Enviar Google Sitemap

Si hemos realizado un sitio web completamente nuevo o cambios importantes en el sitio web existente, a menudo es más fácil volver a enviar el mapa del sitio.

* Inicia sesión en **Search Console**
* Selecciona la propiedad apropiada
* Haz clic en “**Sitemaps**” en la barra de navegación izquierda
* Ingresa la URL de tu sitemap en la casilla de búsqueda “**Agregar un nuevo sitemap**” y haz clic en “**Enviar**“. La ruta del archivo de tu mapa del sitio podría ser la siguiente: https://www.mydomain.org/sitemap.xml
* En la tabla debajo de “**Sitemaps enviados**” ahora puedes ver tu sitemap y el estado “**Correcto**” .
* Si se producen errores, te los mostrará y recibirás consejos sobre cómo solucionarlos.





Ahora que ya has solicitado con éxito la indexación de tu sitio web por Google, es el momento de esperar y ver cómo Google califica el sitio y cómo evolucionan las ubicaciones.