Robots

Il file robots.txt è un semplice **file di testo utilizzato per dare indicazioni ai web crawler**, cioè ai robot dei motori di ricerca, su come accedere e indicizzare i contenuti di un sito web.

Il file robots.txt **si trova nella directory principale** del sito web, accanto ai file HTML principali e alle cartelle dei contenuti.   
La sua posizione standard è <https://www.esempio.com/robots.txt>.

**Fornisce istruzioni ai web crawler su quali aree del sito possono essere esplorate e indicizzate e quali devono essere escluse**, evitando che vengano indicizzate aree non rilevanti, come pagine di amministrazione, file di configurazione, o pagine duplicate.

**Crawler**Noto anche come **web crawler** o **bot** (spesso chiamato anche **spider**), è un software usato dai motori di ricerca per esplorare il web automaticamente. Il suo scopo principale è scoprire nuove pagine e aggiornare i contenuti di quelle già esistenti nel database del motore di ricerca.

Quando un crawler, come Googlebot, visita un sito web, controlla per prima cosa il file robots.txt per vedere quali aree del sito può esplorare o deve evitare.   
Le istruzioni sono create utilizzando alcune regole basilari:

* **User-agent**: Specifica a quale crawler si applicano le istruzioni (ad esempio, User-agent: Googlebot per Google).
* **Disallow**: Indica quali pagine o directory devono essere escluse dall’indicizzazione (ad esempio, Disallow: /private/ per impedire l’accesso alla directory "private").
* **Allow**: Permette a un bot di accedere a specifiche pagine all’interno di una directory che altrimenti sarebbe disabilitata (ad esempio, Allow: /public/info.html).
* **Sitemap**: Inserisce il link alla mappa del sito per aiutare i crawler a trovare e indicizzare le pagine.

## Direttive rare

* **Crawl-delay**: Imposta il tempo di attesa tra le richieste di un bot, utile per siti con server di capacità limitata.
* **Host**: Indica il dominio preferito se il sito è raggiungibile tramite più domini (ad es. .com e .org).
* **Clean-param**: Dice ai motori di ricerca di ignorare determinati parametri URL, riducendo la possibilità di indicizzare contenuti duplicati.  
  Clean-param: sessionid istruisce i bot a ignorare i parametri di sessione negli URL, evitando di considerare le sessioni come contenuti diversi.

# Esempio di **robots.txt** completo

## Direttive generali:

* **User-agent: \*** : Questa direttiva indica che le istruzioni successive si applicano a tutti i bot (crawlers) senza distinzione.
* **Disallow: /** : Questa riga blocca l'accesso a tutte le pagine del sito per tutti i bot.   
  È una direttiva "generale" che verrà parzialmente sovrascritta dalle successive direttive Allow.
* **Allow: /$** : Questa direttiva permette ai bot di accedere alla homepage (l'URL che termina con /), usando $ per limitare l'accesso solo all'URL che finisce immediatamente dopo il nome del dominio, escludendo quindi sottodirectory e pagine aggiuntive.
* **Allow: /public/** : Permette ai bot di accedere a tutte le pagine e risorse che si trovano nella directory /public/, nonostante il precedente Disallow: /.
* **Allow: /products/\*.html$** : Permette l'accesso a tutte le pagine .html nella directory /products/. Qui, \* è un carattere jolly che rappresenta qualunque nome di file, mentre $ assicura che l'URL termini con .html.

## Direttive per specifici bot

* **User-agent: Googlebot** : Le istruzioni che seguono sono specifiche per Googlebot, il crawler di Google.
* **Disallow: /no-google/** : Blocca l'accesso alla directory /no-google/ solo per Googlebot.
* **Allow: /** : Permette a Googlebot di accedere a tutte le altre pagine del sito.
* **User-agent: Googlebot-Image**: Questa sezione è specifica per Googlebot-Image, il crawler di Google che si occupa solo di immagini.
* **Disallow: /private-images/** : Blocca l'accesso a tutte le immagini nella directory /private-images/ solo per Googlebot-Image.
* **User-agent: Googlebot, Bingbot, Yandex :** Questa direttiva si applica ai bot di Google, Bing e Yandex.
* **Disallow: /admin/** : Blocca l'accesso alla directory /admin/ solo per Googlebot, Bingbot e Yandex.

## Altre direttive

* **Crawl-delay: 10** :Imposta un intervallo di 10 secondi tra le richieste che i bot inviano al server per evitare sovraccarichi.
* **Clean-param: sessionid /product/** : Indica ai bot di ignorare il parametro sessionid quando si trovano nella directory /product/. Questo aiuta a ridurre la duplicazione dei contenuti che potrebbe derivare da URL dinamici.
* **Host:** [**www.example.com**](http://www.example.com) : Specifica il dominio preferito per la versione canonica del sito. I crawler daranno priorità a questo dominio rispetto a eventuali alternative come example.com senza www.
* **Sitemap:** [**https://www.example.com/sitemap.xml**](https://www.example.com/sitemap.xml) : Indica la posizione della sitemap XML, aiutando i bot a trovare e comprendere meglio la struttura del sito e i suoi contenuti.

# **User-agent strings**

**Sono sequenze di testo che i web crawler (bot) inviano quando visitano un sito web, identificandosi per far sapere ai server chi sta facendo la richiesta.**

### Identificazione

Ogni user-agent string **contiene informazioni sul tipo di bot** e sul motore di ricerca a cui appartiene (come googlebot o bingbot).  
Questo **aiuta i server a** riconoscere e **applicare le regole del file robots.txt** in modo specifico, consentendo o bloccando l’accesso in base al bot

### Evasione e Mascheramento

Alcuni attori malevoli modificano i loro user-agent string per mascherarsi come bot legittimi, eludendo le restrizioni imposte dai file robots.txt e raccogliendo dati sensibili o eccessivi.

### Differenti User-Agent per funzioni diverse

I motori di ricerca più grandi utilizzano diversi tipi di bot per scopi specifici:

* Indicizzazione dei **contenuti** (come Googlebot per Google)
* Indicizzazione di **immagini** (es. Googlebot-Image)
* **Testing** della compatibilità mobile (es. Googlebot-Mobile)

## Principali User-Agent Token

* **googlebot**: Il crawler principale di Google, incaricato dell’indicizzazione di testi e contenuti generali.
* **bingbot**: Il bot di Microsoft per Bing, che svolge funzioni simili a Googlebot.
* **yahoo! slurp**: Bot del motore di ricerca Yahoo!, ora meno diffuso ma utilizzato in alcuni paesi.
* **baiduspider**: Il crawler di Baidu, ampiamente usato in Cina
* **yandexbot**: Il bot di Yandex, che indicizza contenuti per il motore di ricerca più usato in Russia.
* **duckduckgobot**: Utilizzato da DuckDuckGo, un motore di ricerca che si focalizza sulla privacy

# **Ottimizzazione del file robot.txt**

## Prioritizzare le pagine importanti

Le pagine di maggiore valore, come le principali pagine prodotto, di servizio o di contenuto, **devono essere** **accessibili e indicizzabili** per i motori di ricerca.

## Bloccare il contenuto non necessario

**Le pagine di basso valore SEO o quelle duplicate possono essere bloccate per evitare di disperdere il valore complessivo del sito**, ad esempio pagine di accesso (login), pagine di amministrazione (admin) oppure pagine di test o sviluppo non rilevanti per l’utente

## Ottimizzazione per i dispositivi mobili

**Considerando l’indicizzazione “mobile-first”, il file robots.txt deve** **permettere l’accesso a versioni mobili di contenuti e risorse**.   
È quindi importante non bloccare elementi come CSS e JavaScript, che sono essenziali per una corretta visualizzazione sui dispositivi mobili.

## Test e monitoraggio

**Test regolari** **del file robots.txt assicurano che funzioni come previsto,** bloccando solo le pagine volute e mantenendo visibili quelle essenziali per la SEO.

Gli strumenti di Google, come la funzione di test nel **Google Search Console**, consentono di controllare come i bot leggono il file e di diagnosticare errori che potrebbero ridurre la visibilità del sito.

# Robots meta tag

È un tag HTML **usato per fornire istruzioni specifiche ai motori di ricerca su come gestire una determinata pagina del sito.**Posizionato nella sezione **<head>** del codice HTML di una pagina, questo tag consente di decidere se la pagina deve essere indicizzata, se i link al suo interno devono essere seguiti, e altro ancora.

* **<meta name="robots" content="index, follow">:** indica che la pagina può essere indicizzata e i link possono essere seguiti dai motori di ricerca.
* **<meta name="robots" content="noindex, follow">:** la pagina non sarà indicizzata (non apparirà nei risultati di ricerca), ma i link sulla pagina verranno seguiti.
* **<meta name="robots" content="index, nofollow">:** la pagina sarà indicizzata, ma i link interni non verranno seguiti.
* **<meta name="robots" content="noindex, nofollow">:** la pagina non sarà né indicizzata né i link al suo interno saranno seguiti.
* **<meta name="robots" content="noarchive">:** impedisce ai motori di ricerca di creare una copia cache della pagina.

#### Quando usare **noindex**

* **Pagine interne o di scarso valore SEO:** come pagine di login, carrello o pannelli amministrativi.
* **Contenuti duplicati**: per evitare che contenuti simili siano indicizzati, penalizzando la SEO del sito.
* **Contenuti di bassa qualità:** pagine di poco valore che potrebbero ridurre la qualità complessiva percepita del sito.
* **Contenuti temporanei**: come pagine create per eventi o promozioni temporanee, da rimuovere dopo la fine dell’evento.

#### Quando usare **nofollow**

Valore che può essere aggiunto ai link HTML (tag <a>) per indicare ai motori di ricerca di **non seguire il link** e di **non trasferire valore SEO**



* **Contenuti Esterni Non Affidabili o Generati dagli Utenti**: Link verso contenuti generati dagli utenti, come commenti nei forum o recensioni, possono contenere spam o contenuti di bassa qualità, quindi è preferibile usare nofollow per evitare qualsiasi rischio SEO.
* **Link Sponsorizzati o a Pagamento**: I link sponsorizzati devono avere l'attributo nofollow (o sponsored) per rispettare le linee guida SEO e prevenire penalità. Questo chiarisce che il link non è una raccomandazione organica, ma uno scambio commerciale.
* **Link a Pagine di Bassa Qualità o Non Rilevanti**: Aggiungere nofollow a link verso pagine di scarsa qualità o non pertinenti, come siti non verificati o pagine non rilevanti, aiuta a preservare l’autorità del proprio sito.

# **Robots.txt** vs **Robots Meta Tag**

### Robots meta tag

Offre un controllo più specifico per singole pagine, permettendo di **bloccare l’indicizzazione** senza impedire l’accesso per il crawling.

### Robots.txt

**File di testo** situato nella directory principale del sito, che fornisce regole a livello di sito o directory per gestire l’accesso dei crawler.

È utile per bloccare intere sezioni o file specifici, ma **non impedisce l’indicizzazione** se una pagina è già nota ai motori di ricerca.