Indice

Tutto ciò che devi conoscere sul file robots.txt – e non solo – per approvare o negare ai motori di ricerca l'accesso a file e cartelle sul tuo sito.

Nelle scorse settimane abbiamo visto parecchi elementi cruciali per l'**ottimizzazione SEO del tuo sito web**. Elementi cari e alla base del lavoro di ogni buon consulente SEO. Ti ho parlato del tag title, del tag meta description, dei tag heading h1… tutta roba importantissima ma, diciamolo, a livello operativo, davvero molto semplice da maneggiare per chi utilizza dei CMS come **WordPress.**

Grazie a queste piattaforme user friendly, infatti, ci si può dedicare quasi interamente al SEO copywriting senza mai – o quasi – sporcarsi le mani con il linguaggio HTML o con altri processi complessi e insidiosi: è tutto lì già bell'e pronto!

Occhio però, che non è sempre tutto così facile e immediato. Nossignore: per poter ottimizzare alla perfezione un sito web è necessario anche andare oltre a quello che ci mette

a disposizione di volta in volta WordPress, e persino più in là di quello che ci fa notare puntualmente il plugin **Yoast.**

> Tu, che hai imparato ad ottimizzare nel migliore dei modi i titoli, le parole chiave e le meta-descrizioni del tuo portale, sai come ottimizzare il file **robots.txt?**

E soprattutto, sai perché dovresti investire un po' del tuo tempo per farlo?

Ti dico subito una cosa. Nonostante non tutti conoscano il valore del robots.txt, va specificato che questo è certamente uno dei passaggi chiave per l'ottimizzazione e l'indicizzazione SEO di un sito web.

> Attenzione: un problema di configurazione a livello di questo file, potrebbe danneggiare enormemente il tuo portale, con un forte impatto negativo su **posizionamento** e sul traffico.

Non spaventarti: so che già il nome robots.txt può preoccupare, ma in realtà, se leggerai con attenzione questo post, ogni mistero sarà svelato, e così potrai godere di un sito web realmente più amico e ottimizzato nel dettaglio per i motori di ricerca**.**

Che cos'è il file robots.txt?

Come sicuramente sai, o meglio, come certamente hai già letto in qualche guida SEO, i **motori di ricerca** scansionano in lungo e in largo i contenuti della rete attraverso i suoi bot **crawler.**

Ho forse parlato di bot? Sì, ho proprio parlato di loro, perché sono loro i destinatari del nostro file robots.txt.

> Questo file testuale è uno strumento che dice ai crawler – e quindi a Google, a Yahoo!,

a Bing e a Yandex – quali Url tuo sito web possono scansionare e quali invece devono saltare, durante il loro instancabile lavoro di indicizzazione.

Dire che fino a qui è tutto chiaro, no?

In parole molto semplici, dunque, potremmo guardare al file robots.txt come ad un semaforo che indica le porte chiuse – come vedremo, però, non a chiave – accompagnate, in certi casi, da alcune porte aperte.

Vuoi mettere a tacere qualche directory o pagina? Il Robots.txt ti aiuta a suggerire a big G come apparire su Google.

Se hai letto i miei precedenti articoli sull'ottimizzazione SEO hai già capito che i motori di ricerca, quando indicizzano il mare magnum dei siti online, non guardano a caso di qua e di là. Nossignore: posano invece lo sguardo su **determinati elementi particolarmente descrittivi e rappresentativi,** come il tag title, gli headings tag come il tag h1 e via dicendo. Ma prima di tutto questo, prima di fare qualsiasi altro passaggio, i crawler si mettono a leggere velocissimamente il contenuto del tuo file robots.txt.

Già da questo puoi dunque capire che questo file testuale è davvero essenziale per l'indicizzazione del tuo sito web, perché è proprio a partire da questa lettura che i motori di ricerca creano **la lista di Url da indicizzare**.

Occhei, vedo che la nebbia intorno al robots.txt inizia pian piano a diradarsi. Vedrai che, tra qualche paragrafo, ci sarà un bel cielo sereno. Andiamo avanti?

Ma… è proprio obbligatorio avere il file robots.txt?

Sì, immagino che tu te lo stia domando: ma il file robots.txt **è proprio necessario?** È naturale farsi questa domanda, perché siamo fondamentalmente pigri, e se non afferriamo concretamente lo scopo di un'azione, se non capiamo che è proprio obbligatoria, allora non troviamo la voglia di compierla. Per questo motivo voglio subito toglierti ogni dubbio e renderti immediatamente operativo, dicendoti che sì, **il robots.txt è davvero obbligatorio**, non puoi farne a meno. Se vuoi aumentare le visite al sito sbarrare la porta al crawler di Google rappresenterebbe infatti un grosso, grossissimo errore.

Il motivo è semplice: un crawler che arriva su un sito e non trova questo file assume automaticamente che in quel portale è tutto pubblico, e che quindi **tutte le pagine vanno scansionate e indicizzate,** con le brutte conseguenze che ti spiegherò più avanti.

Bene, ora hai capito che non puoi proprio fare a meno di questo file. Ipotizziamo allora che tu, che adesso ti senta obbligato a realizzarne uno, inizi a farlo controvoglia, e senza metterci troppa attenzione.

Cosa può succedere se il tuo file robots.txt **non è formattato alla perfezione?**

Beh, ovviamente ogni singolo errore può portare a problemi diversi, ma in generale ti posso dire che, nei casi in cui un crawler non riesce a capire il contenuto di un file robots.txt, beh, non farà altro ch**e agire di testa sua,** mettendosi a scansionare a destra e a manca tutte le pagine del tuo portale, un po' come se non esistesse nessun file robots.txt.

C'è però un errore sopra a tutti gli altri che devi assolutamente evitare di fare, ovvero quello di creare un file robots.txt che blocchi totalmente l'accesso ai motori di ricerca: come diretta conseguenza, infatti, i crawler non sfioreranno nessuna delle tue pagine, rimuovendole una dopo l'altra dal proprio **indice**, rendendoti – nel tempo – pressoché i**ntrovabile attraverso i motori di ricerca**.

E questo non lo vuoi di certo, non è vero?

Bene, sai cosa vuol dire tutto questo? Che il file robots.txt va assolutamente realizzato, ma solo e unicamente dopo aver capito alla perfezione come procedere.

Dunque, abbiamo visto cos'è questo particolare file, e qual è la sua funzione. Ma tu – sì, tu –

per quale scopo preciso dovresti decidere di metterci mano?

> Stai leggendo quest'articolo perché vuoi ottimizzare il file robots.txt del tuo blog? Non perderti le mie dritte SEO per i blogger. Leggi l'approfondimento: Come scrivere un blog

Per quali motivi dovresti utilizzare il file robots.txt?

Ciò che ogni esperto SEO non faticherà a dirti è che In base alla funzione principale di questo file, sono **due i più comuni utilizzi** che tu potresti farne per ottimizzare al meglio il tuo portale, ovvero:

A) Il primo e più semplice motivo è quello di **bloccare i motori di ricerca in merito a determinate directory o pagine del tuo sito web.** Immaginiamo che tu voglia chiudere cortesemente la porta in faccia al crawler davanti alla tua pagina 'chi siamo'. Ebbene, nel file robots.txt si leggerà una cosa del tipo:

```
User-agent:*
  Disallow: /Chisiamo
```

Semplice e immediato, no?

B) Immaginiamo che il tuo sito sia davvero grande. Parlo di tante, tantissime pagine, e quindi di un sito in cui **il processo di scansione e di indicizzazione possa essere davvero pesante**. Il primo istinto dei crawler sarà ovviamente quello di scansionare tutte le cartelle, tutte le sottocartelle e tutte le pagine, ma per l'appunto questa operazione può essere piuttosto gravosa, e influenzare così negativamente le performance del portale.

CONCETTO IMPORTANTE: IL CRAWL BUDGET

Questo accade perché i crawler di Google e degli altri motori di ricerca non vanno avanti ad aria, nossignore: utilizzano corrente, server e risorse reali. Ecco perché ad ogni sito, specialmente in fase iniziale, viene assegnato un basso "crawl budget", ovvero un numero abbastanza limitato di pagine scansionabili.

> Se il crawler passa il suo tempo dietro a file e cartelle inutili consumerà – ahimè – il numero di pagine scansionabili, tralasciando quelle importanti che vorresti venissero indicizzate e posizionate.

Ed è qui che entra in gioco il file robots.txt, il quale restringendo l'accesso in determinate aree del sito **alleggerisce il processo di scansione**.

> Bada bene: non si deve mai tagliare fuori il crawler da delle pagine casuali, ma solo da quelle che non sono importanti ai fini della SEO e quindi del posizionamento. Così facendo ridurrai il carico a livello del tuo server e allo stesso tempo farai viaggiare più velocemente il processo di indicizzazione.

Occhio: il robots.txt potrebbe anche essere snobbato dai crawler

Sarebbe naturale pensare che, una volta che si dice ai crawler di non andare a ficcare il naso in una determinata pagina attraverso un '**disallow**' nel nostro file robots.txt, questi se ne stiano bene alla larga. Peccato che non succeda sempre così. E questo perché quelle contenute in questo file testuale sono delle direttive, degli accalorati consigli, delle richieste, ma **non delle regole ferree**.

> Stai pensando di lasciar stare la SEO e diventare social media manager. Fidati… aspetta ancora qualche minuto.

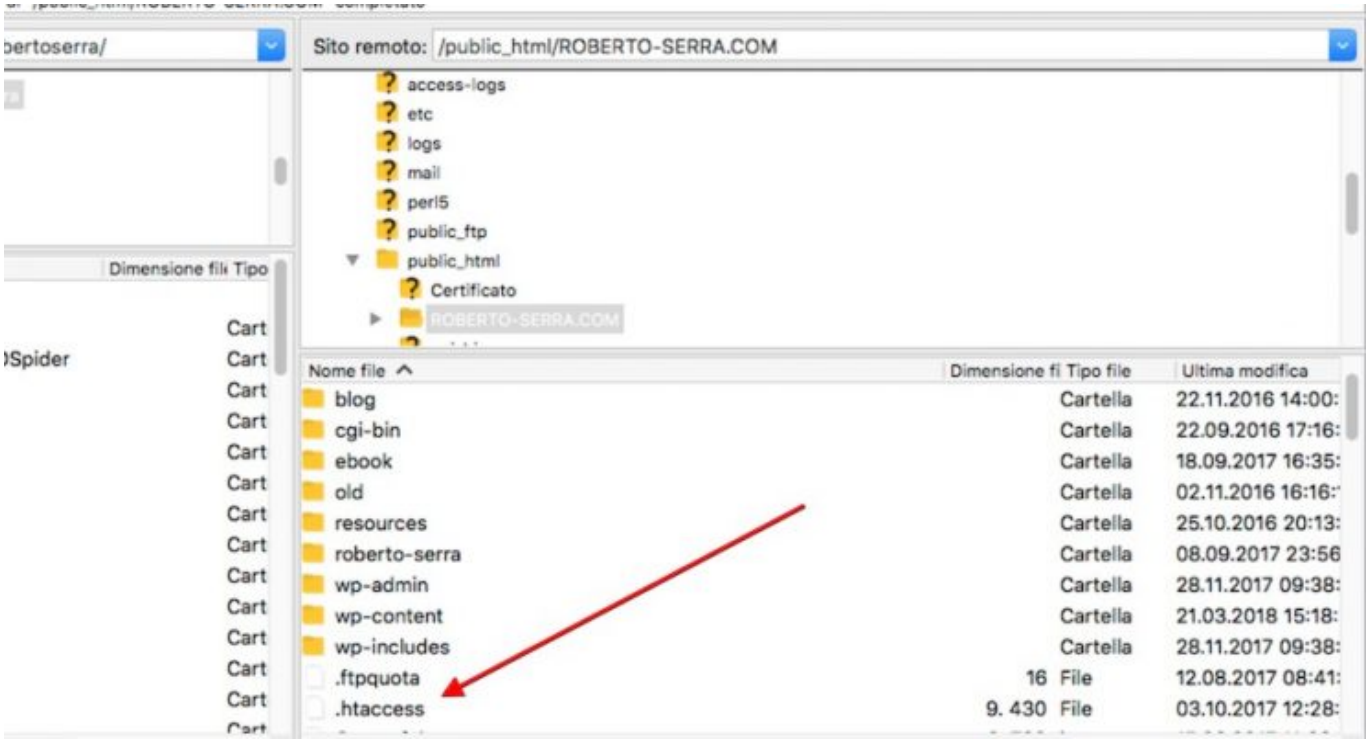
Insomma, i motori di ricerca potrebbero semplicemente decidere di non seguire le tue direttive. Per maggior parte lo faranno, ma talvolta possono decidere altrimenti, e se ne infischiano dei tuoi 'disallow'.

Per questo motivo, se ti ritrovi ad avere dei contenuti che davvero non vuoi indicizzare… beh, per essere certi di non farli finire nell'indice ti consiglio di proteggere quella particolare directory con una **password.** O potresti più semplicemente usare il file htaccess o banalmente inserire l'opzione '**noindex**' nella sezione 'head' della tua pagina.

Che dici? Vuoi vedere come si fa? Bene. Eccoti servito.

Come proteggere una directory con il file .htaccess.

Per prima cosa devi sapere che ti servono due file. Il primo è il file ".htaccess" che trovi nella root del tuo server nella stra-grande maggioranza dei casi.



Il secondo file che ci serve è il file ".htpasswd". A differenza del primo questo dovrai crearlo a manina con un comune editor di testo .txt

I sistemi spesso non consentono la generazione di file che iniziano con il punto. Per questo motivo:

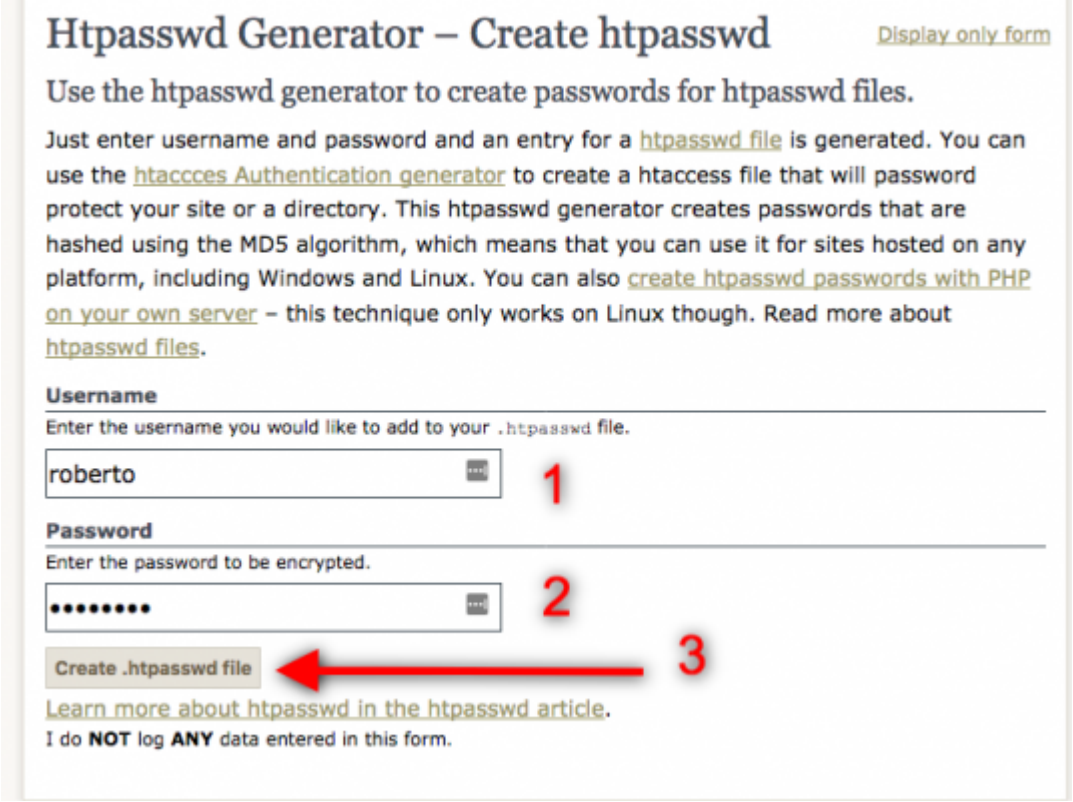
1. nomina il tuo file htpasswd.txt
2. spostalo sul server **sopra la root public_html**
3. rimuovi l'estensione .txt e rinominalo ".htpasswd"

Una volta creato e nominato il file al suo interno dovrai inserire le password che intendi usare per bloccare le cartelle.

In giro per la rete trovi tantissimi tool nati per questo scopo. Io ti segnalo

htaccesstools.com.

Vai su htaccesstools.com e genera una password criptata fornendogli user e password in chiaro.



Ottimo, a questo punto hai la tua password criptata pronta per essere copiata e incollata dentro il file correttamente posizionato **sopra la root public_html.**

Ora che il file htpasswd contiene la tua lista di nomi utente e password abilitati andiamo sul file htaccess e diciamoli che dovrà seguire queste istruzioni.

Subito sopra digitiamo quanto segue:

```
AuthUserFile /home/root-sito/.htpasswd
AuthGroupFile /dev/null
AuthName "directory"
AuthType Basic

<Limit GET POST>
```

```
require valid-user
</Limit>
```

Ora ti basterà spostare/copiare questo file nelle cartelle che vuoi proteggere e il gioco è fatto. Nessun crawler ora potrà accedere.

Perché i motori di ricerca non seguono rigidamente le istruzioni contenute nel file robots.txt?

I motivi per i quali il motore di ricerca può decidere di non ascoltare la tua richiesta espressa nel file robots.txt di evitare una determinata pagina sono tantissimi. Molto spesso delle pagine che tu ha indicato come 'chiuse' nel tuo file robots.txt vengono comunque presentate nei risultati di ricerca a causa di uno o più link diretti verso quelle stesse pagine da altre fonti già indicizzate.

> Se una delle tue pagine indicizzate contiene un link ad una risorsa dichiarata in disallow nel robots.txt i motori di ricerca tenderanno comunque ad indicizzarla. Unica eccezione quando in quest'ultima e contenuta la dichiarazione "meta noindex" nella zona <head>.

Insomma, con il file robots.txt non ci sono mai vere e proprie garanzie – ma i SEO più esperti sanno fin troppo bene che, quando si ha a che fare con i motori di ricerca, quasi nulla è del tutto certo.

Come funziona il robots.txt?

Avrai già iniziato a intuirlo: la struttura interna del robots.txt è davvero semplice, e una volta capito il suo funzionamento, è difficile sbagliare. Devi solo capire cosa sono le varie etichette che si incontrano in questo file, ovvero principalmente:

- User-agent
- Disallow
- Allow

- Crawl-delay
- Sitemap

Andiamo a vedere il loro significato e le loro funzioni!

User-agent: questo valore indica il crawler al quale vengono indirizzate le seguenti direttive. Per riferirsi al crawler generale di Google si indicherà dunque Googlebot, per le sole immagini Googlebot-Image, per Baidu baiduspider, per Bing bingbot e via dicendo. Da sottolineare che, se vuoi riferirti a tutti i crawler, ti basterà inserire un asterisco (*).

Disallow: è la direttiva che fornisce allo User-agent specificato che non deve scansionare una determinata Url. Insomma, è il vero cuore del file, la sua profonda ragion d'essere.

Allow: al contrario di Disallow, questa direttiva esplicita quali pagine o quali sottocartelle possono essere scansionate. Ma perché serve questa direttiva, quando sappiamo che i crawler sono portati a scansionare automaticamente ogni Url fino a richiesta contraria? Ebbene, questa direttiva viene utilizzata soprattutto per dare accesso a delle pagine che sono comprese in directory tacciata dalla direttiva Disallow. Chiaro no?

> Ricordati però che Allow funziona solo e unicamente per il bot di Google.

Ecco un esempio:

```
User-agent: Googlebot
Disallow: /libri
Allow: /libri/ilmiolibro/
```

Crawl-delay: questa speciale direttiva serve per dire ai crawler di aspettare un determinato ammontare di tempo prima di scansionare la prossima pagina del sito web. Il valore da inserire **si intende in millisecondi**. In questo caso, devi sapere che Googlebot fa orecchie da mercante davanti a tale direttiva: insomma, puoi usarla per Bing, Yahoo! e Yandex, ma non per mister G, per il quale invece devi andare a cambiare il setting all'interno della Google Search Console.

Sitemap: questa direttiva viene utilizzata per specificare al motore di ricerca la Sitemap del sito.

Come si realizza un file robots.txt?

Bene: ora sai pressapoco tutto quello che dovresti conoscere su questo importante file. Ma come si crea un file robots.txt a regola d'arte? Beh, come dice il suo stesso nome, si tratta di un file testuale. Non ti servono dunque strumenti particolari, ti basterà avere libero accesso al **notepad** di Windows e al pannello di controllo del tuo sito web.

E… Aspetta. Non è che tu il file robots.txt ce l'hai già?

Beh, scoprirlo non è affatto difficile. Anzi, ti dirò di più: praticamente chiunque potrebbe farlo. Ti basterà aprire il tuo **browser,** digitare il nome del tuo domino e aggiungere la stringa 'robots.txt', ovvero:

https://www.pincopallino.com/robots.txt

Se ti uscirà una pagina bianca con queste parole (o qualcosa di estremamente simile):

*User-agent: **
Allow: /

allora vuol dire che il tuo file robots.txt in realtà esiste già, e che tutto quello che devi fare è modificarlo a dovere.

Come si modifica un file robots.txt?

La prima cosa da fare è scaricare il tuo file robots.txt, utilizzando ad esempio un server FTP, e aprirlo con notepad o con un altro editor di testo. Una volta effettuate le modifiche necessarie, ti basterà ricaricare la nuova versione sul tuo server.

E invece, come si crea dal nulla un file robots.txt? Se non hai trovato alcun file robots.txt nel tuo dominio, allora devi assolutamente crearne uno: crea un file di testo, inserisci le direttive in base alle tue esigenze e caricalo immediatamente sotto alla **root directory** del tuo sito web.

> Occhio: i crawler riconosceranno e dunque useranno il tuo file solo e unicamente se questo si chiamerà esattamente robots.txt. Nomi come 'ilmiorobots.txt' non vanno bene, come nemmeno 'Robots.txt', in quanto il nome del file è **case-sensitive**. Quindi mi raccomando, scrivi tutto in minuscolo!

Vuoi un esempio base di un file robots.txt? Ecco quo:

```
User-agent: *
  Allow: /
  Sitemap: https://esempio.com/sitemap.xml
```

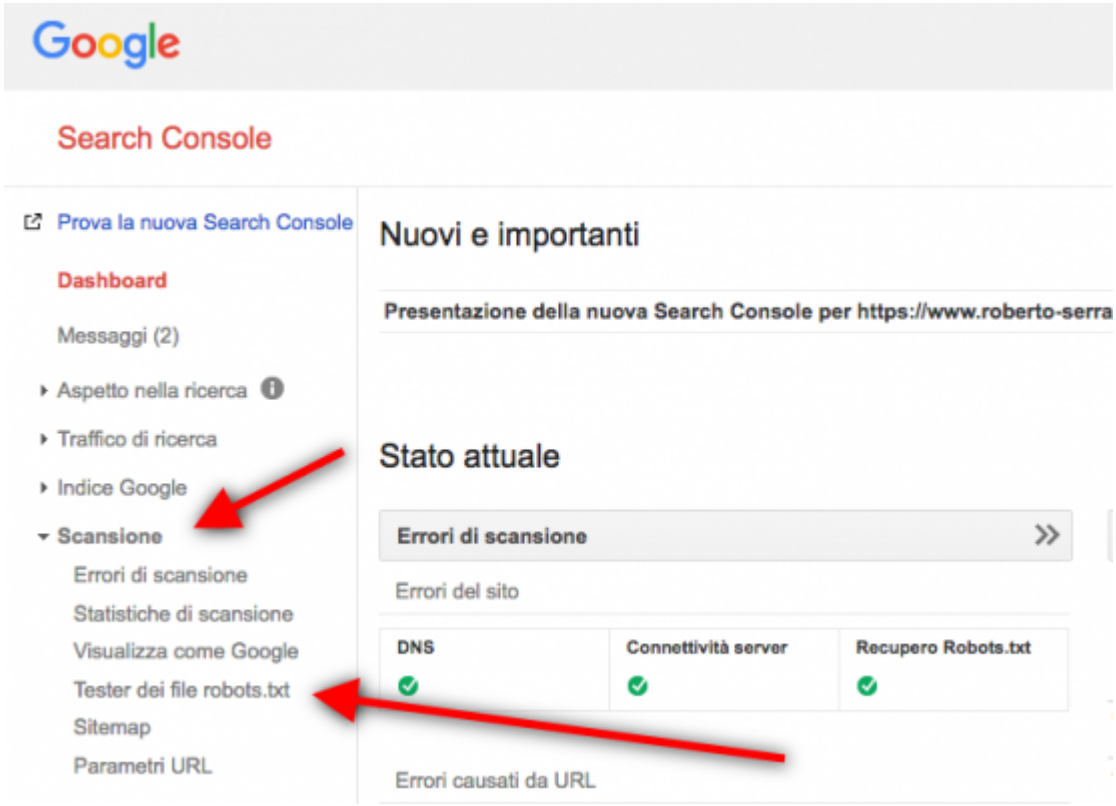
Come avrai certamente già capito, questo file robots.txt, senza alcuna altra specifica, dà il libero accesso a tutto il tuo sito web. Il che, però, non è una buona idea, tant'è che perfino Google, nelle sue linee guida, ci ricorda di '**non consentire la scansione dell'intero sito web**'. E tutti quelli che bazzicano nel mondo SEO sanno bene che, in quei rari casi in cui Google si sbilancia in questo modo, è davvero il caso si ascoltarlo!

Presente il crawl budget di cui ti ho parlato sopra? Ecco.

Come essere certi di avere realizzato un file robots.txt efficace?

Bene, ora potresti aver terminato il tuo robots.txt, inserendo tutti i vari disallow e allow. Come puoi essere certo di aver fatto un bel lavoro?

Ebbene, per tua fortuna non devi affidarti al caso, né pregare alla divinità della SEO – che davvero non so chi possa essere e dove si possa trovare. No, perché Google, nella sua **Search Console**, ha messo a nostra disposizione uno strumento estremamente utile, ovvero il **Tester dei file robots.txt**: ti basterà fare login nel tuo Google Search Console Account



inserire il tuo file e testarlo: se non ci saranno errori, Google ti darà il foglio di via, altrimenti evidenzierà la stringa o le stringhe di testo errate, così da permetterti di individuare l'errore e risolverlo prima di metterlo online.

Il file robots.txt e WordPress

Sì, lo so, appena hai letto questo titoletto ti è venuto un infarto. Tranquillo, tutto quello che hai letto fino ad ora vale anche per te che usi WordPress, né più né meno. Ci sono solamente alcuni particolari dei quali devi tener conto, ovvero:

- Prima del 2012 i siti WordPress dovevano bloccare con il file robots.txt l'accesso alle cartelle wp-admin and wp-includes. Dal 2012 in poi, però, non è più necessario farlo, in quanto è lo stesso Worpress che provvede allo scopo inserendo uno speciale tag 'noindex' nelle pagine interessate.
- Devi inoltre sapere che il file robots di default di WordPress è sempre cosi:

```
User-agent: *
 Disallow: /wp-admin/
 Allow: /wp-admin/admin-ajax.php
```

Nel caso tu – per assurdo – voglia scoraggiare i crawler a indicizzare il tuo intero sito, ti basterà attivare l'opzione nelle impostazioni del tuo CMS, e il risultato sarà questo:

```
User-agent: *
 Disallow: /
```

- WordPress usa dei file **robots.txt virtuali.** Questo significa che non puoi modificarli direttamente: puoi solo crearne uno nuovo e sostituirlo alla tua directory principale.

Cosa non fare sul file robots.txt?

Per essere certo di trasformare il tuo file robots.txt in un alleato per l'ottimizzazione SEO del tuo sito, il tuo primo passo deve essere quello di testarlo attraverso l'apposito tester di Google, in modo da essere certo che quello non finisca per oscurare della pagine che tu vuoi far apparire nei risultati del motore di ricerca.

Altro consiglio fondamentale è di **non bloccare mai le cartelle CSS o JS**: i crawler hanno infatti la necessità di vedere il tuo sito web come lo vedrebbe un utente reale in carne e ossa, e se dunque le tue pagine abbisognano di CSS e di JavaScript per funzionare, quelli non possono essere nascosti ai bot. Chiaro?

Non osare – se non è strettamente necessario – dare direttive diverse ai differenti bot dei motori di ricerca: potresti finire per creare della confusione. A questo, va aggiunto che sarebbe molto difficile tenere aggiornato il file robot.txt con queste specifiche differenenziate. Il mio consiglio, per avere un file robots.txt davvero ottimizzato, è quello di usare lo user-agent:*, così da avere delle direttive uniche per tutti i diversi crawler.

Se il tuo scopo è solo e unicamente quello di bloccare della pagine specifiche dall'essere indicizzate, ti consiglio di non passare attraverso il file robots.txt, quanto invece di utilizzare il 'noindex' nella sezione header delle singole pagine o in alternativa attraverso l'utilizzo del file htaccess.

Conclusione

Visto? La creazione e l'ottimizzazione del tuo file robots.txt non sono per nulla complicate come potevano sembrare! Potrai crearlo e modificarlo in poco tempo, e

potrai essere certo di pubblicarne uno efficiente dopo averlo testato sull'apposito tool di Google. Ora che sai tutto quello che dovevi sapere, ti consiglio di programmare quanto prima la modifica del tuo robots.txt, così da poter dare nuovo slancio all'**ottimizzazione del tuo portale**!

Fai marketing ascoltando, non strillando.
Un abbraccio.

Che mi racconti di bello?