Fichier Robots.txt : les syntaxes pour Altavista, Exalead, Fast, Google, Inktomi et Voila

Retour au sommaire de la lettre

Le fichier Robots.txt permet d'indiquer aux moteurs les zones d'un site web à indexer ou pas. Un article complet lui est consacré dans la zone gratuite du site Abondance (http://docs.abondance.com/robots.html). Nous ne reviendrons pas sur son utilisation et sa syntaxe générale, mais nous avons posé, à son sujet, trois questions aux moteurs de recherche majeurs sur le Web à l'heure actuelle (Altavista, Exalead, Fast, Google, Inktomi et Voila) :

1. Un fichier **robots.txt** est-il **indispensable** ou recommandé sur un site web par rapport à son indexation par votre moteur ? **Que se passe-t-il si ce fichier est absent** ?

2. Si on désire indiquer à TOUS les robots de ne pas indexer un répertoire (exemple "cgi-bin"), la syntaxe sera :
User-agent: *
Disallow: /cgi-bin/
Mais quelle **syntaxe** utiliser pour indiquer spécifiquement à votre robot de ne pas indexer, par exemple, le répertoire "clients" :
User-agent: ????
Disallow: /clients/
Y a -t-il plusieurs orthographes possibles (plusieurs noms de robots) ?

3. Les balises **Meta "Robots"** sont-elles prises en compte par votre moteur à l'heure actuelle ?

Voici les réponses des différents moteurs :

Réponses d'Altavista :

1. La présence de ce fichier est recommandée pour contrôler l'indexation du site, mais elle n'est pas indispensable. Si le fichier est absent, nous faisons le crawl. Cependant si le fichier est present, mais nous ne pouvons pas le lire ou l'acceder (par exemple erreur 403 "Access Forbidden"), nous ne faisons pas le crawl du site.

2. Le nom du robot est "scooter", par exemple:
User-agent: scooter
Disallow: /clients/

3. Oui, par exemple:
<META NAME="ROBOTS" CONTENT="NOINDEX, NOFOLLOW">

Réponses d'Exalead :

1. Le fichiers robots.txt n'est pas indispensable, et s'il est absent ou invalide le moteur considère que tout le site est autorisé. Le fait de savoir s'il est recommandé dépendra du contenu du site : une meilleure indexation du site sera obtenue si le fichier robots.txt bloque l'accès aux pages dont le contenu n'est pas pertinent pour le moteur (pages contenant des logs, pages générées à la volée sans contenu documentaire utile par exemple). En gros le fichier robots.txt est souvent inutile pour les sites simples, et d'autant plus utile que le site est sophistiqué ou généré par des outils sophistiqués.

2. Le robot Exalead répond aujourd'hui au doux nom de "NG" (il s'annonce par "User-Agent: NG/1.0"). Il honore donc les directives préfixées par "User-agent: NG" ou "User-agent: NG/*". Dans la prochaine version ça sera "Exabot".

3. Ce n'est pas supporté par la version qui est actuellement en production, mais c'est prévu dans la prochaine, qui devrait sortir courant septembre.

Réponses de Fast :

1. Nous n'avons besoin d'un fichier Robots.txt que si vous désirez que le robot ne crawle pas certaines parties de votre site. Vous n'êtes pénalisé en aucune façon si le fichier robots.txt n'est pas présent.

2. Utilisez le " User-agent" fourni dans vos logs si vous désirez faire une différenciation entre certains de nos crawlers. Sinon, cette syntaxe convient à tous nos crawlers :
User-Agent: fast
Disallow: /clients/

3. Oui, nous prenons en compte les options "nofollow", "noindex", "noimageindex" et "none".

Réponses de Google :

1. Le fichier n'est pas obligatoire. S'il existe, nous le prenons en compte. S'il n'existe pas, nous indexons le site.

***NDLR** : pour info, indiquons qu'auparavant, lorsque l'accès au fichier Robots.txt d'un site web générait une erreur 403 ("access forbidden"), Google n'indexait aucune page du site. Dorénavant, dans ce cas-là, le moteur fait comme si ce fichier n'existait pas et indexe les pages.*

2. La bonne syntaxe est :
User-Agent: Googlebot
Disallow: /clients/

3. Oui, Google supporte la balise Meta "Robots", mais également les balises "Googlebot" qui lui sont propres. Voir :
http://www.google.com/remove.html
http://www.google.com/webmasters/3.html#B1

Réponses d'Inktomi :

1. Le fichier n'est pas nécessaire. Par défaut, comme la plupart des autres robots, nous estimons que le fait qu'un fichier Robots.txt n'existe pas signifie que nous pouvons crawler le site dans son ensemble.

2. La bonne syntaxe est :
User-agent: slurp
Disallow: /clients/

3. Oui, pas de problèmes, nous prenons en compte la balise Meta "Robots". Voir :

http://www.inktomi.com/slurp.html

Réponses de Voila :

1. L'absence d'un tel fichier ne pose pas de problème. Nous indexons, dans ce cas, le site. Pour info, le fichier robots.txt est absent sur 93% des sites.

2. La syntaxe est :
User-agent: VoilaBot
Disallow: /clients/
A noter que le robot change très bientôt mais son nom devrait être le même (seul son numéro de version sera modifié).

3. Oui, nous suivons les directives "noindex" et "nofollow". Il y a des cas qui ne marchent pas, par ex: A -> B dans A il y a un nofollow, donc on indexe pas B. Par contre dans une page externe, si quelqu'un pointe sur B, et on crawle B.

Et un grand Merci à Mark Barlow (Altavista), Patrice Bertin (Exalead), Kristian Aune (Fast), Nate Tyler (Google), Ron Verheijen (Inktomi) et Pierre Aubert (Voila) d'avoir bien voulu nous donner ces indications.