ROBOTS.TXT

Os robôs dos buscadores são aplicativos que navegam pela internet através dos links encontrados nas páginas, em busca de conteúdo a ser indexado e exibido nos resultados de busca. Porém, você pode optar por não ter algumas de suas páginas exibidas nos resultados de busca, como por exemplo:

Páginas de Login - uma página de login a uma área restrita geralmente não deve ser indexada;
Páginas de conteúdo repetido - Caso você tenha, por exemplo, diversas Landing Pages com conteúdo bastante similar rodando para suas campanhas Google AdWords, deve bloquear as cópias e deixar apenas uma versão ser indexada pelo Google, minimizando o problema do conteúdo duplicado;
Páginas de impressão – Se o site tiver versões para tela e impressão sendo indexadas, elimine a versão para impressão do índice do Google.

O que é robots.txt

Como o próprio nome já diz, robots.txt é um arquivo no formato .txt (bloco de notas) que funciona como um filtro para os robôs dos sites de busca e faz com que os webmasters controlem permissões de acesso a determinadas páginas ou pastas dos sites.

O robots.txt controla qual informação de um site deve ou não deve ser indexada pelos sites de busca. A sintaxe do arquivo é bem simples, e deve ser colocada pelo webmaster responsável pelo site na raiz da hospedagem. O próprio Google usa um arquivo em *http://www.google.com/robots.txt*, e navegar por ele é, no mínimo, curioso.

Sintaxe do Robots.txt

O arquivo robots.txt tem o papel de criar uma política de acesso aos Robots. Para a execução dessas tarefas há palavras reservadas, ou seja, palavras com a função de comandos que permitirão ou não o acesso a determinados diretórios ou páginas de um site. Vejamos os principais comandos do arquivo robots.txt:

User-agent

A função do comando user-agent é listar quais robôs devem seguir as regras indicadas no arquivo robots.txt. Supondo que você deseje somente que o mecanismo de busca do Google siga as definições do arquivo robots.txt, basta indicar o User-agent como Googlebot. Eis as principais opções:

Google: User-agent: Googlebot
Google Imagens: User-agent: Googlebot-images
Google Adwords: User-agent: Adsbot-Google
Google Adsense: User-agent: Mediapartners-Google

Yahoo: User-agent: Slurp
Bing: User-agent: Bingbot
Todos os mecanismos: User-agent: * (ou simplesmente não incluir o comando user-agent)

Disallow

O comando instrui os sites de busca sobre quais diretórios ou páginas não devem ser incluídas no índice. Exemplos:

Disallow: /prod - orienta aos robots a não indexarem pastas ou arquivos que comecem com "prod";
Disallow: /prod/ - orienta aos robots a não indexarem conteúdo dentro da pasta "prod";
Disallow: print1.html - orienta aos robots a não indexarem conteúdo da página print1.html.

Allow

O comando Allow orienta aos robots qual diretório ou página deve ter o conteúdo indexado. Diretórios e páginas são, por definição, sempre permitidos. Assim, este comando deve ser utilizado apenas em situações em que o webmaster bloqueou o acesso a um diretório por meio do comando Disallow, mas gostaria de ter indexado um arquivo ou sub-diretório dentro do diretório bloqueado.

Note, por exemplo, no robots.txt do Google, logo no início, as duas linhas abaixo. O Allow permite que seja indexado o diretório /about abaixo do diretório /catalogs:

```
Disallow: /catalogs
Allow: /catalogs/about
```

Sitemap

Uma outra função permitia pelo robots.txt é a indicação do caminho e nome do sitemap em formato XML do site. A ferramenta para Webmasters do Google, porém, oferece um maior controle e visibilidade para a mesma função - comunicar ao Google onde está o (ou os) arquivos sitemap. Note como o Google submete, em seu robots.txt, diversos sitemaps:

```
Sitemap: http://www.google.com/hostednews/sitemap_index.xml
Sitemap: http://www.google.com/sitemaps_webmasters.xml
Sitemap: http://www.google.com/ventures/sitemap_ventures.xml
Sitemap: http://www.gstatic.com/earth/gallery/sitemaps/sitemap.xml
Sitemap: http://www.gstatic.com/s2/sitemaps/profiles-sitemap.xml
```

Cuidados com o arquivo robots.txt

Como veremos abaixo em exemplos reais de robots.txt, é muito fácil acessar o conteúdo de arquivos robots.txt de qualquer site, inclusive de concorrentes. Assim, cuidado com o que é incluído nesse arquivo. Evite colocar arquivos confidenciais. Nesses casos, o ideal

é utilizar a Meta Tag Robots <meta name="robots" /> (clique aqui para abrir a apostila de Meta Tags).

Aplicação de Robots.txt

Exemplo: o webmaster não deseja que o conteúdo do diretório/docs seja indexado pelos robots, então, bloqueou o acesso ao diretório /docs com o comando "Disallow: /docs", no arquivo robots.txt. Dentro desse diretório, porém, existe um sub-diretório chamado "public", que deve ter seu conteúdo indexado. Para que isso aconteça, basta usar no arquivo robots.txt a instrução "Allow: /docs/public/".

Exemplos reais de Robots.txt

Para olhar exemplos de arquivos robots.txt, saia navegando pela internet e inclua o arquivo /robots.txt na raíz dos sites visitados para verificar se eles utilizam o arquivo robots.txt. Veja abaixo alguns exemplos:

Google - www.google.com.br/robots.txt - alguns sites interessantes listados;
Facebook - www.facebook.com/robots.txt - Veja como este sitemap utiliza áreas separadas para cada Bot (mas sem necessidade, visto que os comandos parecem ser os mesmos para todos);
Casa Branca - www.whitehouse.gov/robots.txt - note que o comando Disallow é usado corretamente para remover áreas de login, como Disallow: /user/password/ e Disallow: /user/login/;
Abradi – www.abradi.com.br/robots.txt – Bloqueia o acesso de robots às áreas administrativas do Wordpress;
COB - www.cob.org.br/robots.txt - Bloqueia o acesso a uma área de uploads, provavelmente de arquivos submetidos por usuários.