Spiderbait! – blocking search engines from your site

Why do it?

One's first reaction on hearing about ways to block search engines from examining some or all of a website is 'Why?' Nearly everything on the web is publicly available anyway; what is the point of keeping it out of reach of a search engine? In fact, there are a variety of reasons why you would want to block some or all search engines from listing some or all of your site. Running from the most altruistic to the most personal, these include:

- Lowering the load on the search engines themselves by steering them away from areas that you know are of no interest to the casual searcher.
- Quickly removing pages or sections of your from the search engine listings when they are removed from the site.
- Saving your own bandwidth by not having to send irrelevant pages from your site to the search engine's computers.
- You may simply dislike the philosophy or disagree with the approach of a particular search engine, and not want to give it access to your site.
- Some content management systems generate large numbers of duplicate files in different directories. ROBOTS.TXT can block search engines from providing access to more than one copy of a file.
- There are 'bad robots' which trawl websites looking for things like email addresses to add to spamming lists. (Unfortunately most of these are unlikely to obey the directives in ROBOTS.TXT and have to be deterred by other means. Some even use the entries in ROBOTS.TXT as an indication of where to look. See http://www.fleiner.com/bots for details.)
- Most website statistics include a record of 'missed pages' – errors from pages looked for but not found. Since the statistics don't discriminate between human users and robots this will include any unblocked pages which the robots have been directed to or not found.
- Preventing the public display of information that may have commercial value to you. For instance, Google lists the first two lines of text on each web page that it finds, but on your site this information may have commercial value to your potential clients or competitors.

How to do it: ROBOTS.TXT

Most large search engines operate by 'trawling' or 'crawling' the Web with what used to be called 'spiders' and are now generally referred to as 'robots'. Both these terms are used figuratively, of course, to refer to software. Essentially these are programs running on the huge computer farms maintained by Google and other search engine providers that – starting with a single web page – visit all of the sites linked to by that page, analyse and record some or all of the content and then use the links on *those*

pages to find another set of pages to visit and analyse. Obviously the size of the search increases exponentially with each new level, and the number of pages coming and going on the web makes it a never-ending task. The exact algorithms used by search engines to trawl the web are closely-guarded commercial secrets, but all of them make provision for back-tracking over previously indexed pages from time to time in order to look for changes.

The ability to 'block' web robots from a particular site or portion of a site relies on an agreement known as the Robots Exclusion Protocol, which was reached in June 1994 by members of the Robots mailing list. It remains entirely voluntary, and there is currently no way to enforce its observance. Recently there has been a movement elaborate the system in order to allow more subtle distinctions that the current all-or-nothing settings: more on this later.

To block a website from being trawled, the user must add a 'ROBOTS.TXT' file to the root directory of the site. The file contains one or more groups of lines, each group relating to a specific search engine. To refer to all search engines the file uses an asterisk, while the root directory of the site is represented by a slash character. Thus the ROBOTS.TXT file below pulls away the welcome mat and requests that all search engines kindly go elsewhere:

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

More specific blocking can be achieved by targeting specific search engines and individual directories or files. Thus:

```
User-agent: Google
Disallow: /private
```

blocks only the Google search engine, and only from the subdirectory 'private', while

```
User-agent: Google
Disallow: /private/mystuff.html
Disallow: /private/bobstuff.html
User-agent: msnbot
Disallow: /notMS
```

blocks the first two files from Google only and the /notMS directory from the Microsoft Live Search tool.

Assuming you want to differentiate between search engines in this way, how can you find out what their robots are called? There is no complete list available, but most current robot names can be found at http://www.robotstxt.org/db.html. Others may be identified by looking at the search logs for your website.

An online syntax checker for ROBOTS.TXT systems can be found at http://tool.motoricerca.info/robots-checker.phtml. Google also provides a ROBOTS.TXT analysis for subscribers to its system.

Blocking individual pages with META tags

An alternative method which allows for more specificity is to block individual web pages through the use of META tags in the page HEAD section. For instance, the

following tag blocks all robots from indexing or following links from that specific page.

```
<META NAME="ROBOTS" CONTENT="NOINDEX, NOFOLLOW">
```

Individual robots can be blocked and particular types of content excluded. The following tag blocks only the Google robot from indexing the images on the page but not the text content:

```
<META NAME="GOOGLEBOT" CONTENT="NOIMAGEINDEX">
```

Other methods

Some content is screened from web robots by default: for instance, if the only way for users to access a page is by entering a valid username and password, then the robot is unlikely to be able to find it. Persistent 'bad' robots can be blocked by your web host at the server level, either by name or by identifying the IP addresses from which they come.

Expanding the system?

Given the (moderate) success of the ROBOTS.TXT method, interested parties are now suggesting that it should be expanded to take in other information about websites. A consortium of news media publishers has put forward a proposal that the system should incorporate further reductions on access to and use of website content. The new proposals, known as Automated Content Access Protocol, would bar search engines – specifically Google – from listing news summaries, headlines and photos from specified pages. The consortium has acquired a number of publishers but the only major search engine to sign up is the French Exalead, which only covers a fraction of the pages indexed by Google. Details can be found on the ACAP site at http://www.the-acap.org, but without support from Google and the other major search engines the proposal is unlikely to go far.