



## Your Website Coach



### ***Egads! I've Been Crawled! – 8/31/06***

Egads, I've been crawled! Carol Mastroianni of the Detroit Regional Chamber asks: "I just had someone say to me, '... he must have been crawling our website for all of the email addresses'. Although I understand search engines crawl the web for pages, I guess I wasn't aware there were other programs picking off information from websites."

Yes, it is true. Web crawling is what search engine spiders (robots) do to index your site. The search engine invokes a program or script (the crawler) that traverses your site and extracts the useful information so it can be indexed by the search engine. In that same vein, there are other applications out there that can target specific information on a site and bring it back to the one who sent the spider out. Perhaps at this point I should clarify that the robot or web crawler does not technically go cruising around the World Wide Web; rather it resides on its own machine and requests web pages just like your web browser. So the next logical question is "What type of information can be gathered?" Basically – anything on your site. For example, there is web crawler software that can traverse an entire site and reconstruct a duplicate site from the data it picked up. The "legitimate" uses (typically by developers) of such a program include:

- ensuring there are no broken or dead links
- creating a backup copy of your website as part of a disaster recovery plan
- version control, i.e. see what parts of your site have been updated

The less-than-sterling uses would include: duplicating another's site to see how it is put together (plagiarism) and grabbing emails to be used in spam. This is off-the-shelf software that can be purchased online or in a store. So, can robots crawl password protected pages? Search engine spiders typically can't (or won't) trespass into protected areas, but there are programs, with questionable integrity available, that do. There are a couple of things you can do to prevent robots from crawling your site. One thing is include the Robot META tag which can tell a crawler to not follow links and not index your site. However, this tag is not well supported by robots any more. Another possibility is to include a robot.txt file on your web server that lists what robots are excluded from crawling your site. Further, that file can also tell those crawlers on the "in" list where they are allowed to crawl. The thing to keep in mind is that in your quest to keep out trespassers, you could also be keeping out legitimate search engines too.

---

Written by Christine Chubenko, [christine@yourwebsitecoach.net](mailto:christine@yourwebsitecoach.net)