Keeping Google Out

This book is about partnering with Google: getting into the index, improving your PageRank, advertising on Google, distributing other people's Google ads on your site, and other ways of building your online business through Google. So a section about rebuffing Google might seem counterproductive. But in the interest of covering all bases, here it is.

Sometimes even publicity-hungry Webmasters want to keep Google away from certain parts of their business. Private pages designed for friends and semiprivate pages created for select visitors shouldn't be indexed for the world at large. Entire sites that are still under development while existing on the Web in a live state might best be excluded from Google.

It's fairly easy to prevent Google from indexing an entire site or selected pages of a site even if the spider crawls your URL. You can prevent Google also from *caching* pages of your site, a process by which Google stores each indexed page on its servers. This section explains how to prevent Google from crawling and caching your site.

Deflecting the crawl

The key to deflecting Google's spider is the *robots.txt file,* also known as the Robots Exclusion Protocol. Google's spider understands and obeys this protocol. The robots.txt file is a short, simple text file that you place in the top-level directory (root directory) of your domain server. (If you lease your Web space from your ISP, not from a dedicated Web host, you probably need administrative help in placing the robots.txt file.)

Create the robots.txt file in Notepad or another text editor, and transfer it as an ASCII text file. It's best not to use Microsoft Word or another word processor to create the robots.txt file. But if you do, remember to save it as a plain text file with the *.txt* file extension. Then make sure you transfer it to your server as a binary file, which is the default setting of many FTP (file transfer protocol) programs.

The robots.txt file contains two instructions:

- **User-agent.** This instruction specifies which search engine crawler must follow the robots.txt instructions. You may specify Google's spider, multiple specific spiders, or all spiders. (The command works for all spiders that seek and acknowledge the robots.txt file.)
- **Disallow.** This line specifies which directories (Web page folders) or specific pages at your site are off-limits to the search engine. You must include a separate `Disallow` line for each excluded directory.