# Make effective use of robots.txt

## Restrict crawling where it's not needed with robots.txt

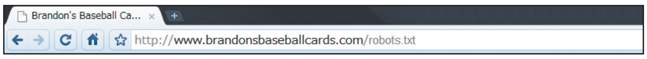
**A "robots.txt" file tells search engines whether they can access and therefore crawl parts of your site** (1). This file, which must be named "robots.txt", is placed in the root directory of your site (2).

You may not want certain pages of your site crawled because they might not be useful to users if found in a search engine's search results. If you do want to prevent search engines from crawling your pages, Google Webmaster Tools has a friendly robots.txt generator to help you create this file. Note that if your site uses subdomains and you wish to have certain pages not crawled on a particular subdomain, you'll have to create a separate robots.txt file for that subdomain. For more information on robots.txt, we suggest this Webmaster Help Center guide on using robots.txt files.

There are a handful of other ways to prevent content appearing in search results, such as adding "NOINDEX" to your robots meta tag, using .htaccess to password protect directories, and using Google Webmaster Tools to remove content that has already been crawled. Google engineer Matt Cutts walks through the caveats of each URL blocking method in a helpful video.

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

(1) All compliant search engine bots (denoted by the wildcard * symbol) shouldn't access and crawl the content under /images/ or any URL whose path begins with /search.



(2) The address of our robots.txt file.

Keep a firm grasp on managing exactly what information you do and don't want being crawled!

Best Practices

## Use more secure methods for sensitive content

You shouldn't feel comfortable using robots.txt to block sensitive or confidential material. One reason is that search engines could still reference the URLs you block (showing just the URL, no title or snippet) if there happen to be links to those URLs somewhere on the Internet (like referrer logs). Also, non-compliant or rogue search engines that don't acknowledge the Robots Exclusion Standard could disobey the instructions of your robots.txt. Finally, a curious user could examine the directories or subdirectories in your robots.txt file and guess the URL of the content that you don't want seen. Encrypting the content or password-protecting it with .htaccess are more secure alternatives.

**Avoid:**

- allowing search result-like pages to be crawled
  - users dislike leaving one search result page and landing on another search result page that doesn't add significant value for them
- allowing URLs created as a result of proxy services to be crawled

Robots Exclusion Standard

A convention to prevent cooperating web spiders/crawlers, such as Googlebot, from accessing all or part of a website which is otherwise publicly viewable.

Proxy service

A computer that substitutes the connection in cases where an internal network and external network are connecting, or software that possesses a function for this purpose.

Links

- **robots.txt generator**
  http://googlewebmastercentral.blogspot.com/2008/03/speaking-language-of-robots.html
- **Using robots.txt files**
  http://www.google.com/support/webmasters/bin/answer.py?answer=156449
- **Caveats of each URL blocking method**
  http://googlewebmastercentral.blogspot.com/2008/01/remove-your-content-from-google.html