The Basics of robots.txt

A robots.txt file instructs automated search engine crawlers which URLs the crawler can access on your site. These files can be used to provide details about which site directories or files should or should not be crawled, how quickly they should be re-crawled, and which bots are welcome or not on your website.

The robots.txt file has a very strict formatting requirement, such as JPG is not the same as jpg, and other notes below. If you ever have any questions, feel free to contact GlowHost support.

Important Details About robots.txt

- The file must be located at the root of the domain (for your main domain - /public_html/), and each subdomain needs its own file.
- The robots.txt protocol is case sensitive.
- It's easy to accidentally block crawling of everything:
    - `Disallow: /` means **disallow** everything.
    - `Disallow:` means **allow** everything.
    - `Allow: /` means **allow** everything.
    - `Allow:` means **disallow** everything.
- The instructions in robots.txt are guidance for bots, not binding requirements — bad bots may ignore your settings.

Setting Up robots.txt

Crawl Delay - A robots.txt file may specify a "crawl delay" directive for one or more user agents, which tells a bot how quickly it can request pages from a website. For example, a crawl delay of 10 specifies that a spider should not request a new page more than every 10 seconds. The lower the crawl delay, the faster the bot will be able to index your site. An example of crawl delay is below:

```
User-agent: *
Crawl-delay: 10
```

Blocking Content - Setting up the allow or disallow directives, as mentioned above, is the primary usage in an attempt to disallow directories or files that should not be searchable. While "officially", the robots.txt standard does not include wildcards, most major search engines understand it, therefore, you can use the wildcard (*****) to block all folders or names in a given path. Note that if you disallow a directory after it's already been indexed by a search engine, this may not remove that content from the search. You will need to go into the search engine's tools to request removal. Also note that search engines may index individual pages within a disallowed folder if the search engine learns about the URL from a non-crawl method, like a link from another site or your sitemap. An example of blocking content is below:

```
Disallow: /archives/
Disallow: /news/do_not_search.html
Disallow: /projects/*/private
Disallow: /images/*.jpg
```

It is strongly recommended against blocking CSS and JS files. If these files are blocked, Google can not render your website properly which may result in search indexing issues.

Sitemapping - Your robots.txt file can also include your XML sitemaps, if applicable. For example:

```
Sitemap: https://www.yoursite.com/sitemap.xml
```

More Detailed Options

Bot-specific Settings - You can also define the above instructions on a per-agent basis. For example, if you wanted to allow Google to search your site quicker than others, you can use the below as an example:

```
User-agent: Googlebot
Crawl-delay: 2
Allow: /project/*/downloads/pdf

User-agent: *
Crawl-delay: 10
Disallow: /project/*/downloads/pdf
```

IMPORTANT: Completely blocking Googlebot and bingbot will negatively affect your search engine indexing and listing.

Common Bot User-Agents - The most common user agents, or bots/crawlers, that will be checking your site are below. You can utilize these user-agents in your robots.txt formatting. Please see below:

- Baidu - **baiduspider**
- Bing - **bingbot**
- Bing - **msnbot**
- Bing (Images and Video) - **msnbot-media**
- Bing (Ads) - **adidxbot**
- Google - **Googlebot**
- Google (Images) - **Googlebot-Image**
- Google (Mobile) - **Googlebot-Mobile**
- Google (News) - **Googlebot-News**
- Google (Video) - **Googlebot-Video**
- Google (Commerce) - **Storebot-Google**
- Google (AdSense) - **Mediapartners-Google**
- Google (AdWords) - **AdsBot-Google**
- Yahoo! - **slurp**
- Yandex - **YandexBot**

You can check the bot activity on your site by clicking the "Awstats" icon inside your cPanel, selecting your domain, and scrolling to "Robots/Spiders visitors".