

5. ROBOTS.TXT

1. MAKE SURE ONLY GOOD PAGES ARE INDEXED

What we're going to do:

Hide pages that might damage our search engine ranking.

Why we're doing it:

Some components can produce thousands of pages that are useless but still get indexed by Google. One particular user installed the Events Calendar component and ended up with empty pages indexed until the year 3200. He also ended up with a heavy penalty from Google.

Google is in the process of prioritizing the way it crawls pages. If you have a lot of junk, Googlebots will crawl your site less frequently. The days of when bigger sites were always better sites are gone. Each page only has a certain amount of Page Rank and link authority to spend. You need to spend it wisely on pages that matter. If your excessive content gets out of hand, you'll be hit by a rankings penalty.

How we do it:

Use your robots.txt to stop Google indexing components that might cause trouble. Robots.txt is located in the root folder of your website.

Disallow: /badseocomponent/

The default Joomla robots.txt looks like this:

```
User-agent: *
Disallow: /administrator/
Disallow: /cache/
Disallow: /components/
Disallow: /editor/
Disallow: /help/
Disallow: /images/
Disallow: /includes/
Disallow: /language/
Disallow: /mambots/
Disallow: /media/
Disallow: /modules/
Disallow: /templates/
Disallow: /installation/
```

- 1) If we remove the /images/ line, there is a good change to Google and Yahoo image searches will index our images and hopefully drive more traffic. Be sure to give your images descriptive names.
- 2) It is not unusual to see PDF and Print pages with ranking more highly than the original content pages. I recommend you unpublish these buttons before you launch your site, but if you must publish them, block /index2.php. This will stop them from being indexed in Google.
- 3) After turning on Search Engine Friendly URLs, we like to make sure that no default Joomla URLs are indexed. Therefore, we add /index.php and /option to the list:

```
User-agent: *  
Disallow: /administrator/  
Disallow: /cache/  
Disallow: /components/  
Disallow: /editor/  
Disallow: /help/  
Disallow: /includes/  
Disallow: /language/  
Disallow: /mambots/  
Disallow: /media/  
Disallow: /modules/  
Disallow: /templates/  
Disallow: /installation/  
Disallow: /index.php  
Disallow: /index2.php  
Disallow: /option
```

This sets you up well for a default Joomla installation.

The following is a list of components that have been known to cause SEO problems with Joomla:

ExtCalendar:

<http://forum.joomla.org/index.php/topic,1227.0.html>

Events Calendar:

<http://forum.joomla.org/index.php/topic,13294.0.html>

VirtueMart:

[http://www.alledia.com/blog/search-engine-optimisation-\(seo\)/virtuemart-sometimes-causes-seo-problems/](http://www.alledia.com/blog/search-engine-optimisation-(seo)/virtuemart-sometimes-causes-seo-problems/)

MosTree

<http://www.alledia.com/blog/search-engine-optimisation-%28seo%29/using-robots.txt-to-keep-your-joomla-pages-under-control/>

Amazon Products Feed Bridge

HotProperty

JomComment

MosKnowledgebase

All of these are great components and if handled correctly, they can actually help your SEO. However, I'd recommend turning off all PDF and Print buttons, plus all RSS Feeds.

For example, we use the Amazon Products Feed Bridge on many sites. It's useful for visitors, but because the same Amazon products appear on so many other sites, it's useless for visitors. So, I simply open up the robot.txt file and stop Google from indexing the component by adding code to tell Google where NOT to look.

- 1) Carefully and regularly monitor the pages you have indexed in Google. Check for any components that have too many pages indexed. One tool we use frequently is WebCEO. Using their "Indexed Pages" tool we can easily scroll through all the pages we have in Google and spot any patterns of troublesome pages.
- 2) Check your components carefully when you set them up. If any component produces many pages without any effort on your part, change the settings to minimize those pages, or use robots.txt.
- 3) Turn off your RSS feeds unless you really believe they're going to be useful. Don't have RSS published just because its "Web 2.0" and its cool. If you

decided you do need RSS in place, use robots.txt to stop the search engines from indexing the feeds.

Finally, a warning: Search Engines do not always listen to robots.txt. If a component is really generating a large number of useless pages and robots.txt doesn't help, it might be best to remove that component.

Further Reading:

- <http://thesitewizard.com/archive/robotstxt.shtml>

2. STOP BAD PAGES FROM BEING INDEXED

What we're going to do:

Make sure we only install components that create real value for our users.

Why we're doing it:

Lots of components create extra pages, but these are often of very low value.

How we do it:

Try to avoid components that generate any of the following:

- **Tags**, especially user-generated ones. Each page only had a certain amount of authority and link power to distribute. Make sure you use that precious resource wisely rather than linking to pages full of tags.

- **RSS Feeds**, unless your users are really going to use them. Generally one RSS Feed is sufficient for most sites.
- **Social Networking links**. Lets face it...no-one is going to Digg you just because you have a link in place. They'll do it because you have great content. Otherwise you're just making the owners of Digg and Technorati rich.

If you must install them, use robots.txt mentioned in the previous section to stop search engines from indexing their pages.

Further reading:

- <http://www.seobook.com/archives/002022.shtml>
- <http://www.seobook.com/archives/002021.shtml>