Basic Search Engine Optimisation

Robots.txt file

The 'robots.txt' file is a small text file that sits in the root folder of your website. Having a robots.txt file probably won't affect your SEO much but it takes literally seconds to create one and it could influence how search engines see your website. At a bare minimum, a robots.txt file should include the following:

*User-agent: **
Allow: /

This entry tells all search engines that they are allowed to index every page on your site. You can specify individual spiders to not index certain parts, but they don't have to take this advice so might spider the whole site anyway. If a spammy spider is scraping the content of your website, telling it that it is not allowed will make no difference.

The one important thing I would advise is do not reveal any hidden urls in your robots file. I've seen examples where people have used disallow: /admin – which tells spiders not to include a link to your admin folder, but also tells potential hackers where your admin folder is located.

The only time I would advise putting more than the basic minimum into your site is when content is duplicated, e.g. if you have two paths to reach the same url. So in my shop example, I can go to www.mykesbikes.com/newarrivals/shinybike which could show the same content as www.mykesbikes.com/kids-bikes/shinybike . In this instance I might disallow the /newarrivals folder. Duplicate content can hurt your rankings but you can add special metatags called 'canonical URLs' to point Google at the correct url for the content. Canonical urls are covered later in chapter 6 "Duplicate Content Penalty."

Recently, Google has started indexing search result urls, e.g. mykekesbikes.com/search.aspx?s=search-keywords. If this is happening, you can disallow the search results pages by using:

*User-agent: **
Disallow: /?s=

Sitemaps

There are two types of sitemaps on a website. The html sitemap and the xml sitemap. The HTML sitemap is a list of links in hierarchical form designed to help users navigate complex websites. They also help search engine spiders find parts of your site that might be deeply linked.

The other type of sitemap, an xml sitemap is a file in the root of your website – often called sitemap.xml – that is structured in a formal way using the sitemaps.org guidelines. This file contains a list of your websites pages along with the following information:

1. url of the page
2. date last modified
3. change frequency – always, hourly, daily, weekly, monthly, yearly or never
4. Priority – how important this page is compared to other pages on the site, range is from 0.0 to 1.0. Default priority is 0.5. Homepage should always be set to 1.0

There are plenty of free sitemap generators online if you don't fancy writing it out by hand (who would?) A quick Google should be able to find you one easily. Sitemap generators simply spider your website and generate the xml file for you to download.

Once you have created your sitemap you can submit it to Google webmaster tools, and also Bing webmaster tools, although we would not advise you waste much time on Bing as it is not used by many people. Most search engines should be able to spider your site without the sitemap file, and if they can't then there is something wrong with your site linking structure.