Crawlability

This lesson covers crawlability. We'll explain what crawlability is, and why it is important to understand if you own or maintain a website. In addition, we'll explore how the Yoast SEO plugin takes care of a lot of aspects of crawlability for you.

What is crawlability?

Ranking in the search engines requires flawless technical SEO. To most people, this sounds a little scary. Luckily, the Yoast SEO plugin takes care of (almost) everything related to technical SEO on your website. Still, if you really want to get most out of your website, some basic knowledge of technical SEO is a must. One of the most important concepts of technical SEO is crawlability, so we'll start this technical SEO module by explaining crawlability.

The crawler

Let's revisit the concept of the crawler. A search engine like Google consists of a crawler, an index and an algorithm. The crawler – also called spider, robot or simply bot – follows links. When the crawler finds your site, it'll read your posts and pages and add the content to a gigantic database, called the index. This index is updated every time the crawler comes around your website and finds a new or revised version of it. Depending on how important Google deems your site and the amount of changes you make on your website, the crawler comes around more or less often.

Crawlability

But what exactly is crawlability? Crawlability has to do with the possibilities Google has to crawl your website. These possibilities can be restricted in a number of ways. You can block the crawler from crawling and indexing your website or certain pages on your website. If your

website or a page on your website is blocked, you're saying to Google's crawler: "do not come here". Your site or the respective page won't turn up in the search results in most of these cases.

Why block the crawler?

But why would you not want your website or page on your website to be crawled and indexed? Some of the pages on your site serve a purpose, but that purpose isn't ranking in search engines or even getting traffic to your site. For example, you wouldn't want people to find your admin and login pages in Google. People also don't want to land on a thank you page, a page that serves no other purpose than to thank the customer for purchasing something or subscribing to a newsletter.

What could prevent Google from crawling your site?

Let's discuss three methods to prevent Google from crawling or indexing your website.

1. The robots.txt file

You can generate a .txt file (a text file) named robots.txt to tell Google not to crawl a page. Before a search engine bot crawls any page it hasn't encountered before, it will open the robots.txt file for that site. The robots.txt file will tell the crawler which URLs on that site it's allowed to visit. So, using the robots.txt file, you can tell a spider where it cannot go on your site.

A robots.txt file always has the same URL: *http://example.com/robots.txt*. You can simply use a text editor to create your text file and upload it to the URL we just mentioned. We discuss the contents of a robots.txt file in our Technical SEO course, which is part of the Yoast SEO academy Premium subscription.

However, your robots.txt can't forbid a search engine to show a URL in its search results. This means that blocking the crawler on a certain page does not mean that URL will not show up in the search results. If the search engine finds enough links to that URL, it will include it, it will just not know what's on that page.

2. The HTTP header

It's also possible to use the HTTP header to prevent search engines from crawling and indexing a page. The HTTP header contains a status code, which is a message the server sends when a request made by a browser can or cannot be fulfilled. If this status code says that a page doesn't exist, the search engine won't crawl the page.

There are several status codes, with different meanings. If the status code is – for example – 200, the page exists and Google can crawl the page. However, if the status code is is 307, the page has been redirected to another URL and Google won't crawl the current URL.

3. Robots meta tags

The last method we'll discuss is using robots meta tags on your pages. You can use robots meta tags to prevent Google from indexing a page. Please note that Google will in fact crawl the page. You can, however, forbid Google from indexing the page. Robots meta tags are short pieces of code that tell Google what it can and can't do. We won't go into too much detail, but let's explore the options.

There are quite a lot of robot meta tag values, but we'll stick to the basics. To prevent Google from adding a page to its index, you can use the *noindex* value on that page. Google will then crawl the page, but it won't add it to its index. The opposite value of *noindex* is *index*. Another useful robots meta tag is the *nofollow* value. If you've been paying attention, you already know that a crawler follows links on a page. The *nofollow* value tells the crawler to not follow any links on a specific page at all. The opposite of the *nofollow* value is the *follow* value. You don't have to manually set *index* and *follow* values. They are the default values for any page the crawler will encounter. Let's look at a specific code example. If you want to disallow crawlers from indexing and following your page, this is the code you want to put into the `<head>` of your page:

```
<meta name="robots" content="noindex, nofollow">
```

Of course, you can play around with the values to reflect the situation you

want to achieve. The flow chart in Image 1 might help you understand the process crawlers follow when attempting to index a page.

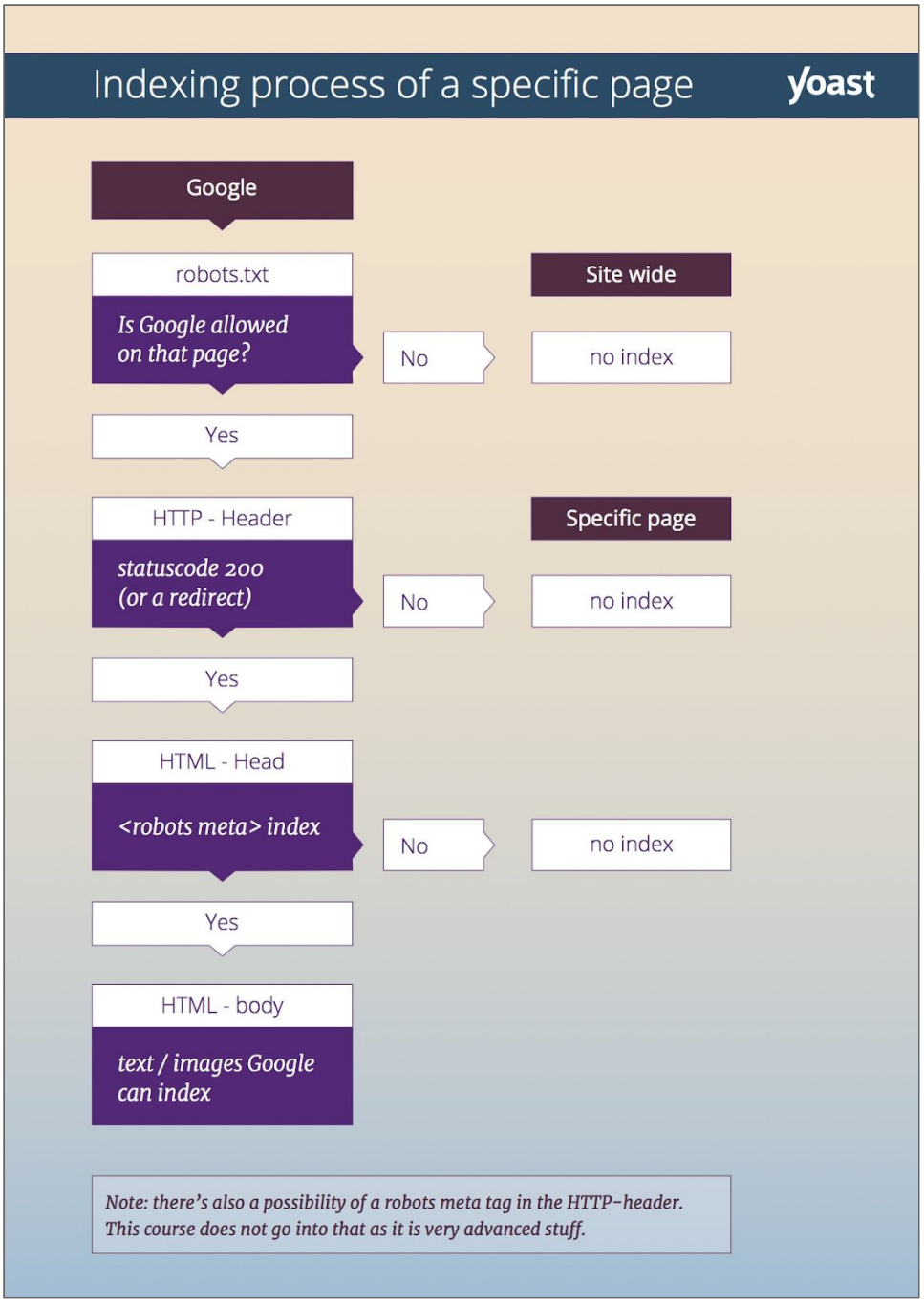


Image 1: Indexing process of a specific page

Crawlability and the Yoast SEO plugin

We understand that crawlability can be a bit tedious if you don't have a technical background. That's why we take care of a lot of aspects of

crawlability for you in our Yoast SEO plugin. Let's take a look what the plugin does, and see what options you have to make sure that Google indexes just what you want it to show in the search results.

Adding a *noindex* tag for different types of content

Yoast SEO allows you to determine whether you want each of the different content types your site has, to show up in the search results. These settings can be found on the Content Types tab of the Search Appearance settings of the plugin. For every type of content, we ask you whether you want search engines to put it in the search results. If you select "no", we add a *noindex* robots meta tag to those pages, making sure that bots will not put that type of content into their index. In the Search Appearance settings, you can do this for posts, pages, categories, tags, archives and any custom post and taxonomy types.

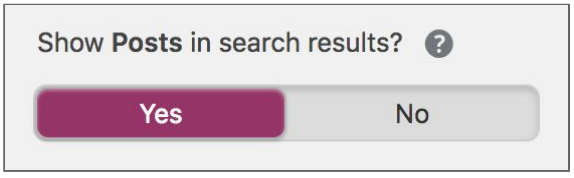


Image 2: Determining whether you want to show posts in the search results in Yoast SEO

Adding a *noindex* tag for specific posts or pages

Say that you've told the search engines to show all of your posts in the search results, via the Search Appearance settings, but you have one specific post that you don't want to appear in Google. This could be the case when it's an old article that you're not very proud of, for example. Luckily, the Yoast SEO plugin also allows you to noindex specific posts. You can do this on the post editor of that specific post. On the Advanced tab of the Yoast SEO meta box, you find the same option as we've discussed before.

Besides determining whether you want to show this specific post in the search result, you can also determine whether to allow search engines to follow links on that page. If you choose not to allow them, Yoast SEO adds a *nofollow* meta tag to the page.



Image 3: Determining whether you want to show a specific post in the search results

Editing the robots.txt file

If you want to move beyond the simple crawlability settings, the Yoast SEO plugin also allows you to edit your own robots.txt file. Editing your robots.txt file is advanced stuff, and thus beyond the scope of this course. You can learn more about this in our Technical SEO course, part of the Yoast SEO academy Premium subscription.

Check your indexability

Yoast SEO also provides you with several possibilities to quickly check your indexability. The easiest and fastest way is to go to the Yoast SEO Posts Overview on your WordPress dashboard. There, you'll find a tool called the *Indexability check by Ryte*. This tool gives you feedback on whether the homepage of your site is indexable. If it's not, you need to fix this immediately. If it is, you know that search engines are able to index at least part of your site.

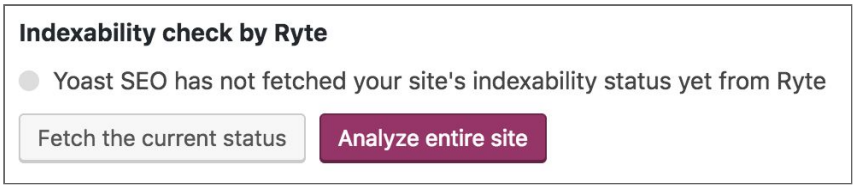


Image 4: Indexability check by Ryte

Google Search Console

The last plugin feature that's important when it comes to crawlability is the possibility to connect the plugin with Google Search Console. On the Search Console settings page, you can connect with Google Search Console and check all the crawl errors that Google encountered when

crawling your site. This is a great way to check if individual pages have crawlability issues. There are two basic types of errors Google can encounter:

- **Site errors** that affect your entire site. Think along the lines of connectivity issues with your web server, and problems fetching your robots.txt file.
- **URL errors** that affect a specific page on your website. Googlebot tried to crawl the URL but did not succeed somehow. It was able to connect to your server and then request the URL. But after that, something went wrong.

Solving these errors makes it much easier for Google to crawl your site, which can have a positive effect on your rankings.

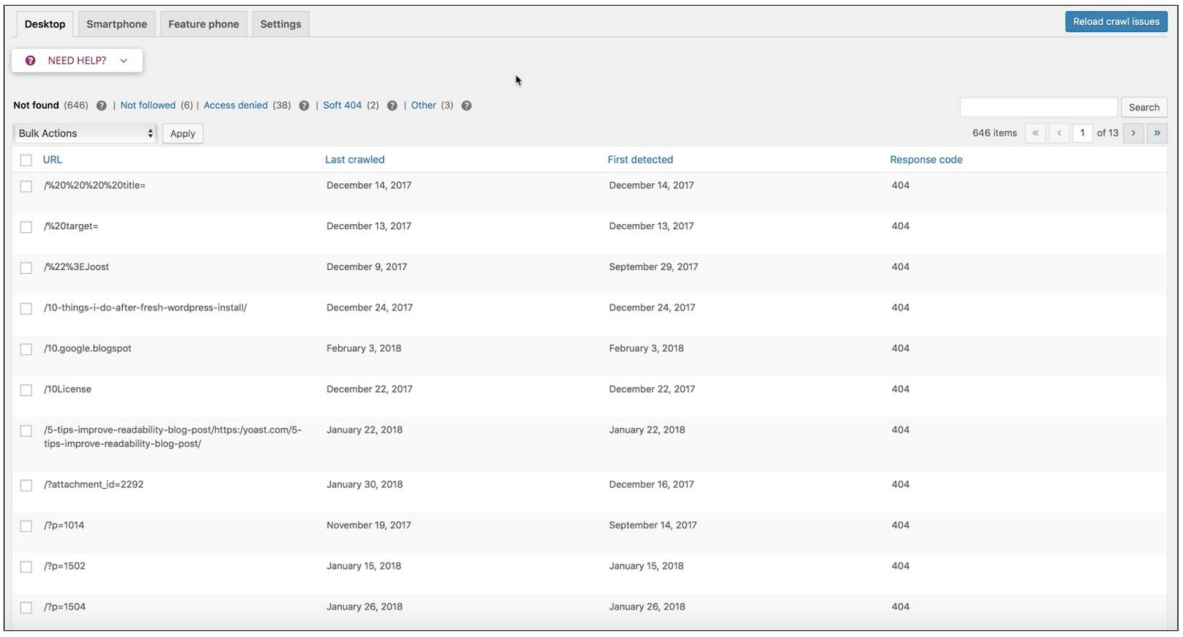


Image 5: Search Console in Yoast SEO

Conclusion

We've seen that crawlability has to do with the possibilities Google has to crawl your website, and that these possibilities can be restricted in a number of ways. We've discussed three methods to prevent Google's bot from crawling or indexing your page or website: the robots.txt file, the HTTP header, and robots meta tags. Finally, we've seen how the Yoast SEO plugin takes care of a lot of aspects of crawlability for you.