Search Engine Optimisation

In fact, so much so, that entire businesses capitalise on improving a domains SEO “ranking”. At an abstract view, search engines will “prioritise” those domains that are easier to index.

**Crawlers**

it will index the entire contents of the domain, looking for keywords and other miscellaneous information. crawlers attempt to traverse, termed as crawling, every URL and file that they can find! Say if “mywebsite.com” had the same keywords as before (“Apple", “Banana” and “Pear”), but also had a URL to another website “anotherwebsite.com”, the crawler will then attempt to traverse everything on that URL (anotherwebsite.com) and retrieve the contents of everything within that domain respectively.

Name the key term of what a "Crawler" is used to do

Top of Form



Correct Answer

Bottom of Form

What is the name of the technique that "Search Engines" use to retrieve this information about websites?

Top of Form



Correct Answer

Bottom of Form

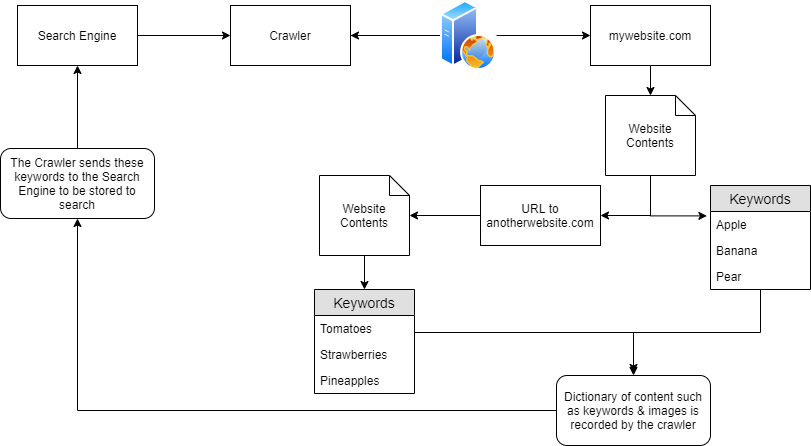
What is an example of the type of contents that could be gathered from a website?

Top of Form



Correct Answer

Bottom of Form



BeepBoop-Robots.txt

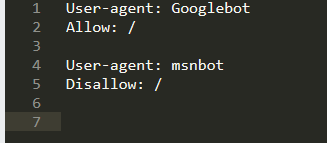
file is the first thing indexed by "Crawlers" when visiting a website. file must be served at the root directory - specified by the webserver itself. The text file defines the permissions the "Crawler" has to the website

Here we have a few keywords...

|  |  |
| --- | --- |
| Keyword | Function |
| User-agent | Specify the type of "Crawler" that can index your site (the asterisk being a wildcard, allowing  **all "User-agents"** |
| Allow | Specify the directories or file(s) that the "Crawler" **can** index |
| Disallow | Specify the directories or file(s) that the "Crawler" **cannot**index |
| Sitemap | Provide a reference to where the sitemap is located |

What if we Only Wanted Certain "Crawlers" to Index our Site?

We can stipulate so, such as in the picture below:



In this case:

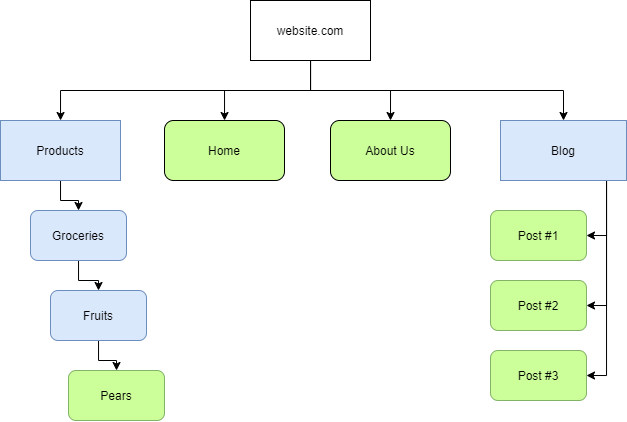
1. The "Crawler" "Googlebot" is allowed to index the entire site ("Allow: /")

2. The "Crawler" "msnbot" is not allowed to index the site (Disallow: /")

Crawlers also know the differences between sub-directories, directories and files.

**Sitemaps**

“Sitemaps” are indicative resources that are helpful for crawlers, as they specify the necessary routes to find content on the domain.



blue rectangles represent the route to nested-content, similar to a directory I.e. “Products” for a store. Whereas, the green rounded-rectangles represent an actual page.

Sitemaps” are XML formatted. presence of "Sitemaps" holds a fair amount of weight in influencing the "optimisation" and favorability of a website. Resources like "Sitemaps" are extremely helpful for "Crawlers" as the necessary routes to content are already provided! All the crawler has to do is scrape this content - rather than going through the process of manually finding and scraping

What is the typical file structure of a "Sitemap"?

Top of Form



Correct Answer

Bottom of Form

What real life example can "Sitemaps" be compared to?

Top of Form



Correct Answer

Bottom of Form

Name the keyword for the path taken for content on a website

Top of Form



Correct Answer

Bottom of Form

**Dorking**

if we wanted to narrow down our search query, we can use quotation marks. Google will interpret everything in between these quotation marks as exact and only return the results of the exact phrase provided We can use terms such as “site” (such as bbc.co.uk) and a query (such as "gchq news") to search the specified site for the keyword we have provided to filter out content that may be harder to find otherwise. For example, using the “site” and "query" of "bbc" and "gchq", we have modified the order of which Google returns the results.

A few common terms we can search and combine include:

|  |  |
| --- | --- |
| Term | Action |
| filetype: | Search for a file by its extension (e.g. PDF) |
| cache: | View Google's Cached version of a specified URL |
| intitle: | The specified phrase MUST appear in the title of the page |

* What would be the format used to query the site bbc.co.uk about flood defences

site: bbc.co.uk flood defences

Correct Answer

* What term would you use to search by file type?

filetype:

Correct Answer

* What term can we use to look for login pages?

intitle: login