Technical SEO Audit

The Crawlability and Indexability Checklist

The Search Initiative

The Crawlability & Indexability Checklist

Robots.txt

- The file is stored within the root directory of your domain. e.g. `yourdomain.com/robots.txt`.
- Include all the files and directories you don't want to index using the `Disallow` directive.
- Specify the files or directories you want to index using the `Allow` directive.
- Ensure that each directive (i.e. rule) is on a new line.
- Ensure that your file uses each user-agent only once to avoid confusion.
- If your website has multiple subdomains, use a separate robots.txt file for each one.
- Test your robots.txt using Google's own tester via Google Search Console.

The Search Initiative

The Crawlability & Indexability Checklist

XML Sitemaps

- Include any page or file that you want search engines to crawl and index on your website.
- File size should not exceed 50MB.
- It should not contain more than 50,000 URLs.
- Create a separate Video Sitemap for video content.
- For websites that publish news articles on a regular basis, create a separate Google News Sitemap.
- For any images you want crawled, create a separate Image Sitemap.
- Only indexable pages are added to your sitemap.
- Your robots.txt file includes a reference to your sitemap
- Submit your XML sitemap(s) to Google Search Console > Sitemaps > paste in the site URL of your sitemap (i.e. “sitemap.xml”) > Submit

The Search Initiative

The Crawlability & Indexability Checklist

Index Bloating

Below are some simple checks you can make in order to identify the types of pages that cause index bloating.

Once you've carried out these checks, follow this awesome guide from Ahrefs on how to go about removing them from Google's index.

Pagination

- **site:yourdomain.com inurl:page** - this returns all pages indexed that contain "**page**" in the URL.
- **site:yourdomain.com inurl:p=** - this returns all pages indexed that contain "**p=**" in the URL.

Tags

- **site:yourdomain.com inurl:tag** – this returns all indexed URLs that contain "**tag**"
- **site:yourdomain.com inurl:/tag/** – this returns all indexed URLs that contain "/**tag**/"

HTTP Pages

- **site:yourdomain.com inurl:http://** – this returns URLs that contain "**http://**" in the URL
- **site:yourdomain.com -inurl:https://** – this returns URLs that do not contain "**https://**" in the URL

The Search Initiative

The Crawlability & Indexability Checklist

eCommerce Empty Category Pages

- **`site:yourdomain.com "0 products found"`** - this returns all pages that contain the text "**`0 products found`**"

eCommerce Sorting & Views

- **`site:yourdomain.com inurl:price`** – this returns URLs that contain "**`price`**" in the URL
- **`site:yourdomain.com inurl:size`** – this returns URLs that contain "**`size`**" in the URL
- **`site:yourdomain.com inurl:color`** – this returns URLs that contain "**`color`**" in the URL
- **`site:yourdomain.com inurl:brand`** – this returns URLs that contain "**`brand`**" in the URL

Others

Here's a list of other possible checks for index bloat that you should look out for:

- Pages with and without trailing slashes
- Various pages that serve the same content.
- Pages with and without capital letters
- Pages for different device types
- Pages with AMP and non-AMP

The Search Initiative

The Crawlability & Indexability Checklist

Meta Robots Tag

- Meta robots tag should be in the **<head>** section.
- Use **<meta name="robots" content="noindex">** for pages tyou don't want to be index.
- Use **<meta name="googlebot" content="noindex">** for pages you don't want Googlebot to index.
- Parameters should not conflict with each other, Google will use the most restrictive parameter by default.
- If page isn't indexed, check whether the meta robots tag has been set with the "**noindex**" parameter.

X-Robots Tag

- If a page still isn't indexed, check to see whether an X-Robots tag has been set with the "**noindex**" parameter.
- To block specific pages, add the following syntax to the HTTP header of that page: **Header("X-Robots-Tag: noindex", true);**
  - You can also use the same directives as you would for the Meta Robots Tag i.e. noindex, nofollow, index, follow etc.
- To block file types from being indexed without having to list all of them in your robots.txt file, use the following syntax in your .htaccess: **<FilesMatch ".(doc|pdf)$">Header set X-Robots-Tag "noindex, noarchive, nosnippet"</FilesMatch>**

The Search Initiative

The Crawlability & Indexability Checklist

Canonical Tag

- Use absolute URLs instead of relative paths.
- Use lowercase URLs instead of capitalized URLs
- Use the correct version of your domain i.e. HTTPS or HTTP.
- Canonicalize your homepage
- Don't use multiple canonicals
- `rel=canonical` should appear in the `<head>` or HTTP header
- Canonicalized (master) pages shouldn't be blocked
- Non-canonicalized pages shouldn't appear in your sitemap

HTTP Status Codes

301 Moved Permanently

You should use a 301 redirect:

- If you are changing the URL of a page or subfolder
- If you are changing from a subdomain to a subfolder
- If you are moving your entire website to a new domain
- If you are switching from HTTP to HTTPS
- When switching from www to non-www & vice versa

When auditing 301 redirects, you should look out for:

- 404 pages that redirect to the homepage
- Redirect chains
- Infinite redirect chains
- Broken 301 redirects
- HTTP pages that don't redirect to HTTPS

The Search Initiative

The Crawlability & Indexability Checklist

302 Found / Moved Temporarily

- [ ] Ensure any 302 redirects haven't been left for a long time
- [ ] Ensure that any 302 redirects are implemented for temporary redirects only

404 Not Found

- [ ] Ensure your site returns a custom 404 page
- [ ] Implement a redirect for pages that have moved

Soft 404s

- [ ] Check for soft 404s
- [ ] If the page no longer exists - It should return a 404 (not found) or 410 (gone) response
- [ ] If the page has moved - You should return a 301 redirect

Crawl Errors

Look at the Index Coverage report on Google Search Console. **Check for and fix the following crawl errors:**

- [ ] Server error (5xx)
- [ ] Redirect error
- [ ] Submitted URL blocked by robots.txt
- [ ] Submitted URL marked 'noindex'
- [ ] Submitted URL seems to be a Soft 404
- [ ] Submitted URL returns unauthorized request (401)
- [ ] Submitted URL not found (404)
- [ ] Submitted URL returned 403
- [ ] Submitted URL blocked due to another 4xx issue

The Search Initiative

Want to make sure your SEO is working for you?

- Want help figuring out the best SEO strategy for your site?
- Don't want to commit to monthly payments for an ongoing SEO management service?
- Interested in a straight-forward master plan you can take into action right now?

If you answered "Yes" to any of these questions, then we'd like to talk to you!

CLICK HERE: To Get A SEO Audit From the Search Initiative

The Search Initiative