Delete your pages and rank higher in search – Index bloat and technical optimization 2019



If you're looking for a way to optimize your site for technical SEO and rank better, consider deleting your pages.

I know, crazy, right? But hear me out.

We all know Google can be slow to index content, especially on new websites. But occasionally, it can aggressively index anything and everything it can get its robot hands on whether you want it or not. This can cause terrible headaches, hours of clean up, and subsequent maintenance, especially on large sites and/or ecommerce sites.

Our job as search engine optimization experts is to make sure Google and other search engines can first find our content so that they can then understand it, index it, and rank it appropriately. When we have an excess of indexed pages, we are not being clear with how we want search engines to treat our pages. As a result, they take whatever action they deem best which sometimes translates to indexing more pages than needed.

Before you know it, you're dealing with index bloat.

What is the index bloat?

Put simply, index bloat is when you have too many low-quality pages on your site indexed in search engines. Similar to bloating in the human digestive system (disclaimer: I'm not a doctor), the result of processing this excess content can be seen in search engines indices when their information retrieval process becomes less efficient.

Index bloat can even make your life difficult without you knowing it. In this puffy and uncomfortable situation, Google has to go through much more content than necessary (most of the times low-quality and internal duplicate content) before they can get to the pages you want them to index.

Think of it this way: Google visits your XML sitemap to find 5,000 pages, then crawls all your pages and finds even more of them via internal linking, and ultimately decides to index 30,000 URLs. This comes out to an indexation excess of approximately 500% or even more.

But don't worry, diagnosing your indexation rate to measure against index bloat can be a very simple and straight forward check. You simply need to cross-reference which pages you want to get indexed versus the ones that Google is indexing (more on this later).

The objective is to find that disparity and take the most appropriate action. We have two options:

1. Content is of good quality = Keep indexability
2. Content is of low quality (thin, duplicate, or paginated) = noindex

You will find that most of the time, index bloat results in removing a relatively large number of pages from the index by adding a "NOINDEX" meta tag. However, through this indexation analysis, it is also possible to find pages that were missed during the creation of your XML sitemap(s), and they can then

be added to your sitemap(s) for better indexing.

Why index bloat is detrimental for SEO

Index bloat can slow processing time, consume more resources, and open up avenues outside of your control in which search engines can get stuck. One of the objectives of SEO is to remove roadblocks that hinder great content from ranking in search engines, which are very often technical in nature. For example, slow load speeds, using noindex or nofollow meta tags where you shouldn't, not having proper internal linking strategies in place, and other such implementations.

Ideally, you would have a 100% indexation rate. Meaning every quality page on your site would be indexed – no pollution, no unwanted material, no bloating. But for the sake of this analysis, let's consider anything above 100% bloat. Index bloat forces search engines to spend more resources (which are limited) than needed processing the pages they have in their database.

At best, index bloat causes inefficient crawling and indexing, hindering your ranking capability. But index bloat at worst can lead to keyword cannibalization across many pages on your site, limiting your ability to rank in top positions, and potentially impacting the user experience by sending searchers to low-quality pages.

To summarize, index bloat causes the following issues:

1. Exhausts the limited resources Google allocates for a given site
2. Creates orphaned content (sending Googlebot to dead-ends)
3. Negatively impacts the website's ranking capability
4. Decreases the quality evaluation of the domain in the

eyes of search engines

Sources of index bloat

1. Internal duplicate content

Unintentional duplicate content is one of the most common sources of index bloat. This is because most sources of internal duplicate content revolve around technical errors that generate large numbers of URL combinations that end up indexed. For example, using URL parameters to control the content on your site without proper canonicalization.

Faceted navigation has also been one of the “thorniest SEO challenges” for large ecommerce sites, as Portent describes, and has the potential of generating billions of duplicate content pages by overlooking a simple feature.

2. Thin content

It’s important to mention an issue introduced by the Yoast SEO plugin version 7.0 around attachment pages. This WordPress plugin bug led to “Panda-like problems” in March of 2018 causing heavy ranking drops for affected sites as Google deemed these sites to be lower in the overall quality they provided to searchers. In summary, there is a setting within the Yoast plugin to remove attachment pages in WordPress – a page created to include each image in your library with minimal content – the epitome of thin content for most sites. For some users, updating to the newest version (7.0 then) caused the plugin to overwrite the previous selection to remove these pages and defaulted to index all attachment pages.

This then meant that having five images per blog post would lead to 5x-ing the number of indexed pages with 16% of actual quality content per URL, causing a massive drop in domain

value.

3. Pagination

Pagination refers to the concept of splitting up content into a series of pages to make content more accessible and improve user experience. This means that if you have 30 blog posts on your site, you may have ten blog posts per page that go three pages deep. Like so:

- https://www.example.com/blog/
- https://www.example.com/blog/page/2/
- https://www.example.com/blog/page/3/

You'll see this often on shopping pages, press releases, and news sites, among others.

Within the purview of SEO, the pages beyond the first in the series will very often contain the same page title and meta description, along with very similar (near duplicate) body content, introducing keyword cannibalization to the mix. Additionally, the purpose of these pages is for a better browsing user experience for users already on your site, it doesn't make sense to send search engine visitors to the third page of your blog.

4. Under-performing content

If you have content on your site that is not generating traffic, has not resulted in any conversions, and does not have any backlinks, you may want to consider changing your strategy. Repurposing content is a great way to maximize any value that can be salvaged from under-performing pages to create stronger and more authoritative pages.

Remember, as SEO experts our job is to help increase the overall quality and value that a domain provides, and improving content is one of the best ways to do so. For this, you will need a content audit to evaluate your own individual

situation and what the best course of action would be.

Even a 404 page that results in a 200 Live HTTP status code is a thin and low-quality page that should not be indexed.

Common index bloat issues

One of the first things I do when auditing a site is to pull up their XML sitemap. If they're on a WordPress site using a plugin like Yoast SEO or All in One SEO, you can very quickly find page types that do not need to be indexed. Check for the following:

- Custom post types
- Testimonial pages
- Case study pages
- Team pages
- Author pages
- Blog category pages
- Blog tag pages
- Thank you pages
- Test pages

To determine if the pages in your XML sitemap are low-quality and need to be removed from search really depends on the purpose they serve on your site. For instance, sites do not use author pages in their blog, but still, have the author pages live, and this is not necessary. "Thank you" pages should not be indexed at all as it can cause conversion tracking anomalies. Test pages usually mean there's a duplicate somewhere else. Similarly, some plugins or developers build custom features on web builds and create lots of pages that do not need to be indexed. For example, if you find an XML sitemap like the one below, it probably doesn't need to be indexed:

- https://www.example.com/tcb_symbols_tax-sitemap.xml

Different methods to diagnose index bloat

Remember that our objective here is to find the greatest contributors of low-quality pages that are bloating the index with low-quality content. Most times it’s very easy to find these pages on a large scale since a lot of thin content pages follow a pattern.

This is a quantitative analysis of your content, looking for volume discrepancies based on the number of pages you have, the number of pages you are linking to, and the number of pages Google is indexing. Any disparity between these numbers means there’s room for technical optimization, which often results in an increase in organic rankings once solved. You want to make these sets of numbers as similar as possible.

As you go through the various methods to diagnose index bloat below, look out for patterns in URLs by reviewing the following:

- URLs that have /dev/
- URLs that have “test”
- Subdomains that should not be indexed
- Subdirectories that should not be indexed
- A large number of PDF files that should not be indexed

Next, I will walk you through a few simple steps you can take on your own using some of the most basic tools available for SEO. Here are the tools you will need:

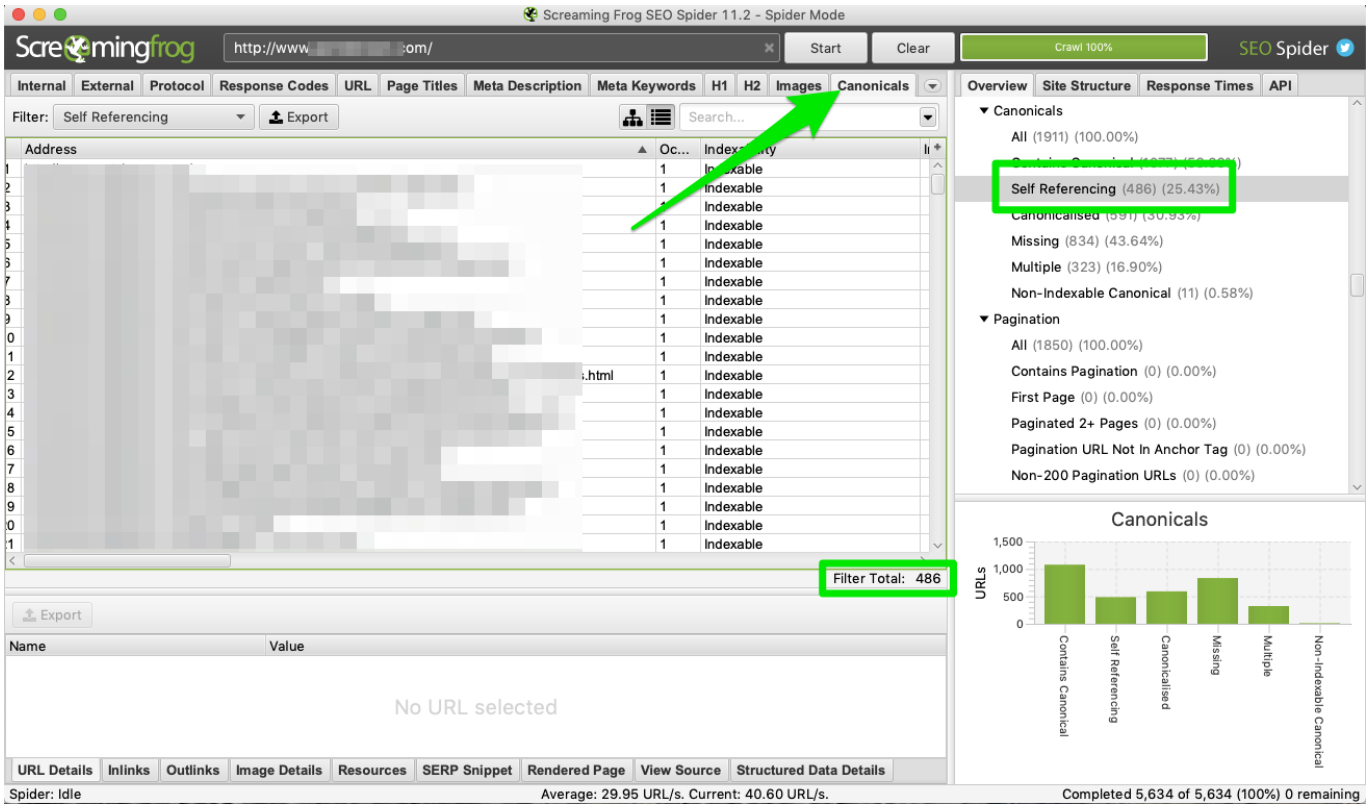
- Paid Screaming Frog
- Verified Google Search Console
- Your website’s XML sitemap
- Editor access to your Content Management System (CMS)
- Google.com

As you start finding anomalies, start adding them to a

spreadsheet so they can be manually reviewed for quality.

1. Screaming Frog crawl

Under Configuration > Spider > Basics, configure Screaming Frog to crawl (check “crawl all subdomains”, and “crawl outside of start folder”, manually add your XML sitemap(s) if you have them) for your site in order to run a thorough scan of your site pages. Once the crawl has been completed, take note of all the indexable pages it has listed. You can find this in the “Self-Referencing” report under the Canonicals tab.

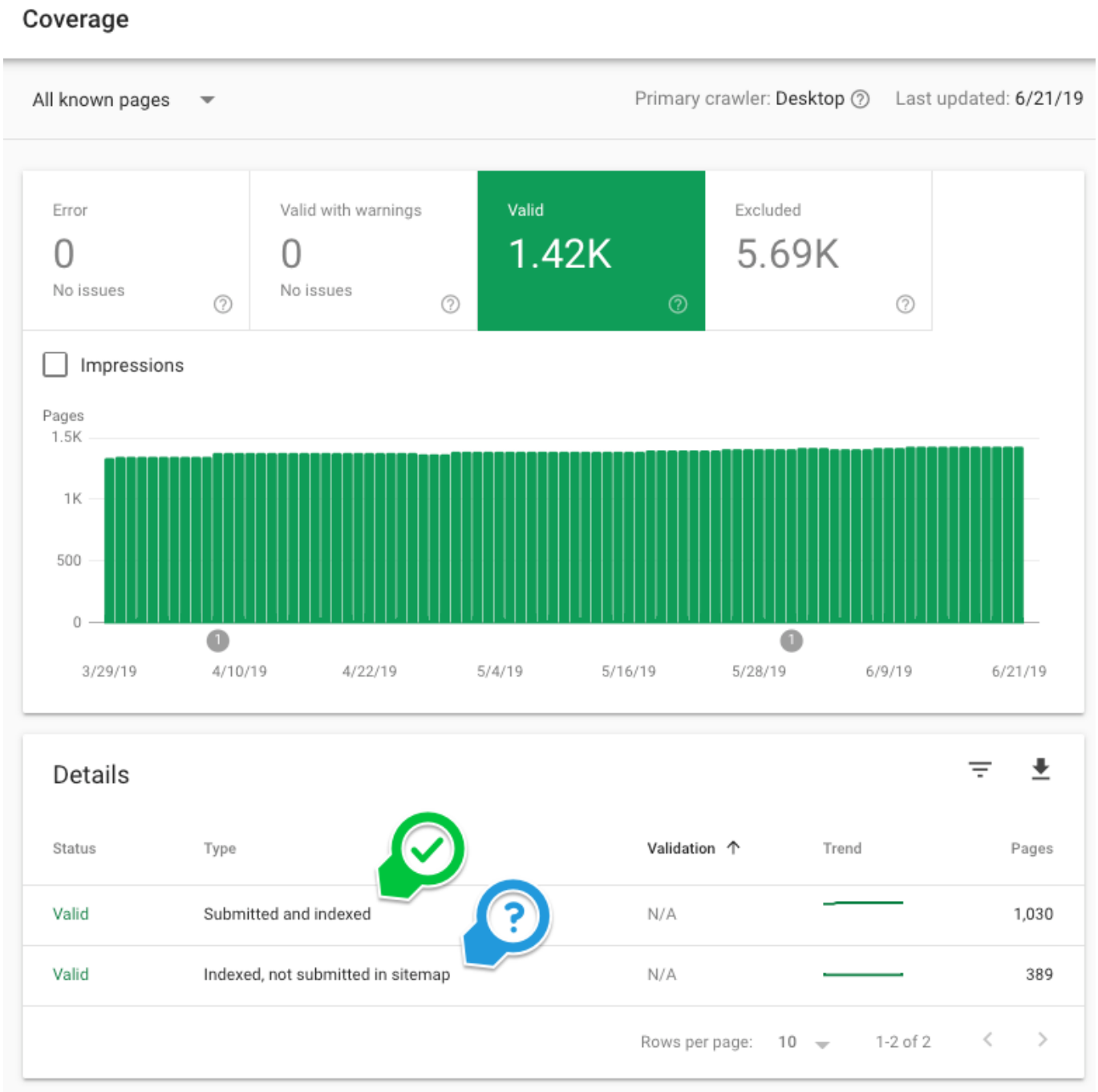


Take a look at the number you see. Are you surprised? Do you have more or fewer pages than you thought? Make a note of the number. We’ll come back to this.

2. Google’s Search Console

Open up your Google Search Console (GSC) property and go to the Index > Coverage report. Take a look at the valid pages. On this report, Google is telling you how many total URLs they

have found on your site. Review the other reports as well, GSC can be a great tool to evaluate what the Googlebot is finding when it visits your site.



How many pages does Google say it’s indexing? Make a note of the number.

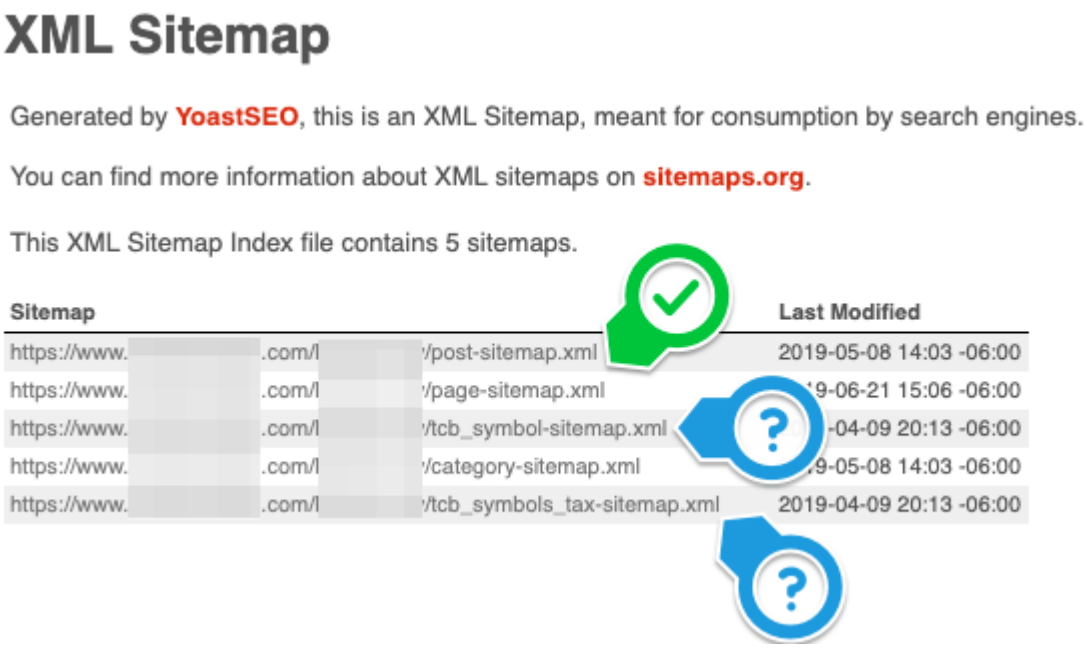
Share this article

Related articles

3. Your XML sitemaps

This one is a simple check. Visit your XML sitemap and count the number of URLs included. Is the number off? Are there unnecessary pages? Are there not enough pages?

Conduct a crawl with Screaming Frog, add your XML sitemap to the configuration and run a crawl analysis. Once it’s done, you can visit the Sitemaps tab to see which specific pages are included in your XML sitemap and which ones aren’t.



Make a note of the number of indexable pages.

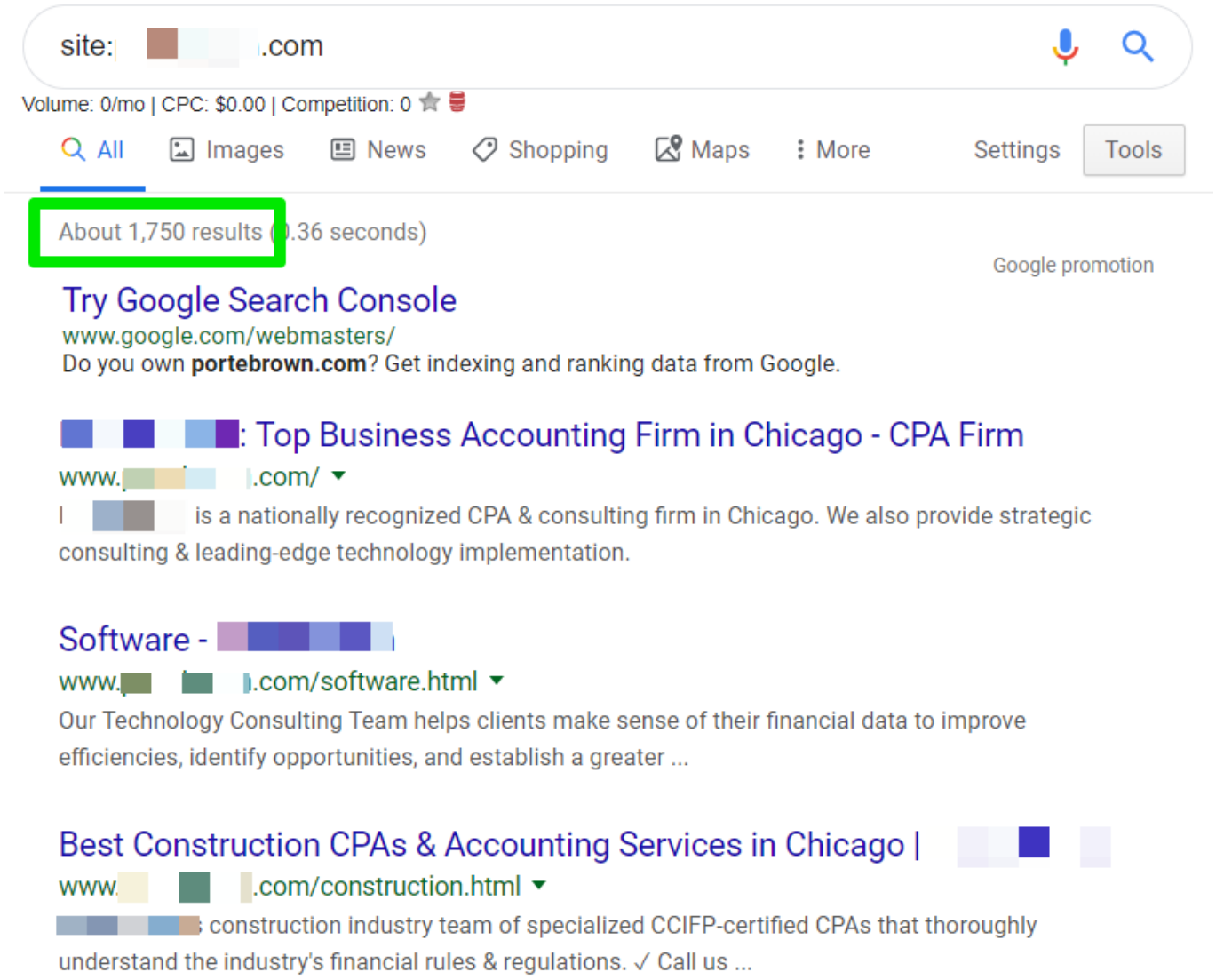
4. Your own Content Management System (CMS)

This one is a simple check too, don't overthink it. How many pages on your site do you have? How many blog posts do you have? Add them up. We're looking for quality content that provides value, but more so in a quantitative fashion. It doesn't have to be exact as the actual quality a piece of content has can be measured via a content audit.

Make a note of the number you see.

5. Google

At last, we come to the final check of our series. Sometimes Google throws a number at you and you have no idea where it comes from, but try to be as objective as possible. Do a "site:domain.com" search on Google and check how many results Google serves you from its index. Remember, this is purely a numeric value and does not truly determine the quality of your pages.



Make a note of the number you see and compare it to the other numbers you found. Any discrepancies you find indicates symptoms of an inefficient indexation. Completing a simple quantitative analysis will help direct you to areas that may not meet minimum qualitative criteria. In other words, comparing numeric values from multiple sources will help you find pages on your site that contain a low value.

The quality criteria we evaluate against can be found in Google's Webmaster guidelines.

How to resolve index bloat

Resolving index bloat is a slow and tedious process, but you have to trust the optimizations you're performing on the site

and have patience during the process, as the results may be slow to become noticeable.

1. Deleting pages (Ideal)

In an ideal scenario, low-quality pages would not exist on your site, and thus, not consume any limited resources from search engines. If you have a large number of outdated pages that you no longer use, cleaning them up (deleting) can often lead to other benefits like fewer redirects and 404s, fewer thin-content pages, less room for error and misinterpretation from search engines, to name a few.

The less control you give search engines by limiting their options on what action to take, the more control you will have on your site and your SEO.

Of course, this isn’t always realistic. So here are a few alternatives.

2. Using Noindex (Alternative)

When you use this method at the page level please don’t add a site-wide noindex – happens more often than we’d like), or within a set of pages, is probably the most efficient as it can be completed very quickly on most platforms.

- Do you use all those testimonial pages on your site?
- Do you have a proper blog tag/category in place, or are they just bloating the index?
- Does it make sense for your business to have all those blog author pages indexed?

All of the above can be noindexed and removed from your XML sitemap(s) with a few clicks on WordPress if you use Yoast SEO or All in One SEO.

3. Using Robots.txt (Alternative)

Using the robots.txt file to disallow sections or pages of your site is not recommended for most websites unless it has been explicitly recommended by an SEO Expert after auditing your website. It’s incredibly important to look at the specific environment your site is in and how a disallow of certain pages would affect the indexation of the rest of the site. Making a careless change here may result in unintended consequences.

Now that we’ve got that disclaimer out of the way, disallowing certain areas of your site means that you’re blocking search engines from even reading those pages. This means that if you added a noindex, and also disallowed, Google won’t even get to read the noindex tag on your page or follow your directive because you’ve blocked them from access. Order of operations, in this case, is absolutely crucial in order for Google to follow your directives.

4. Using Google Search Console’s manual removal tool (Temporary)

As a last resort, an action item that does not require developer resources is using the manual removal tool within the old Google Search Console. Using this method to remove pages, whole subdirectories, and entire subdomains from Google Search is only temporary. It can be done very quickly, all it takes is a few clicks. Just be careful of what you’re asking Google to deindex.

A successful removal request lasts only about 90 days, but it can be revoked manually. This option can also be done in conjunction with a noindex meta tag to get URLs out of the index as soon as possible.

Conclusion

Search engines despise thin content and try very hard to filter out all the spam on the web, hence the never-ending search quality updates that happen almost daily. In order to appease search engines and show them all the amazing content we spent so much time creating, webmasters must make sure their technical SEO is buttoned up as early in the site's lifespan as possible before index bloat becomes a nightmare.

Using the different methods described above can help you diagnose any index bloat affecting your site so you can figure out which pages need to be deleted. Doing this will help you optimize your site's overall quality evaluation in search engines, rank better, and get a cleaner index, allowing Google to find the pages you're trying to rank quickly and efficiently.

Pablo Villalpando is a Bilingual SEO Strategist for Victorious. He can be found on Twitter

Related reading



Understand your site's critical

HTTP codes to recognize errors that demand attention and find opportunities to he



lp im pr ov e yo ur SE O.

A ne w re se ar ch st ud y re ve al s th at Go og le re vi ew st ar s fo rm

social proof for your brand, boosting your site’s CTR by over 35 percent.



The powe

rf
ul
dr
iv
in
g
so
ur
ce
of
an
y
bu
si
ne
ss
is
it
s
co
mp
et
it
io
n.
Th
re
e
co
mp
et
it
iv
e
re
se
ar



chtoolsthatprovidehighlyactionableinsights.

Googleneverconfirmedits

existence, but many SEOs believe that Google Sandbox exists. So, does it exist in

2019? If yes, how to avoid getting affected?

Delete your pages and rank higher in search – Index bloat and technical optimization 2019