Quest for more coverage: Making a case for larger exports from Google Search Console's Coverage reporting

In 2018 Google revamped the Index Coverage report in Google Search Console. It was a huge improvement from the previous version, and SEOs around the globe rejoiced. With the upgrade, the new Coverage reporting greatly expanded the amount of information for site owners, including errors, indexing levels, and urls that were being excluded from indexing. Google also started providing 16 months of data, which was significant increase from what we had previously.

The update was awesome, but there was one glaring issue: the limited export capability from the Coverage reporting, which is especially tough for medium to larger-scale sites. Currently, there is a **one thousand row limit** per report.

Sure, smaller sites can often export all of their data per report, but many larger-scale sites easily blow past that limit per category.

For example, a site with just two hundred pages could easily export all indexed pages, and typically all excluded pages by category. But sites with hundreds of thousands, or millions of pages, are often severely limited by just one thousand rows per report. And that can hamper a site owner’s ability to fully identify patterns of problems across the entire site, export more URLs by category, and then of course, address all of those problems in a timely manner.

Google Search Console should provide more data per report. I’ll provide specific cases where more data would have helped

site owners, and I'll provide a way for you to get involved (so we can learn how the overall SEO community feels about the current exporting limitation in Google Search Console).

To accomplish that, I have set up a poll (which is embedded below) where you can make your voice heard. Depending on feedback from SEOs, site owners, etc., maybe we can convince Google to increase the export ability from Google Search Console's Coverage reporting.

An important conversation at the Webmaster Conference in Mountain View

In November of 2019 I was fortunate enough to attend the Google webmaster conference at Google headquarters in Mountain View. It was an awesome one-day conference where we were able to see presentations from various Googlers about what they were working on across Search. In addition, Google held one of their Product Fairs where Googlers set up kiosks showing off projects they are working on.

As soon as the fair started, I scanned the room checking out the titles at each kiosk. One immediately stood out for me… It read, “Google Search Console.” So, I ran over a few people as I rapidly approached the stand. I was determined to speak with someone about gaining more data from the new and improved Coverage reporting (or even an API).

I ended up having a great conversation with an engineer from the Google Search Console product team. I made my case for gaining more data as he listened to my points, while also countering some of what I was saying. It was extremely interesting to hear Google’s view, be challenged to provide real-word cases of why we would need more data, etc.

I explained that crawling sites via third-party tools is incredibly important and helpful, but for larger-scale sites,

it’s often not feasible to continually crawl massive sites hunting for problems. Also, just crawling a site doesn’t provide how **Google is actually treating** URLs at scale. For example, are they being categorized as “crawled, not indexed,” soft 404s, being canonicalized to other urls, etc.? You still need to inspect those urls to find out (and you can’t easily do that in bulk).

The combination of larger exports from Google Search Console’s Coverage reporting and crawling through popular tools like DeepCrawl, Screaming Frog, and Sitebulb is a much better approach. Basically, Google has the data already, so maybe it could provide more of that data so site owners can get a stronger view of those problems.

Google: Coverage is for identifying patterns versus full (or large) exports

The overwhelming reason I was hearing from the Google Search Console product team member about why the reporting doesn't provide larger exports is that the Coverage reporting is supposed to provide **patterns** and not full exports of data. I totally get that, and it does provide some patterns, but my point was that Google Search Console can't possibly surface **all patterns** for larger-scale, complex sites. The engineer working on Google Search Console was great, and heard me out, but I didn't leave that conversation feeling we would see an API any time soon… although I was hopeful we might see one in the future.

Here is part of a tweet thread I shared after the webmaster conference. This specific tweet was about a Coverage API:

> *Regarding a coverage report API, I spoke w/an engineer working on GSC. I didn't get the feeling we would see an API anytime soon. He said GSC would be more valuable for providing patterns of issues & not providing every single URL for each issue. I don't agree, but that's me. :) pic.twitter.com/2vjeiHo0it*
>
> *— Glenn Gabe (@glenngabe) November 8, 2019*

SEO nirvana – larger exports or a Coverage reporting API:

After that conversation, and continuing to work on larger-scale sites, I kept running into situations where more data more would have greatly helped my efforts (and the site owners I was helping). I couldn't stand seeing tens of thousands,

hundreds of thousands, or even millions of urls in various Coverage reports only to export just one thousand rows. Talk about underwhelming…

So what am I looking for? What would help SEOs and site owners working on larger-scale sites? There's a two-part answer to this, and to be honest, I would be happy with either one.

1) A Coverage API
This would be incredible. Just like you can use the Search Console API now to bulk export data from the Performance reporting, you would be able to do that for the Coverage reporting (and by category).

For example:

- Need to export all URLs categorized as "crawled, not indexed"? Boom, you got it.
- How about exporting all Soft 404s? No problem. Beep, bop, boop. They're exported.
- Is Google ignoring rel=canonical across many URLs and choosing a different canonical? Pfft… they're exported.
- Want to export all URLs indexed in a specific directory? YOUR WISH IS MY COMMAND!

I get it, this is a big ask of the Google Search Console product team, but it would be incredible. And if an API isn't easily doable, then there's a second option.

2) Larger Exports (e.g. 100K per report)
Although this would still be limiting for many larger-scale

sites, 100K rows of data per report would be much, much better than just one thousand rows. If sites could export 100K rows per report in the Coverage reporting, site owners could surface more patterns based on seeing more URLs that are being categorized as a certain issue. I think I speak for most people working on larger-scale sites when I say this would be an amazing upgrade in Google Search Console.

Who wouldn’t love seeing this in Google Search Console?

Case studies: real-word scenarios

of when more data would have greatly helped site owners

To help demonstrate the limitations that larger-scale sites run into, I decided to provide a few cases below that I have worked on personally. They are meant to underscore the point that one thousand rows per report is often not enough to surface all of the patterns necessary on a large and complex site. There may be problems located all over the site, while the reporting might only show a handful of those areas. And again, just crawling sites on your own doesn't give you the full picture (how Google is actually handling those URLs). Inspecting the URLs in Google Search Console is required and you cannot do that in bulk.

After covering the cases, I provided a poll where you can make your voice heard. I would love to hear what the larger SEO community thinks about gaining more data from the Coverage reporting. Thank you in advance for your participation!

Indexed URLs (both submitted and NOT submitted)

- The most obvious benefit here is to understand all of the pages that are currently indexed (and within certain directories), which can be challenging for larger sites. This can help site owners understand which specific pages are indexed, in which sections of the site, etc. There is currently NO WAY to truly understand every page that's indexed on a site or in a directory.
- Being able to see indexing levels **across page types or site sections** can help site owners understand potential problems with those page types or areas of the site. Having a stronger view of indexing levels would help on this front.
- As a stopgap, I typically recommend site owners set up

every major directory as a property in GSC to gain more visibility into indexing levels. That works ok, but it's still not great for larger-scale sites.

- For example, here is **one directory** from a larger-scale site that was struggling with mass duplicate content problems. It would be great to see all URLs that were being indexed in this section so their dev team could easily dig in:

- And beyond just the total, it would be amazing to export all URLs that are categorized as "indexed, but **not** submitted in sitemap." Those can often be URLs that sites don't know are indexable, which could be why they aren't included in xml sitemaps. Being able to bulk export that category would be amazing.
- Also, this would help site owners figure out which canonical URLs on a large-scale site are **not** being submitted in sitemaps and then make sure they are. So, a larger export could help them track down problems, while also helping them submit the right URLs in xml sitemaps.
- For example, here's a site I'm helping with 59% of its indexed pages as "indexed, not submitted in sitemap."

“Crawled, not indexed” and “Discovered, not indexed”

- These categories can often signal quality problems and/or crawl budget issues. It’s super important for site owners to dig into these categories and understand what’s there.
- By the way, crawling your own sites will **not** help here since you won’t know if pages are categorized as “Crawled, not indexed,” or “Discovered, not indexed.” You can test them in Google Search Console by URL using the URL inspection tool, but that won’t work in bulk.
- When digging into these two categories, I often find low-quality or thin content there, or pockets of a site that Google just isn’t liking for some reason. The ability to fully export this data would be incredible.
- As a recent example, a client had 256K URLs categorized as “Crawled, not indexed” and they spanned several categories across the site. The reporting did **not** cover

all patterns from the site. It would be extremely helpful to export all of those URLs so my client's dev team and content team could go through them to identify key problems (and then build a plan for fixing them).
- And that same client has over 1M urls categorized as "Discovered, not indexed." Again, it would be great to see those URLs (or at least more of them), and across page types, in order to better analyze the situation.

Canonicalization problems

- There are several categories in Google Search Console that flag when Google is choosing a different canonical URL. This is extremely important for site owners to understand. For example, "Duplicate, Google chose different canonical than user" and "Duplicate, submitted url not selected as canonical." The ability to export all of the URLs (or more of them) and analyze each situation to surface patterns would be great. For now, you can only see one thousand URLs and there's no way that can fully represent some larger-scale sites that might have millions of pages indexed (or more).
- Since rel=canonical is a hint, Google can choose to

ignore it and index URLs anyway. And when that happens, many URLs can end up getting indexed that shouldn't, but site owners often have no idea that's going on. And that can lead to problems (like many lower quality pages getting indexed when site owners thought they were being canonicalized). Being able to export all URLs that fall into these two categories would be helpful for understanding how Google is handling certain situations.

- As an example, I was auditing a client's site containing a lot of reviews and they were canonicalizing many pages across the site, but to pages with **different content**. They automatically assumed those pages would not be indexed. Since Google views rel=canonical as a **hint**, it ended up ignoring rel=canonical across many of those pages. The problem was that the pages were thin and lower quality and many were being indexed. It's a large site and not all of the URLs in that page type were being handled this way. So, being able to export a full list would have been incredible.

- Side note: Here is a post I wrote about this situation (Google ignoring rel=canonical resulting in many pages getting indexed when they shouldn’t be).

Soft 404s

- I know Google Search Console has improved this category recently, which is great, but it’s still important for site owners to understand where soft 404s reside. For larger and more complex sites, those can be located all over a site, and in different categories. The ability to export this data in bulk can help them hunt down those issues and handle accordingly.
- This can also help sites that have products out of stock or that have expired deals or campaigns. For larger sites, they may have no idea those pages are still on the site returning 200 codes, eating up crawl budget, etc.
- I have also found situations where only some of those pages are being treated as soft 404s by Google, while many others are actually being indexed. And those pages have no shot at meeting or exceed user expectations.
- For example, a client had over 261K soft 404s until the recent Google Search Console changes rolled out in early January. It would have been helpful to be able to export those soft 404s. It’s a huge site, with several feeds of data being used, and it’s often hard for the dev team to track down pages like this.

Join me on the quest for more coverage

The one thousand row limit in Google Search Console's coverage reporting can be extremely frustrating to deal with for larger sites. By providing an API or even larger exports, site owners would be able to track down more patterns from across their sites. And that could quicken up the process of rectifying those problems.

If you manage a larger site, you might have also experienced this frustration. And that's why I want to hear from the larger SEO community about this. Below, I've embedded a poll so you can make your voice heard. It won't take long to fill out, and maybe, just maybe, we can share the results with Google while making a case for receiving larger exports from Google Search Console's Coverage reporting. And if the overall SEO community doesn't feel there's a need for larger exports from Google Search Console's Coverage reporting, I'll accept that, move on, and stare at one thousand rows in Google Search Console for the foreseeable future.

Thank you in advance for your participation!

Opinions expressed in this article are those of the guest author and not necessarily Search Engine Land. Staff authors are listed here.

About The Author

Glenn Gabe is an SEO consultant at G-Squared Interactive (GSQi) and has over 24 years of experience in digital marketing. Glenn currently helps clients with algorithm update recovery, technical SEO audits, SEO for website redesigns and migrations, and SEO training. During his career, Glenn has helped companies across a wide range of industries including publishing, e-commerce, health and medical, finance, startups, military, education, non-profits, real estate, affiliate marketing, and more.

http://feeds.searchengineland.com/~r/searchengineland/~3/ftx6lovMNBw/quest-for-more-coverage-making-a-case-for-larger-exports-from-google-search-consoles-coverage-reporting-345937