Crawled – Currently Not Indexed: A Coverage Status Guide

Google's Index Coverage report is absolutely fantastic because it gives SEOs clearer insights into Google's crawling and indexing decisions. Since its roll-out, we use it almost daily at Go Fish Digital to diagnose technical issues at scale for our clients.

Within the report, there are many different "statuses" that provide webmasters with information about how Google is handling their site content. While many of the statuses provide some context around Google's crawling and indexation decisions, one remains unclear: "Crawled – currently not indexed".

Since seeing the "Crawled – currently not indexed" status reported, we've heard from several site owners inquiring about its meaning. One of the benefits of working at an agency is being able to get in front of a lot of data, and because we've seen this message across multiple accounts, we've begun to pick up on trends from reported URLs.

Google's definition

Let's start with the official definition. According to Google's official documentation, this status means: "The page was crawled by Google, but not indexed. It may or may not be indexed in the future; no need to resubmit this URL for crawling."

So, essentially what we know is that:

1. Google is able to access the page

2. Google took time to crawl the page
3. After crawling, Google decided not to include it in the index

The key to understanding this status is to think of reasons why Google would “consciously” decide against indexation. We know that Google isn’t having trouble finding the page, but for some reason it feels users wouldn’t benefit from finding it.

This can be quite frustrating, as you might not know why your content isn’t getting indexed. Below I’ll detail some of the most common reasons our team has seen to explain why this mysterious status might be affecting your website.

1. False positives

Priority: Low

Our first step is to always perform a few spot checks of URLs flagged in the “Crawled – currently not indexed” section for indexation. It’s not uncommon to find URLs that are getting reported as excluded but turn out to be in Google’s index after all.

For example, here’s a URL that’s getting flagged in the report for our website: https://gofishdigital.com/meetup/

However, when using a site search operator, we can see that the URL is actually included in Google’s index. You can do this by appending the text “site:” before the URL.

If you’re seeing URLs reported under this status, I recommend starting by using the site search operator to determine whether the URL is indexed or not. Sometimes, these turn out to be false positives.

Solution: Do nothing! You’re good.

2. RSS feed URLs

Priority: Low

This is one of the most common examples that we see. If your site utilizes an RSS feed, you might be finding URLs appearing in Google’s “Crawled – currently not indexed” report. Many times these URLs will have the “/feed/” string appended to the end. They can appear in the report like this:

Google finding these RSS feed URLs linked from the primary page. They'll often be linked to using a "rel=alternate" element. WordPress plugins such as Yoast can automatically generate these URLs.

Solution: Do nothing! You're good.

Google is likely selectively choosing not to index these URLs, and for good reason. If you navigate to an RSS feed URL, you'll see an XML document like the one below:

While this XML document is useful for RSS feeds, there’s no need for Google to include it in the index. This would provide a very poor experience as the content is not meant for users.

3. Paginated URLs

Priority: Low

Another extremely common reason for the “Crawled – currently not indexed” exclusion is pagination. We will often see a good number of paginated URLs appear in this report. Here we can see some paginated URLs appearing from a very large e-commerce site:

Solution: Do nothing! You’re good.

Google will need to crawl through paginated URLs to get a complete crawl of the site. This is its pathway to content such as deeper category pages or product description pages. However, while Google uses the pagination as a pathway to access the content, it doesn’t necessarily need to index the paginated URLs themselves.

If anything, make sure that you don’t do anything to impact the crawling of the individual pagination. Ensure that all of your pagination contains a self-referential canonical tag and is free of any “nofollow” tags. This pagination acts as an avenue for Google to crawl other key pages on your site so you’ll definitely want Google to continue crawling it.

4. Expired products

Priority: Medium

When spot-checking individual pages that are listed in the report, a common problem we see across clients is URLs that contain text noting “expired” or “out of stock” products. Especially on e-commerce sites, it appears that Google checks to see the availability of a particular product. If it determines that a product is not available, it proceeds to exclude that product from the index.

This makes sense from a UX perspective as Google might not want to include content in the index that users aren’t able to purchase.

However, if these products are actually available on your site, this could result in a lot of missed SEO opportunity. By excluding the pages from the index, your content isn't given a chance to rank at all.

In addition, Google doesn't just check the visible content on the page. There have been instances where we've found no indication within the visible content that the product is not available. However, when checking the structured data, we can see that the "availability" property is set to "OutOfStock".

It appears that Google is taking clues from both the visible content and structured data about a particular product's availability. Thus, it's important that you check both the content and schema.

Solution: Check your inventory availability.

If you're finding products that are actually available getting listed in this report, you'll want to check all of your products that may be incorrectly listed as unavailable. Perform a crawl of your site and use a custom extraction tool like Screaming Frog's to scrape data from your product pages.

For instance, if you want to see at scale all of your URLs with schema set to "OutOfStock", you can set the "Regex" to: "availability":"<="" p="">

This: <="" p="">"class="redactor-autoparser-object">http://schema.org/OutOfStock" should automatically scrape all of the URLs with this property:

You can export this list and cross-reference with inventory data using Excel or business intelligence tools. This should quickly allow you to find discrepancies between the structured data on your site and products that are actually available. The same process can be repeated to find instances where your visible content indicates that products are expired.

5. 301 redirects

Priority: Medium

One interesting example we've seen appear under this status is destination URLs of redirected pages. Often, we'll see that Google is crawling the destination URL but not including it in the index. However, upon looking at the SERP, we find that Google is indexing a redirecting URL. Since the redirecting URL is the one indexed, the destination URL is thrown into the "Crawled – currently not indexed" report.

The issue here is that Google may not be recognizing the redirect yet. As a result, it sees the destination URL as a “duplicate” because it is still indexing the redirecting URL.

Solution: Create a temporary sitemap.xml.

If this is occurring on a large number of URLs, it is worth taking steps to send stronger consolidation signals to Google. This issue could indicate that Google isn’t recognizing your redirects in a timely manner, leading to unconsolidated content signals.

One option might be setting up a “temporary sitemap”. This is a sitemap that you can create to expedite the crawling of these redirected URLs. This is a strategy that John Mueller has previously recommended.

To create one, you will need to reverse-engineer redirects that you have created in the past:

1. Export all of the URLs from the “Crawled – currently not indexed” report.
2. Match them up in Excel with redirects that have been previously set up.
3. Find all of the redirects that have a destination URL in the “Crawled – currently not indexed” bucket.
4. Create a static sitemap.xml of these URLs with Screaming Frog.
5. Upload the sitemap and monitor the “Crawled – currently not indexed” report in Search Console.

The goal here is for Google to crawl the URLs in the temporary sitemap.xml more frequently than it otherwise would have. This will lead to faster consolidation of these redirects.

6. Thin content

Priority: Medium

Sometimes we see URLs included in this report that are extremely thin on content. These pages may have all of the technical elements set up correctly and may even be properly internally linked to, however, when Google runs into these URLs, there is very little actual content on the page. Below is an example of a product category page where there is very little unique text:

This product listing page was flagged as “Crawled – Currently Not Indexed”. This may be due to very thin content on the page.
This page is likely either too thin for Google to think it’s useful or there is so little content that Google considers it to be a duplicate of another page. The result is Google removing the content from the index.

Here is another example: Google was able to crawl a testimonial component page on the Go Fish Digital site (shown above). While this content is unique to our site, Google probably doesn't believe that the single sentence testimonial should stand alone as an indexable page.

Once again, Google has made the executive decision to exclude the page from the index due to a lack of quality.

Solution: Add more content or adjust indexation signals.

Next steps will depend on how important it is for you to index these pages.

If you believe that the page should definitely be included in the index, consider adding additional content. This will help Google see the page as providing a better experience to users.

If indexation is unnecessary for the content you're finding, the bigger question becomes whether or not you should take the additional steps to strongly signal that this content *shouldn't* be indexed. The "Crawled –currently not indexed" report is indicating that the content is eligible to appear in Google's index, but Google is electing not to include it.

There also could be other low quality pages to which Google is not applying this logic. You can perform a general “site:” search to find indexed content that meets the same criteria as the examples above. If you’re finding that a large number of these pages are appearing in the index, you might want to consider stronger initiatives to ensure these pages are removed from the index such as a “noindex” tag, 404 error, or removing them from your internal linking structure completely.

7. Duplicate content

Priority: High

When evaluating this exclusion across a large number of clients, this is the highest priority we’ve seen. If Google sees your content as duplicate, it may crawl the content but elect not to include it in the index. This is one of the ways that Google avoids SERP duplication. By removing duplicate content from the index, Google ensures that users have a larger variety of unique pages to interact with. Sometimes the report will label these URLs with a “Duplicate” status (“Duplicate, Google chose different canonical than user”). However, this is not always the case.

This is a high priority issue, especially on a lot of e-commerce sites. Key pages such as product description pages often include the same or similar product descriptions as many other results across the Web. If Google recognizes these as too similar to other pages internally or externally, it might exclude them from the index all together.

Solution: Add unique elements to the duplicate content.

If you think that this situation applies to your site, here’s how you test for it:

1. Take a snippet of the potential duplicate text and paste it into Google.
2. In the SERP URL, append the following string to the end: “&num=100”. This will show you the top 100 results.
3. Use your browser’s “Find” function to see if your result appears in the top 100 results. If it doesn’t, your result might be getting filtered out of the index.
4. Go back to the SERP URL and append the following string to the end: “&filter=0”. This should show you Google’s unfiltered result (thanks, Patrick Stox, for the tip).
5. Use the “Find” function to search for your URL. If you see your page now appearing, this is a good indication that your content is getting filtered out of the index.
6. Repeat this process for a few URLs with potential duplicate or very similar content you’re seeing in the “Crawled – currently not indexed” report.

If you’re consistently seeing your URLs getting filtered out of the index, you’ll need to take steps to make your content more unique.

While there is no one-size-fits-all standard for achieving this, here are some options:

1. Rewrite the content to be more unique on high-priority pages.

2. Use dynamic properties to automatically inject unique content onto the page.
3. Remove large amounts of unnecessary boilerplate content. Pages with more templated text than unique text might be getting read as duplicate.
4. If your site is dependent on user-generated content, inform contributors that all provided content should be unique. This may help prevent instances where contributors use the same content across multiple pages or domains.

8. Private-facing content

Priority: High

There are some instances where Google's crawlers gain access to content that they shouldn't have access to. If Google is finding dev environments, it could include those URLs in this report. We've even seen examples of Google crawling a particular client's subdomain that is set up for JIRA tickets. This caused an explosive crawl of the site, which focused on URLs that shouldn't ever be considered for indexation.

The issue here is that Google's crawl of the site isn't focused, and it's spending time crawling (and potentially indexing) URLs that aren't meant for searchers. This can have massive ramifications for a site's crawl budget.

Solution: Adjust your crawling and indexing initiatives.

This solution is going to be entirely dependent on the situation and what Google is able to access. Typically, the first thing you want to do is determine how Google is able to discover these private-facing URLs, especially if it's via your internal linking structure.

Start a crawl from the home page of your primary subdomain and see if any undesirable subdomains are able to be accessed by Screaming Frog through a standard crawl. If so, it's safe to say that Googlebot might be finding those exact same pathways. You'll want to remove any internal links to this content to cut Google's access.

The next step is to check the indexation status of the URLs that should be excluded. Is Google sufficiently keeping all of them out of the index, or were some caught in the index? If Google isn't indexing a large amount of this content, you might consider adjusting your robots.txt file to block crawling immediately. If not, "noindex" tags, canonicals, and password protected pages are all on the table.

Case study: duplicate user-generated content

For a real-world example, this is an instance where we diagnosed the issue on a client site. This client is similar to an e-commerce site as a lot of their content is made up of product description pages. However, these product description pages are all user-generated content.

Essentially, third parties are allowed to create listings on this site. However, the third parties were often adding very short descriptions to their pages, resulting in thin content. The issue occurring frequently was that these user-generated product description pages were getting caught in the "Crawled – currently not indexed" report. This resulted in missed SEO opportunity as pages that were capable of generating organic traffic were completely excluded from the index.

When going through the process above, we found that the client's product description pages were quite thin in terms of unique content. The pages that were getting excluded only appeared to have a paragraph or less of unique text. In

addition, the bulk of on-page content was templated text that existed across all of these page types. Since there was very little unique content on the page, the templated content might have caused Google to view these pages as duplicates. The result was that Google excluded these pages from the index, citing the “Crawled – currently not indexed” status.

To solve for these issues, we worked with the client to determine which of the templated content didn’t need to exist on each product description page. We were able to remove the unnecessary templated content from thousands of URLs. This resulted in a significant decrease in “Crawled – currently not indexed” pages as Google began to see each page as more unique.

Conclusion

Hopefully, this helps search marketers better understand the mysterious “Crawled – currently not indexed” status in the Index Coverage report. Of course, there are likely many other reasons that Google would choose to categorize URLs like this, but these are the most common instances we’ve seen with our clients to date.

Overall, the Index Coverage report is one of the most powerful tools in Search Console. I would highly encourage search marketers to get familiar with the data and reports as we

routinely find suboptimal crawling and indexing behavior, especially on larger sites. If you've seen other examples of URLs in the "Crawled – currently not indexed" report, let me know in the comments!

http://tracking.feedpress.it/link/9375/13331422