

Google Responds To Site That Lost Ranks After Googlebot DDoS Crawl via @sejournal, @martinibuster

Google's John Mueller answered a question about a site that received millions of Googlebot requests for pages that don't exist, with one non-existent URL receiving over two million hits, essentially DDoS-level page requests. The publisher's concerns about crawl budget and rankings seemingly were realized, as the site subsequently experienced a drop in search visibility.

NoIndex Pages Removed And Converted To 410

The 410 Gone server response code belongs to the family 400 response codes that indicate a page is not available. The 404 response means that a page is not available and makes no claims as to whether the URL will return in the future, it simply says the page is not available.

The 410 Gone status code means that the page is gone and likely will never return. Unlike the 404 status code, the 410 signals the browser or crawler that the missing status of the resource is intentional and that any links to the resource should be removed.

The person asking the question was following up on a question they posted three weeks ago on Reddit where they noted that they had about 11 million URLs that should not have been

discoverable that they removed entirely and began serving a 410 response code. After a month and a half Googlebot continued to return looking for the missing pages. They shared their concern about crawl budget and subsequent impacts to their rankings as a result.

Mueller at the time forwarded them to a Google support page.

Rankings Loss As Google Continues To Hit Site At DDOS Levels

Three weeks later things have not improved and they posted a follow-up question noting they've received over five millions requests for pages that don't exist. They posted an actual URL in their question but I anonymized it, otherwise it's verbatim.

The person asked:

"Googlebot continues to aggressively crawl a single URL (with query strings), even though it's been returning a 410 (Gone) status for about two months now.

In just the past 30 days, we've seen approximately 5.4 million requests from Googlebot. Of those, around 2.4 million were directed at this one URL:

<https://example.net/software/virtual-dj/> with the ?feature query string.

We've also seen a significant drop in our visibility on Google during this period, and I can't help but wonder if there's a connection – something just feels off. The affected page is:

<https://example.net/software/virtual-dj/?feature=...>

The reason Google discovered all these URLs in the first place is that we unintentionally exposed them in a JSON payload generated by Next.js – they weren't actual links on

the site.

We have changed how our “multiple features” works (using ?mf querystring and that querystring is in robots.txt)

Would it be problematic to add something like this to our robots.txt?

*Disallow: /software/virtual-dj/?feature=**

Main goal: to stop this excessive crawling from flooding our logs and potentially triggering unintended side effects.”

Google’s John Mueller confirmed that it’s Google’s normal behavior to keep returning to check if a page that is missing has returned. This is Google’s default behavior based on the experience that publishers can make mistakes and so they will periodically return to verify whether the page has been restored. This is meant to be a helpful feature for publishers who might unintentionally remove a web page.

Mueller responded:

“Google attempts to recrawl pages that once existed for a really long time, and if you have a lot of them, you’ll probably see more of them. This isn’t a problem – it’s fine to have pages be gone, even if it’s tons of them. That said, disallowing crawling with robots.txt is also fine, if the requests annoy you.”

Caution: Technical SEO Ahead

This next part is where the SEO gets technical. Mueller cautions that the proposed solution of adding a robots.txt could inadvertently break rendering for pages that aren’t supposed to be missing.

He’s basically advising the person asking the question to:

- Double-check that the ?feature= URLs are not being used at all in any frontend code or JSON payloads that power important pages.
- Use Chrome DevTools to simulate what happens if those URLs are blocked – to catch breakage early.
- Monitor Search Console for Soft 404s to spot any unintended impact on pages that should be indexed.

John Mueller continued:

“The main thing I’d watch out for is that these are really all returning 404/410, and not that some of them are used by something like JavaScript on pages that you want to have indexed (since you mentioned JSON payload).”

It’s really hard to recognize when you’re disallowing crawling of an embedded resource (be it directly embedded in the page, or loaded on demand) – sometimes the page that references it stops rendering and can’t be indexed at all.

If you have JavaScript client-side-rendered pages, I’d try to find out where the URLs used to be referenced (if you can) and block the URLs in Chrome dev tools to see what happens when you load the page.

If you can’t figure out where they were, I’d disallow a part of them, and monitor the Soft-404 errors in Search Console to see if anything visibly happens there.

If you’re not using JavaScript client-side-rendering, you can probably ignore this paragraph :-).”

The Difference Between The Obvious Reason And The Actual Cause

Google’s John Mueller is right to suggest a deeper diagnostic to rule out errors on the part of the publisher. A publisher

error started the chain of events that led to the indexing of pages against the publisher's wishes. So it's reasonable to ask the publisher to check if there may be a more plausible reason to account for a loss of search visibility. This is a classic situation where an obvious reason is not necessarily the correct reason. There's a difference between being an obvious reason and being the actual cause. So Mueller's suggestion to not give up on finding the cause is good advice.

Read the original discussion here.

Featured Image by Shutterstock/PlutusART

<https://www.searchenginejournal.com/google-responds-to-site-that-lost-ranks-after-googlebot-ddos-crawl/548692/>