http://www.coincoin.fr.eu.org/?Googlebot-is-now-aggressively

Googlebot is now aggressively crawling syndication feeds

- 6- Webographie -

Date de mise en ligne : mercredi 5 mars 2014

Copyright © L'Imp'Rock Scénette (by @_daffyduke_) - Tous droits réservés

I'm not sure how long this has been going on (I only noticed it recently) but Googlebot, Google's search crawler, is now aggressively crawling syndication feeds. By 'aggressively crawling' I mean two things. First, it is fetching the feeds multiple times a day; one of my feeds was fetched 46 times in one 24-hour period. Second and worse, it's not using conditional GET.

I've written before about why web spiders should not crawl syndication feeds and I still believe everything I wrote back then (even though I've significantly reduced the number of feeds I advertise since those days). My feed URLs are all marked 'nofollow', a declaration that Googlebot generally respects. And even if Google was going to crawl syndication feeds, the minimum standard is implementing conditional GET instead of repeatedly spamming fetch requests; the latter is the kind of thing that gets you banned here.

I might very reluctantly accept Googlebot crawling a few syndication feed URLs if they properly implemented conditional GET. Then it might be a reasonable move to find updated content (although Googlebot accesses my sitemap much less frequently) and I'd passively go along with the 800 pound gorilla of search traffic. But without conditional GET it's my strong opinion that this is abuse plain and simple, and I have no interest in cooperation.

So, in short: I suggest that you check your syndication feed logs to see if Googlebot is pounding on them too and if it is, block it from accessing those URLs. I doubt Google is going to change its behavior any time soon or even notice, but at least you can avoid donating your site resources to an abusive crawler.

(As I expected, Googlebot is paying absolutely no attention to days of 403 responses on the feed URLs it's trying to fetch. It keeps on trying to fetch them at great volume, to the tune of 245 requests so far today for 11 different URLs.)

Sidebar: Some more details

First, this really is Googlebot; it comes from Google IP address ranges and from specific IPs with crawl-*.googlebot.com reverse DNS such as 66.249.66.130.

Second, in the past Googlebot has shown signs of supporting conditional GET on syndication feeds. I have historical logs that show Googlebot getting 304's on syndication feed URLs.

Third, based on historical logs I have for my personal website, this appears to have started happening there around January 13th. There are sporadic requests for feed URLs before then, but January 13th is when things light up with multiple requests a day.

(2 comments.)

Cet article est repris du site http://utcc.utoronto.ca/~cks/space/...