Google Can Detect Duplicate Content Before Crawling It by @MattGSouthern

Google's John Mueller revealed in a Webmaster Central hangout this week that Googlebot is capable of recognizing duplicate content before it has been crawled.

A question was submitted by a site owner wondering if and when Google would consider a French version of a page to be a duplicate of the English version.

Can Google determine when multiple pages have the same content in different languages? If so, how is that handled in search results?

In Mueller's response he revealed that, in some instances, Google can detect when pages share the same content without even having to crawl the pages. This is something worth being aware of, especially when it comes to the URL structure of pages.

> *"What sometimes happens is we kind of proactively recognize that something is probably a duplicate, even before crawling it. So this happens when we see that the difference, for example, is within the URL somewhere in a place where we've generally noticed that the content shown in this part of the URL is not so relevant to the content that's shown on the page.*
>
> *So that could could be something like you have a language parameter that you can set to any kind of term, and we might've gone through and tried something like*

> *“language=English,” “language=French,” “language=German,” … if we find that all these pages show the English content, except for maybe “language=Spanish” that chose the Spanish version, then we might assume that this language parameter is actually irrelevant to this page, and then we might miss that one page that actually has unique content.”*

Let’s unpack this and look at it from a broader perspective. Forget languages for a second. This particular example dealt with languages, but what Mueller had to say can apply to content of the same language as well.

What Mueller is saying here is Google may determine a page has duplicate content if it shares similar URL parameters with pages that are no different from each other.

Obviously this is not an ideal situation, as there may be instances where there are pages with unique content that have similar URL parameters as pages that are exact duplicates.

Site owners can avoid running into the problem of having unique content dismissed as duplicate by paying attention to how URL parameters are generated by their site.

Mueller admits that it may not always be the webmaster’s fault when pages are treated as duplicates– sometimes Google as its own “bugs” as well.

The original question, along with Mueller’s response, can be seen in the video below starting at the 27:38 mark.

[embedded content]

http://tracking.feedpress.it/link/13962/8495283