Search Off the Record - 64th episode

[00:00:00] ♪ [music] ♪

[00:00:10] **John Mueller:** [00:00:10] Hello, and welcome to another episode of Search Off the Record, a podcast coming to you from the Google Search team, discussing all things Search and maybe having some fun along the way, we'll see. My name is John and I'm joined today by Gary, who's also from the Search Relations team of which I'm also a part of.

[00:00:31] Today, we're going to be talking about quality and some of the effects on the rest of Search. What do you think, Gary?

[00:00:40] **Gary Illyes:** [00:00:40] Do you remember the good old days when we... it was just the two of us and we were sitting in the old building here in Google Zurich, or of Google Zurich, and it was just the two of us and it was always awkward because we were wearing our headphones and just doing very, very important work?

[00:01:00] **John Mueller:** [00:01:00] High quality.

[00:01:01] **Gary Illyes:** [00:01:01] Like playing Dig Dug.

[00:01:04] **John Mueller:** [00:01:04] Playing Dig Dug. Yeah, that was fun. Yeah, good times. But we had our own little cube, which was kind of good.

[00:01:12] **Gary Illyes:** [00:01:12] That was fantastic, I want that cube back.

[00:01:15] **John Mueller:** [00:01:15] [laughs] It was. But it was also a weird time because you had to go to the office everyday. Such weird expectations!

[00:01:24] **Gary Illyes:** [00:01:24] That was good quality time. So you said something about quality.

[00:01:28] **John Mueller:** [00:01:28] Yeah.

[00:01:28] **Gary Illyes:** [00:01:28] What was that?

[00:01:29] **John Mueller:** [00:01:29] Yeah, so I was kind of wondering, what are the general effects of quality on Search? We talk about ranking, but where else could play a role?

[00:01:39] **Gary Illyes:** [00:01:39] Quality is a big topic. It's a very big and complex topic. It affects pretty much everything that the Search systems do. I don't know how are we going to cover it in 20 minutes or 25 minutes or whatever, but we can try.

[00:01:56] **John Mueller:** [00:01:56] Okay.

[00:01:57] **Gary Illyes:** [00:01:57] What was the question by the way?

[00:01:58] **John Mueller:** [00:01:58] What is the effect of quality on Search? We talk about ranking and everyone thinks like, "Oh, high quality means ranking, but what else does it play a role in?"

[00:02:08] **Gary Illyes:** [00:02:08] [sighs] Yeah, everything, quite literally. From Sitemaps to ranking everything.

[00:02:14] **John Mueller:** [00:02:14] From Sitemaps, okay. So crawling as well, or?

[00:02:17] **Gary Illyes:** [00:02:17] Yeah. For scheduling, yep. Indexing, yep. Index selection, yep. Ranking, yep. So yeah, it's a very, very big topic, and of course, different systems are affected differently but... or they are using that quality differently but... using that quality differently but still, they are using it and it's affecting it.

[00:02:38] **John Mueller:** [00:02:38] Okay. But how does quality play a role with crawling for example? Is it... I don't know, "We're going to crawl a quality page first?"

[00:02:50] **Gary Illyes:** [00:02:50] Let's start with a new site, potatopeelers.com.

[00:02:53] [both chuckle]

[00:02:56] **Gary Illyes:** [00:02:56] Let's say that it's a brand new site. It it only has the homepage so far, and it has some content on it. It's not gibberish. It's not the highest quality but it's content. And usually, we give the benefit of doubt to homepages, and we are generally eager to crawl it without any biases, I guess, or predictions or whatever.

[00:03:25] So we go to the homepage and then we troll it. We probably start indexing it and unless it's utter crap, we are going to most likely index it and then serve it at least for the domain query.

[00:03:43] And then, if potatopeelers.com publishes yet another page, let's say that we discovery through Sitemaps because people don't know about it yet. Or we go back, revisit the homepage and then we see a link go to that page, then we are going to look back at how was the quality of the homepage, or the content of the homepage, and make a prediction about whether we want to crawl that new page that we just discovered or not.

[00:04:17] **John Mueller:** [00:04:17] So, by making a prediction, what do you mean? Googlebot is going to toss a coin or...

[00:04:24] **Gary Illyes:** [00:04:24] Well, it's not Googlebot. I think this is a big misconception that Googlebot is also smart, but it's not. Basically, Googlebot is pretty much just a Wget or a Curl instance running into cloud. It's fetching things from the Internet, but it's not doing much. But before that, there's a system called a Crawl Scheduler which we talked about it in a very, very early episode, I think episode five or something, that tries to make predictions about what to crawl and when from the Internet, from whole Internet. And that will also make predictions based on the site.

[00:04:59] **John Mueller:** [00:04:59] Okay.

[00:05:00] **Gary Illyes:** [00:05:00] So basically, if we know that we have one thousand URLs that we have to crawl from a site, then it will try to build an ordered list of URLs that we should crawl. And the order will be the priority of the crawl, like...

[00:05:19] **John Mueller:** [00:05:19] Okay.

[00:05:19] **Gary Illyes:** [00:05:19] ... start crawling from the top, because we think that that's the most important thing based on the quality that we saw on previous pages that we crawled.

[00:05:30] **John Mueller:** [00:05:30] Mmm-Hmm.

[00:05:31] **Gary Illyes:** [00:05:31] And then the further down you go on that list, the lower the priority is, either because it's not changing often, the content is not changing often, or because the content is less high quality.

[00:05:48] **John Mueller:** [00:05:48] [laughs] Lower quality.

[00:05:49] **Gary Illyes:** [00:05:49] Oh, thank you. Thank you for fixing my England. But yeah, crawl scheduling, that's where things play a role.

[00:05:59] **John Mueller:** [00:05:59] I have an interesting anecdote, or at least I find it interesting.

[00:06:03] **Gary Illyes:** [00:06:03] Okay.

[00:06:03] **John Mueller:** [00:06:03] Back before I joined Google, I would create test sites to try things out.

[00:06:09] **Gary Illyes:** [00:06:09] Sure, same.

[00:06:10] **John Mueller:** [00:06:10] Because with test sites, you kind of see what is happening. And I made one site where I added, I don't know, a couple hundred links to new pages on there. And when Google, Googlebot, Google whatever, all of these Google systems back then, it was one big thing, or at least to me, when Google discovered all of these links, it crawled them in alphabetical order.

[00:06:38] **Gary Illyes:** [00:06:38] What?

[00:06:38] **John Mueller:** [00:06:38] So they were randomly on the page but they were crawled in alphabetical order. You could look in the log file and it'd be like...

[00:06:45] [both laugh]

[00:06:45] **John Mueller:** [00:06:45] Everything with A, B, C and D. I imagine that's not the case anymore, but I thought that was pretty fun.

[00:06:53] **Gary Illyes:** [00:06:53] [laughs] I guess so. I mean back then, the Internet was different and we keep going back to how different the Internet was before we joined. That was like 15 years ago or whatever. And yeah, it was different, and you had to crawl the Internet differently. So I imagine that alphabetical order might have worked just fine, because instead of having trillions of URLs on the Internet, you had about seven.

[00:07:22] **John Mueller:** [00:07:22] A couple more maybe, yeah.

[00:07:24] **Gary Illyes:** [00:07:24] Maybe more than seven, yeah. That's fun. I don't think we are using the alphabets to organize our crawl tables anymore.

[00:07:37] **John Mueller:** [00:07:37] [laughs] I mean, maybe it was also just like, "Oh, all of these links are equally junk." It's like, "We will just use one random order that we come up with."

[00:07:47] **Gary Illyes:** [00:07:47] Equally junk? What were you publishing, John?

[00:07:49] **John Mueller:** [00:07:49] It's like test pages, you know, to try things out.

[00:07:53] **Gary Illyes:** [00:07:53] Sure, with AdSense on them, right?

[00:07:56] **John Mueller:** [00:07:56] I don't know, it's been a long time.

[00:07:58] **Gary Illyes:** [00:07:58] Admit it!

[00:08:00] **John Mueller:** [00:08:00] [sighs] [laughs] So when you say predict, what is Google trying to predict when it comes to... in the crawl scheduler?

[00:08:10] **Gary Illyes:** [00:08:10] John, pay attention, we are talking about quality.

[00:08:13] **John Mueller:** [00:08:13] Oh, they're predicting the quality.

[00:08:16] **Gary Illyes:** [00:08:16] Among other things, yes. But the quality is the most important thing that we are trying to predict.

[00:08:21] **John Mueller:** [00:08:21] Okay.

[00:08:21] **Gary Illyes:** [00:08:21] Because that's driving most of the decisions that we make, including index selection. Basically whether we should index something or not, or crawl something or not and so on.

[00:08:34] So yeah, the quality, that's the biggest player in the group. And then there's also change frequency. Like for example, if you have a pattern, URL pattern on the site where you are storing your site's legalese, like the terms of service and the privacy policy and whatever, those pages tend to change not. [laughs]

[00:08:55] **John Mueller:** [00:08:55] Yeah.

[00:08:55] **Gary Illyes:** [00:08:55] Maybe once a year or once every five years. So we don't have to crawl that pattern all that often. And then we can focus on the other parts of your site, let's say.

[00:09:08] **John Mueller:** [00:09:08] Okay.

[00:09:08] **Gary Illyes:** [00:09:08] So that would also feed into those predictions that we made, like whether we have to crawl something or not.

[00:09:14] **John Mueller:** [00:09:14] So would the change frequency be related to quality or are those two completely separated things coming together?

[00:09:23] **Gary Illyes:** [00:09:23] They are two distinct signals that just goes into scheduling.

[00:09:27] **John Mueller:** [00:09:27] Okay.

[00:09:28] **Gary Illyes:** [00:09:28] Because we are talking about scheduling, but the most important is quality. It's always quality. And I think externally, people don't necessarily want to believe it, but the quality, that's the biggest driver for most of the indexing and crawling decisions that we make, or our systems making at least.

[00:09:49] **John Mueller:** [00:09:49] Cool. You mentioned the legalese pages like terms of service. Would those be lower quality, or would they be...

[00:09:58] **Gary Illyes:** [00:09:58] No.

[00:09:58] **John Mueller:** [00:09:58] ... equally high quality?

[00:10:00] **Gary Illyes:** [00:10:00] They can be equally high quality. It really depends how the page is written and what the content is and how the content is, versus like, "Oh, this is a legal topic, I don't want to crawl it, oh!"

[00:10:13] **John Mueller:** [00:10:13] [laughs]

[00:10:14] **Gary Illyes:** [00:10:14] Googlebot is not picky about the topic.

[00:10:17] **John Mueller:** [00:10:17] Okay.

[00:10:18] **Gary Illyes:** [00:10:18] And nor is our indexing or crawl scheduling. So basically, at that stage, we don't even know that something is about privacy or something. We just crawl it and then analyze it. Try to estimate the quality of the content, come up with a bunch of numbers that represent quality, and then use that in future crawls.

[00:10:44] **John Mueller:** [00:10:44] Cool. You mentioned that when legal pages are a separate part of a site, then for change frequency, we would try to take that into account. Does that also work with quality? Could you have one part of your site be more high quality, another part of the site lower quality?

[00:11:03] **Gary Illyes:** [00:11:03] Yeah, I think so. If you think about it you I don't know, potatopeelers.com/garybrand, and it's obvious the highest possible quality, all the pages under that pattern are the highest quality, and then, you have potatopeelers.com/gibberishpotatopeelers.

[00:11:27] **John Mueller:** [00:11:27] Okay.

[00:11:27] **Gary Illyes:** [00:11:27] Eventually, we might learn that URLs under that pattern hold content that people will never actually search for, so why would we index it?

[00:11:37] **John Mueller:** [00:11:37] Okay.

[00:11:37] **Gary Illyes:** [00:11:37] Or crawl it?

[00:11:39] And since we're talking about potato peelers, do you think the quality of the product plays a role, or is it just the quality of the text? If you have a really good potato peeler, but you have, I don't know, a cat walked over the keyboard and wrote the landing page for it?

[00:11:55] **Gary Illyes:** [00:11:55] If the content is gibberish, then it's highly unlikely that we are going to index it or crawl it. If, well, it's the first page, then we are probably going to crawl it, but I think we are not going to index it because we still have to be able to extract the words that you use on the page, and if the cat was kind enough to walk over your keyboard, then that might actually generate just gibberish, basically how my sites are, and it would be probably very hard to get it indexed, just because we can't actually extract anything from the page that would make sense.

[00:12:31] **John Mueller:** [00:12:31] So you might get lucky and that we crawl it, but then you get unlucky because we notice it's actually not that useful after all, and we don't use it for indexing.

[00:12:38] **Gary Illyes:** [00:12:38] Yeah, yeah. I guess so.

[00:12:40] **John Mueller:** [00:12:40] Okay.

[00:12:41] **Gary Illyes:** [00:12:41] And then, once we learned that we can apply that to... or eventually, we can apply that on the whole URL pattern, so /gibberishpotatopeelers might actually be treated differently than /garybrand, because the pages under gibberishpotatopeeler tend to be lower quality.

[00:13:01] **John Mueller:** [00:13:01] Okay.

[00:13:02] **Gary Illyes:** [00:13:02] And of course, we can also apply this on, like you have UGC, User Generated Content, but let's say that you have User Generated Content on your site and it's restricted to one particular pattern like /ugc/john and /gary and /whatever. Then eventually, we might learn that the vast majority of the content there is not the highest of quality, and then we might crawl less from there.

[00:13:34] And I think there were quite a few SEOs who ran experiments with this, and they moved out the lowish quality UGC off of the main site, and then they started seeing an uptick in how we crawled...

[00:13:47] **John Mueller:** [00:13:47] Okay.

[00:13:47] **Gary Illyes:** [00:13:47] ... certain patterns on the site. So I think people already... in a sense, people already know about this, it's just not very obvious most of the time.

[00:13:59] **John Mueller:** [00:13:59] Okay. So it also sounds like you can improve the quality of your site. It's not like you're caught with gibberish wants, and then Google will never look at your site again.

[00:14:10] **Gary Illyes:** [00:14:10] Sure, I mean my sites are all a testament to that. You can always improve the site and then get out of the crawling jail, I guess.

[00:14:21] **John Mueller:** [00:14:21] [laughs]

[00:14:23] **Gary Illyes:** [00:14:23] Or whatever... I can't come up with a better name. But yeah, if you're removing low quality content from your site, then that will ultimately improve the rest of the site.

[00:14:34] I think the hardest part is trying to figure out what is lower quality, especially if you have a massive site, or a site that's been around for thousands of years like webmasterworld.com, or... what's Barry's site? Barry Schwartz's site? searchengineroundtable.com.

[00:14:55] If you have one of those sites, then it's very hard to go back and tried to figure out what are the pages that we might consider lower quality, even if we have documentation about what we consider quality content.

[00:15:11] **John Mueller:** [00:15:11] One thing I've seen SEOs do is look at their analytics and say everything that gets fewer than, I don't know, seven clicks is low quality and I'll delete it. Does that make sense, or is that...

[00:15:25] **Gary Illyes:** [00:15:25] I mean seven in general, that makes perfect sense, I think.

[00:15:28] **John Mueller:** [00:15:28] Seven? Okay.

[00:15:29] **Gary Illyes:** [00:15:29] Seven is a good number.

[00:15:30] **John Mueller:** [00:15:30] Okay. I mean, it's obviously tricky with a really large website, and if you're also still producing new content. Do you think like for Barry's site, I feel kind of sorry for him now that we're discussing his site and he's not here. [laughs]

[00:15:47] **Gary Illyes:** [00:15:47] John, don't feel sorry for him, he loves the attention.

[00:15:51] **John Mueller:** [00:15:51] But I mean...

[00:15:52] **Gary Illyes:** [00:15:52] We should have him back.

[00:15:54] **John Mueller:** [00:15:54] We should have a back, that's true, yeah. He's great.

[00:15:56] **Gary Illyes:** [00:15:56] He's fun.

[00:15:58] **John Mueller:** [00:15:58] Like for his site, do you think it makes sense to go through all of the old content and delete or improve the content, or should he just focus on making sure the new content is high quality?

[00:16:09] **Gary Illyes:** [00:16:09] I think in case of these sites, like webmasterworld or searchengineroundtable, it doesn't actually matter that much anymore, because they are so established that they get direct visitors anyway, and people are looking for these sites anyway, regardless of what we are doing. They are linking to these sites a lot, so we see that people actually look for these sites.

[00:16:38] Like for example, when you go to, I don't know, randomsite.com, you see a blog post about SEO, and then that blog post is linking out to searchengineroundtable, for example, that's a good hint for us that that target's site, searchengineroundtable, might be important.

[00:16:58] And the more links you see from normal sites, not profile pages and random gibberish sites like johnwoo.com...

[00:17:05] **John Mueller:** [00:17:05] [gasps] [laughs]

[00:17:07] **Gary Illyes:** [00:17:07] Oh, I'm sorry, did I say that out loud? [laughs] These links that people litter on the Internet on normal places, not weird places, they can actually be very helpful estimating how important something is for getting in the index. And so, for these sites like webmasterworld or searchengineroundtable, it doesn't really matter anymore that in the past, they might have had some lower quality content, UGC or not, because people are linking to these sites. You don't even have to tell people like that, "Oh, please sir, give me one more link!"

[00:17:43] **John Mueller:** [00:17:43] So I guess we don't disclose what individual factors we look at for quality, but I have a few that people ask me from time to time. Can I ask you?

[00:17:53] **Gary Illyes:** [00:17:53] Okay.

[00:17:54] **John Mueller:** [00:17:54] Okay. Duplicate content.

[00:17:57] **Gary Illyes:** [00:17:57] Generally, no. People are afraid of duplicate content way too much, and it's normal to have some duplication on the site and don't worry too much about it. Try not to over duplicate it, because then, you might actually bring down your site accidentally, because we are going to be very excited about crawling all those new URLs that we discover, but otherwise, I wouldn't be too worried about duplicate content.

[00:18:22] **John Mueller:** [00:18:22] Okay. Affiliate links.

[00:18:25] **Gary Illyes:** [00:18:25] Meh.

[00:18:26] **John Mueller:** [00:18:26] Meh? Not a problem. Okay, good.

[00:18:28] **Gary Illyes:** [00:18:28] It depends on the content. Like if you're just copying something from the affiliate marketplace that the product owner gave you and you are just taking that and then you're putting it on your site and then placing a link, then that's not that useful, you're not creating helpful content, you're just copy-pasting something.

[00:18:47] But if you actually reviewed the product, like the potato peeler, and you took pictures, and then you even stress tested the blade, for example, and then you wrote the thing about that, then that's helpful. It's unique helpful content. So you should be able to earn a buck out of that thing and then placing an affiliate link somewhere in there. I think it's fine.

[00:19:19] **John Mueller:** [00:19:19] Okay.

[00:19:19] **Gary Illyes:** [00:19:19] But of course, individual country laws, you might need to disclose things, but from our perspective, it's fine.

[00:19:27] **John Mueller:** [00:19:27] Cool. Translated content. If you translate something into Swiss German or normal German.

[00:19:34] **Gary Illyes:** [00:19:34] Automatically translated content. I would strongly prefer if people review the translations, because Lizzy, our tech writer, and I deal a lot with translations, and we use machine translation with the review, and we see stuff. We've seen things. [laughs]

[00:19:53] **John Mueller:** [00:19:53] Machines.

[00:19:55] **Gary Illyes:** [00:19:55] So I would prefer if people reviewed what these machine translations give to people.

[00:20:00] **John Mueller:** [00:20:00] Okay, cool. I feel we could have hours of discussion on the topic of quality.

[00:20:06] **Gary Illyes:** [00:20:06] Oh, yeah.

[00:20:07] **John Mueller:** [00:20:07] But we have to wrap it up. Alright, well, thank you for joining me here, Gary, this was fun. We should do this from time to time.

[00:20:15] **Gary Illyes:** [00:20:15] It was, wasn't it.

[00:20:16] **John Mueller:** [00:20:16] Yes.

[00:20:17] ♪ [music] ♪

[00:20:21] **John Mueller:** [00:20:21] We've been having fun with these podcast episodes, and I hope you, the listener, has been finding them both entertaining and insightful as well. Feel free to drop me a note on Twitter or chat with us at one of the next events that we go to if you have any thoughts. And of course, don't forget to like and subscribe. Thank you and goodbye.

[00:20:42] ♪ [music] ♪