Analyzing “How Google Search Works” Changes from Google

Google has made some new substantial changes to their “How Google Search Works” search documents for website owners. And as always when Google makes changes to important documents with impact on SEO, such as How Search Works and the Quality Rater Guidelines, there are some key insights SEOs can gleam from the new changes Google has made.

Of particular note, Google detailing how it views a “document” as potentially comprising of more than one webpage, what Google considers primary and secondary crawls, as well as an

update to their reference of “more than 200 ranking factors” which has been present in this document since 2013.

But here are the changes and what they mean for SEOs.

Crawling

Google has greatly expanded this section.

They made a slight change to wording, with “some pages are known because Google has already crawled them before” changed to “some pages are known because Google has already visited them before.” This is a fairly minor change, primarily because Google decided to include an expanded section detailing what crawling actually is.

Google removed:

> *This process of discovery is called crawling.*

The removal of the crawling definition was simply because it was redundant. In Google’s expanded crawling section, they included a much more detailed definition and description of crawling instead.

The added definition:

> *Once Google discovers a page URL, it visits, or crawls, the page to find out what’s on it. Google renders the page and analyzes both the text and non-text content and overall visual layout to decide where it should appear in Search results. The better that Google can understand your site, the better we can match it to people who are looking for your content.*

There is still a great debate on how much page layout is taken into account. There was the page layout algo that was released many years, in order to penalize content that was

pushed well below the fold in order to increase the odds a visitor might click on an advertisement that appeared above the fold instead. But with more traffic moving to mobile, and the addition of mobile first indexing, the importance of above and below the fold for on page layout seemingly was less important.

When it comes to page layout and mobile first, Google says:

> *Don't let ads harm your mobile page ranking. Follow the Better Ads Standard when displaying ads on mobile devices. For example, ads at the top of the page can take up too much room on a mobile device, which is a bad user experience.*

But in How Google Search Works, Google is specifically calling attention to the "overall visual layout" with "where it should appear in Search results."

It also brings attention to "non-text" content. While the most obvious of this refers to image content, the referral to it is quite open ended. Could this refer to OCR as well, which we know Google has been dabbling in?

Improving Your Crawling

Under the "to improve your site crawling" section, Google has expanded this section significantly as well.

Google has added this point:

> *Verify that Google can reach the pages on your site, and that they look correct. Google accesses the web as an anonymous user (a user with no passwords or information). Google should also be able to see all the images and other elements of the page to be able to understand it correctly. You can do a quick check by typing your page URL in the Mobile-Friendly test tool.*

This is a good point – so many new site owners end up accidentally blocking Googlebot from crawling or not realizing their site is set to be only viewable by logged in users only. This makes it clear that site owners should try viewing their site without also being logged into it, to see if there are any unexpected accessibility or other issues that aren't note when logged in as an admin or high level user.

Also recommending site owners check their site via the Mobile-Friendly testing tool is good, since even seasoned SEOs use the tool to quickly see if there are Googlebot specific issues with how Google is able to see, render and crawl a specific webpage – or a competitor's page.

Google expanded their specific note about submitting a single page to the index.

> *If you've created or updated a single page, you can submit an individual URL to Google. To tell Google about many new or updated pages at once, use a sitemap.*

Previously, it just mentioned submitting changes to a single page using the submit URL tool. This just adds clarification to those who are newer to SEO that they do not need to submit every single new or updated pages to Google individually, but that using sitemaps is the best way to do that. There have definitely been new site owners who add each page to Google using that tool because they don't realize sitemaps is a thing. But part of this is that WordPress is such a prevalent way to create a new website, yet it does not have native support for sitemaps (yet), so site owners need to either install a specific sitemaps plugin or use one of the many SEO tool plugins that offer sitemaps as a feature.

This new change also highlights using the tool for creating pages as well, instead of just the previous reference of "changes to a single page."

Google has also made a change to the section about “if you ask Google to crawl only one page” section as well. They are now referencing what Google views as a “small site” – according to Google, a smaller site is one with less than 1,000 pages.

Google also stresses the importance of a strong navigation structure, even for sites it considers “small.” It says site owners of small sites can just submit their homepage to Google, “provided that Google can reach all your other pages by following a path of links that start from your homepage.”

With so many sites being on WordPress, it is less likely that there will be random orphaned pages that are not accessible by following links from the homepage But depending on the specific WordPress theme used, sometimes there can be orphaned pages from pages being added but not manually added to the pages menu… in these cases, if a sitemap is used as well, those pages shouldn’t be missed even if not directly linked from the homepage.

In the “get your page linked to by another page” section, Google has added that links in “advertisements links that you pay for in other sites, links in comments, or other links that don’t follow the Google Webmaster Guidelines won’t be followed *by Google.*” A small change, but Google is making it clear that it is a Google specific thing that these links won’t be followed, but they might be followed by other search engines.

But perhaps the most telling part of this is at the end of the crawling section, Google adds:

> *Google doesn’t accept payment to crawl a site more frequently, or rank it higher. If anyone tells you otherwise, they’re wrong.*

It has long been an issue with scammy SEO companies to guarantee first positioning on Google, to increase rankings or requiring payment to submit a site to Google. And with the

ambiguous Google Partner badge for AdWords, many use the Google Partners badge to imply they are certified by Google for SEO and organic ranking purposes. That said, most of those who are reading the How Search Works probably are already aware of this. But nice to see Google add this in writing again, for times when SEOs need to prove to clients that there is not a “pay to win” option, outside of AdWords, or simply to show someone who might be falling for some scammy SEO company’s claims of Google rankings.

The Long Version

Google then gets into what they call the “long version” of How Google Search Works, with more details on the above sections, covering more nuances that impact SEO.

Crawling

Google has changed how they refer to the “algorithmic process”. Previously, it stated “Googlebot uses an algorithmic process: computer programs determine which sites to crawl, how often and how many pages to fetch from each site.” Curiously, they removed the reference to “computer programs”, which provoked the question about which computer programs exactly Google was using.

The new updated version simply states:

> *Googlebot uses an algorithmic process to determine which sites to crawl, how often, and how many pages to fetch from each site.*

Google also updated the wording for the crawl process, changing that it is “augmented with sitemap data” to “augmented by sitemap” data.

Google also made a change where it referenced that Googlebot

“detects” links and changed it to “finds” links, as well as changes from Googlebot visiting “each of these websites” to the much more specific “page”. This second change makes it more accurate since Google visiting a website won’t necessarily mean it crawls all links on all pages. The change to “page” makes it more accurate and specific for webmasters.

Previously it read:

> *As Googlebot visits each of these websites it detects links on each page and adds them to its list of pages to crawl.*

Now it reads:

> *When Googlebot visits a page it finds links on the page and adds them to its list of pages to crawl.*

Google has added a new section about using Chrome to crawl:

> *During the crawl, Google renders the page using a recent version of Chrome. As part of the rendering process, it runs any page scripts it finds. If your site uses dynamically-generated content, be sure that you follow the JavaScript SEO basics.*

By referencing a recent version of Chrome, this addition is clarifying the change from last year where Googlebot was finally upgraded to the latest version of Chromium for crawling, an update from Google only crawling with Chrome 41 for years.

Google also notes it runs “any page scripts it finds,” and advises site owners to be aware of possible crawl issues as a result of using dynamically-generated content with the use of JavaScript, specifying that site owners should ensure they follow their JavaScript SEO basics.

Google also details the primary and secondary crawls,

something that has garnered much confusion since Google revealed primary and secondary crawls, but Google's details in this How Google Search Works documents detail it differently than how some SEOs previously interpreted it.

Here is the entire new section for primary and secondary crawls:

> ***Primary crawl / secondary crawl***
>
> *Google uses two different crawlers for crawling websites: a mobile crawler and a desktop crawler. Each crawler type simulates a user visiting your page with a device of that type.*
>
> *Google uses one crawler type (mobile or desktop) as the primary crawler for your site. All pages on your site that are crawled by Google are crawled using the primary crawler. The primary crawler for all new websites is the mobile crawler.*
>
> *In addition, Google recrawls a few pages on your site with the other crawler type (mobile or desktop). This is called the secondary crawl, and is done to see how well your site works with the other device type.*

In this section, Google refers to primary and secondary crawls as being specific to their two crawlers – the mobile crawler and the desktop crawler. Many SEOs think of primary and secondary crawling in reference to Googlebot making two passes over a page, where javascript is rendered on the secondary crawl. So while Google clarifies their use of desktop and mobile Googlebots, the use of language here does cause confusion for those who use this to refer to the primary and secondary crawls for javascript purposes. So to be clear, Google's reference to their primary and secondary crawl has nothing to do with javascript rendering, but only to how they use both mobile and desktop Googlebots to crawl and check a

page.

What Google is clarifying in this specific reference to primary and secondary crawl is that Google is using two crawlers – both mobile and desktop versions of Googlebot – and will crawl sites using a combination of both.

Google did specifically state that new websites are crawled with the mobile crawler in their "Mobile-First Indexing Best Practices" document, as of July 2019. But this is the first time it has made an appearance in their How Google Search Works document.

Google does go into more detail about how it uses both the desktop and mobile Googlebots, particularly for sites that are currently considered mobile first by Google. It wasn't clear just how much Google was checking desktop versions of sites if they were mobile first, and there have been some who have tried to take advantage of this by presenting a spammier version to desktop users, or in some cases completely different content. But Google is confirming it is still checking the alternate version of the page with their crawlers.

So sites that are mobile first will see some of their pages crawled with the desktop crawler. However, it still isn't clear how Google handles cases where they are vastly different, especially when done for spam reasons, as there doesn't seem to be any penalty for doing so, aside from a possible spam manual action if it is checked or a spam report is submitted. And this would have been a perfect opportunity to be clearer about how Google will handle pages with vastly different content depending on whether it is viewed on desktop or on mobile. Even in the mobile friendly documents, Google only warns about ranking differences if content is on the desktop version of the page but is missing on the mobile version of the page.

How does Google find a page?

Google has removed this section entirely from the new version of the document.

Here is what was included in it:

> ***How does Google find a page?***
>
> *Google uses many techniques to find a page, including:*
>
> - *Following links from other sites or pages*
> - *Reading sitemaps*

It isn't clear why Google removed this specifically. It is slightly redundant, but it was missing the submitting a URL option as well.

Improving Your Crawling

Google makes the use of hreflang a bit clearer, especially for those who might just be learning what hreflang is and how it works by providing a bit more detail.

Formerly it said "Use hreflang to point to alternate language pages." Now it states "Use hreflang to point to alternate versions of your page in other languages."

Not a huge change, but a bit clearer.

Google has also added two new points, providing more detail about ensuring Googlebot is able to access all the content on the page, not just the content (words) specifically.

First, Google added:

> *Be sure that Google can access the key pages, and also the important resources (images, CSS files, scripts) needed to render the page properly.*

So Google is stressing about ensuring Google can access all the important content. And it is also specifically calling attention to other types of elements on the page that Google wants to also have access to in order to properly crawl the page, including images, CSS and scripts. For those webmasters who went through the whole “mobile first indexing” launch, they are fairly familiar with issues surrounding blocking files, especially CSS and scripts, something that some CMS had blocked Googlebot from crawling by default.

But for newer site owners, they might not realize this was possible, or that they might be doing it. It would have been nice to see Google add specific information on how those newer to SEO can check for this, particularly for those who also might not be clear on what exactly “rendering” means.

Google also added:

> *Confirm that Google can access and render your page properly by running the URL Inspection tool on the live page.*

Here Google does add specific information about using the URL Inspection tool in order to see what site owners are blocking or content that is causing issues when Google tries to render it. I think these last two new points could have been combined, and made slightly clearer for how site owners can use the tool to check for all these issues.

Indexing

Google has made significant changes to this section as well. And Google starts off with making major changes to the first paragraph. Here is the original version:

> *Googlebot processes each of the pages it crawls in order to compile a massive index of all the words it sees and their location on each page. In addition, we process information included in key content tags and attributes, such as <title>*

> *tags and alt attributes.*

The updated version now reads:

> *Googlebot processes each page it crawls in order to understand the content of the page. This includes processing the textual content, key content tags and attributes, such as <title> tags and alt attributes, images, videos, and more.*

Google no longer states it processes pages to "compile a massive index of all the words it sees and their location on each page." This was always a curious way for them to call attention to the fact they are simply indexing all words it comes across and their position on a page, when in reality it is a lot more complex than that. So it definitely clears that up.

They have also added that they are processing "textual content" which is basically calling attention to the fact it indexes the words on the page, something that was assumed by everyone. But it does differentiate between the new addition later in the paragraph regarding images, videos and more.

Previously, Google simply made reference to attributes such as title and alt tags and attributes. But now it is getting more granular, specifically referring to "images, videos and more." However, this does mean Google is considering images, videos and "more" to understand the content on the page, which could affect rankings.

Improving your Indexing

Google changed "read our SEO guide for more tips" to "Read our basic SEO guide and advanced user guide for more tips."

What is a document?

Google has added a massive section here called "What is a

document?” It talks specifically about how Google determines what is a document, but also includes details about how Google views multiple pages with identical content as a single document, even with different URLs, and how it determines canonicals.

First, here is the first part of this new section:

> ***What is a “document”?***
>
> *Internally, Google represents the web as an (enormous) set of documents. Each document represents one or more web pages. These pages are either identical or very similar, but are essentially the same content, reachable by different URLs. The different URLs in a document can lead to exactly the same page (for instance, example.com/dresses/summer/1234 and example.com?product=1234 might show the same page), or the same page with small variations intended for users on different devices (for example, example.com/mypage for desktop users and m.example.com/mypage for mobile users).*
>
> *Google chooses one of the URLs in a document and defines it as the document’s canonical URL. The document’s canonical URL is the one that Google crawls and indexes most often; the other URLs are considered duplicates or alternates, and may occasionally be crawled, or served according to the user request: for instance, if a document’s canonical URL is the mobile URL, Google will still probably serve the desktop (alternate) URL for users searching on desktop.*
>
> *Most reports in Search Console attribute data to the document’s canonical URL. Some tools (such as the Inspect URL tool) support testing alternate URLs, but inspecting the canonical URL should provide information about the alternate URLs as well.*
>
> *You can tell Google which URL you prefer to be canonical, but Google may choose a different canonical for various reasons.*

So the tl:dr is that Google will view pages with identical or near-identical content as the same document, regardless of how many of them there are. For seasoned SEOs, we know this as internal duplicate content.

Google also states that when Google determines these duplicates, they may not be crawled as often. This is important to note for site owners that are working to de-duplicate content which Google is considering duplicate. So it would be more important to submit these URLs to be recrawled, or give those newly de-duplicated pages links from the homepage in order to ensure Google recrawls and indexed the new content, so Google de-dupes them properly.

It also brings up an important note about desktop versus mobile, that Google will still likely serve the desktop version of a page instead of the mobile version for desktop users, when a site has two different URLs for the same page where is designed for mobile users and the other for desktop. While many websites have changed to serving the same URL and content for both using responsive design, some sites still run two completely different sites and URLs for desktop and mobile users.

Google also mentions that you can tell Google the URL you prefer Google to use as the canonical, but states they can chose a different URL "for various reasons." While Google doesn't detail specifics about why Google might choose a different canonical than the one the site owner specifies, it is usually due to http vs https, if a page is included in a sitemap or not, page quality, if the pages appear to be completely different and should not be canonicalized, or due to significant incoming links to the non-canonical URL.

Google has also included definitions for many o the terms used by SEOs and in Google Search Console.

> ***Document:*** *A collection of similar pages. Has a canonical URL,*

> *and possibly alternate URLs, if your site has duplicate pages. URLs in the document can be from the same or different organization (the root domain, for example "google" in www.google.com). Google chooses the best URL to show in Search results according to the platform (mobile/desktop), user language‡ or location, and many other variables. Google discovers related pages on your site by organic crawling, or by site-implemented features such as redirects or <link rel=alternate/canonical> tags. Related pages on other organizations can only be marked as alternates if explicitly coded by your site (through redirects or link tags).*

Again, Google is talking about the fact a single document can encompass more than just a single URL, as Google will consider a single document to potentially have many duplicate or near duplicate pages as well as pages assigned via canonical. Google makes specific mention about "alternates" that appear on other sites, that can only be considered alternates if the site owner specifically codes it. And that Google will choose the best URL from within the collection of documents to show.

But it fails to mention that Google can consider pages duplicate on other sites and will not show those duplicates, even if they aren't from the same sites, something that site owners see happen frequently when someone steals content and sometimes sees the stolen version ranking over the original.

There was a notation added for the above, dealing with hreflang.

> *‡Pages with the same content in different languages are stored in different documents that reference each other using hreflang tags; this is why it's important to use hreflang tags for translated content.*

Google shows that it doesn't include identical content under the same "document" when it is simply in a different language,

which is interesting. But Google is tressing the importance of using hreflang in these cases.

> ***URL:*** *The URL used to reach a given piece of content on a site. The site might resolve different URLs to the same page.*

Pretty self explanatory, although it does have reference to the fact different URLs can be resolved to the same page, presumably such as with redirects or alias.

> ***Page:*** *A given web page, reached by one or more URLs. There can be different versions of a page, depending on the user's platform (mobile, desktop, tablet, and so on).*

Also pretty self explanatory, bringing up the specifics that some site owners can be served different versions of the same page, such as if they try and view the same page on a mobile device versus a desktop computer.

> ***Version:*** *One variation of the page, typically categorized as "mobile," "desktop," and "AMP" (although AMP can itself have mobile and desktop versions). Each version can have a different URL (example.com vs m.example.com) or the same URL (if your site uses dynamic serving or responsive web design, the same URL can show different versions of the same page) depending on your site configuration. Language variations are not considered different versions, but different documents.*

Simply clarifying with greater details the different versions of a page, and how Google typically categorizes them as "mobile," "desktop," and "AMP".

> ***Canonical page or URL:*** *The URL that Google considers as most representative of the document. Google always crawls this URL; duplicate URLs in the document are occasionally crawled as well.*

Google states here again that non-canonical pages are not crawled as often as the main canonical that a site owner assigns to a group of pages they want canonical. Google does not include specific mention here that they sometimes chose a different page as the canonical one, even if there is a specific page designated as the canonical one.

> ***Alternate/duplicate page or URL:*** *The document URL that Google might occasionally crawl. Google also serves these URLs if they are appropriate to the user and request (for example, an alternate URL for desktop users will be served for desktop requests rather than a canonical mobile URL).*

The key takeaway here is that Google "might" occasionally crawl the site's duplicate or alternative page. And here they stress that Google will serve these alternative URLs "if they are appropriate." It is unfortunate they don't go into greater detail in why they might serve these pages instead of the canonical, outside of the mention of desktop versus mobile, as we have seen many cases where Google picks a different page to show other than the canonical for a myriad of reasons.

Google also fails to mention how this impacts duplicate content found on other sites, we we do know Google will crawl those less often as well.

> ***Site:*** *Usually used as a synonym for a website (a conceptually related set of web pages), but sometimes used as a synonym for a Search Console property, although a property can actually be defined as only part of a site. A site can span subdomains (and even domains, for properly linked AMP pages).*

Interesting to note here what they consider a website – a conceptually related set of webpages – and how it related to the usage of a Google Search Console property, as "a property can actually be defined as only part of a site."

Google does make mention that AMP, which technically appear on a different domain, are considered part of the main site.

Serving Results

Google has made a pretty interesting specific change here in regards to their ranking factors. Previously, Google stated:

> *Relevancy is determined by over 200 factors, and we always work on improving our algorithm.*

Google has now updated this "over 200 factors" with a less specific one.

> *Relevancy is determined by hundreds of factors, and we always work on improving our algorithm.*

The 200 factors in the How Google Search Works dates back to 2013 when the document was launched, although then it also made reference to PageRank ("Relevancy is determined by over 200 factors, one of which is the PageRank for a given page") which Google removed when they redesigned their document in 2018.

While Google doesn't go into specifics on the number anymore, it can be assumed that a significant number of ranking factors have been added since 2013 when this was first claimed in this document. But I am sure some SEOs will be disappointed we don't get a brand new shiny number like "over 500" ranking factors that SEOs can obsess about.

Final Thoughts

There are some pretty significant changes made to this document that SEOs can get a bit of insight from.

Google's description of what it considers a document and how

it relates to other identical or near-identical pages on a site is interesting, as well as Google's crawling behavior towards the pages within a document it considers as alternate pages. While this behavior has often been noted, it is more concrete information on how site owners should handle these duplicate and near-duplicate pages, particularly when they are trying to un-duplicate those pages and see them crawled and indexed as their own document.

They added a lot of useful advice for newer site owners, which is particularly helpful with so many new websites coming online this year due to the global pandemic. Things such as checking a site without being logged in, how to submit both pages and sites to Google, etc.

The mention of what Google considers a "small site" is interesting because it gives a more concrete reference point for how Google sees large versus small sites. For some, a small site could mean under 30 pages and the idea of a site with millions of pages being unfathomable. And the reinforcement of a strong navigation, even for "small sites" is useful for showing site owners and clients who might push for navigation that is more aesthetic than practical for both usability and SEO.

The primary and secondary crawl additions will probably cause some confusion for those who think of primary and secondary in terms of how Google processes scripts on a page when it crawls it. But it is nice to have more concrete information on how and when Google will crawl using the alternate version of Googlebot for sites that are usually crawled with either the mobile Googlebot or the desktop one.

Lastly, the change from the "200 ranking factors" to a less specific, but presumably much higher number of ranking factors will disappoint some SEOs who liked having some kind of specific number of potential ranking factors to work out.

The following two tabs change content below.
Jennifer Slegg is a longtime speaker and expert in search engine marketing, working in the industry for almost 20 years. When she isn’t sitting at her desk writing and working, she can be found grabbing a latte at her local Starbucks or planning her next trip to Disneyland. She regularly speaks at Pubcon, SMX, State of Search, Brighton SEO and more, and has been presenting at conferences for over a decade.

Latest posts by Jennifer Slegg (see all)

http://www.thesempost.com/analyzing-google-search-works-changes-google/