https://developers.google.com/search/docs/beginner/how-search-works

How Google Search Works (for beginners)

Google gets information from many different sources, including:

- Web pages
- User-submitted content such as Google My Business and Maps user submissions
- Book scanning
- Public databases on the internet
- Many other sources

However, this page focuses on web pages. Google follows three basic steps to generate results from web pages:

- Crawling
- Indexing
- Serving (and ranking)

Crawling

The first step is finding out what pages exist on the web. There isn't a central registry of all web pages, so Google must constantly search for new pages and add them to its list of known pages. Some pages are known because Google has already visited them before. Other pages are discovered when Google follows a link from a known page to a new page. Still other pages are discovered when a website owner submits a list of pages (a *sitemap*) for Google to crawl. If you're using a managed web host, such as Wix or Blogger, they might tell Google to crawl any updated or new pages that you make.

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.

To improve your site crawling:

- 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.
- 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.

- **If you ask Google to crawl only one page, make it your home page.** Your home page is the most important page on your site, as far as Google is concerned. To encourage a complete site crawl, be sure that your home page (and all pages) contain a good site navigation system that links to all the important sections and pages on your site; this helps users (and Google) find their way around your site. For smaller sites (less than 1,000 pages), making Google aware of only your homepage is all you need, provided that Google can reach all your other pages by following a path of links that start from your homepage.
- Get your page linked to by another page that Google already knows about. **However,** be warned 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.

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

Indexing

After a page is discovered, Google tries to understand what the page is about. This process is called *indexing*. Google analyzes the content of the page, catalogs images and video files embedded on the page, and otherwise tries to understand the page. This information is stored in the *Google index*, a huge database stored in many, many (many!) computers.

To improve your page indexing:

- Create short, meaningful page titles.
- Use page headings that convey the subject of the page.
- Use text rather than images to convey content. Google can understand some image and video, but not as well as it can understand text. At minimum, annotate your video and images with alt text and other attributes as appropriate.

Serving (and ranking)

When a user types a query, Google tries to find the most relevant answer from its index based on many factors. Google tries to determine the highest quality answers, and factor in other considerations that will provide the best user experience and most appropriate answer, by considering things such as the user's location, language, and device (desktop or phone). For example, searching for "bicycle repair shops" would show different answers to a user in Paris than it would to a user in Hong Kong. Google doesn't accept payment to rank pages higher, and ranking is done programmatically.

To improve your serving and ranking:

- Make your page fast to load, and mobile-friendly.
- Put useful content on your page and keep it up to date.
- Follow the Google Webmaster Guidelines, which help ensure a good user experience.
- Read more tips and best practices in our SEO starter guide.

- You can find more information here, including the guidelines that we provide to our quality raters to ensure that we're providing good results.

An even longer version

Want more in-depth information about how Search works? Read our Advanced guide to how Google Search works.

https://developers.google.com/search/docs/advanced/guidelines/how-search-works

Advanced: How Search Works

Understanding how Google Search crawls, indexes, and serves content is important when you're debugging issues and anticipating Search behavior on your site.

Crawling

Crawling is the process by which Googlebot visits new and updated pages to be added to the Google index.

We use a huge set of computers to fetch (or "crawl") billions of pages on the web. The program that does the fetching is called Googlebot (also known as a robot, bot, or spider). Googlebot uses an algorithmic process to determine which sites to crawl, how often, and how many pages to fetch from each site.

Google's crawl process begins with a list of web page URLs, generated from previous crawl processes, augmented by Sitemap data provided by website owners. When Googlebot visits a page it finds links on the page and adds them to its list of pages to crawl. New sites, changes to existing sites, and dead links are noted and used to update the Google index.

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.

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.

How does Google know which pages not to crawl?

- Pages blocked in robots.txt won't be crawled, but still might be indexed if linked to by another page. Google can infer the content of the page by a link pointing to it, and index the page without parsing its contents.
- Google can't crawl any pages not accessible by an anonymous user. Thus, any login or other authorization protection will prevent a page from being crawled.
- Pages that have already been crawled and are considered duplicates of another page, are crawled less frequently.

Improve your crawling

Use these techniques to help Google discover the right pages on your site:

- Submit a sitemap.
- Submit crawl requests for individual pages.
- Use a simple, human-readable, and logical URL paths for your pages and provide clear and direct internal links within the site.
- If you use URL parameters on your site for navigation, for instance if you indicate the user's country in a global shopping site, use the URL parameters tool to tell Google about important parameters.
- Use robots.txt wisely: Use robots.txt to indicate to Google which pages you'd prefer Google to know about or crawl first, in order to protect your server load, not as a method to block material from appearing in the Google index.
- Use hreflang to point to alternate versions of your page in other languages.
- Clearly identify your canonical page and alternate pages.
- View your crawl and index coverage using the Index Coverage Report.
- Be sure that Google can access the key pages, and also the important resources (images, CSS files, scripts) needed to render the page properly.
- Confirm that Google can access and render your page properly by running the URL Inspection tool on the live page.

Indexing

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. Googlebot can process many, but not all, content types. For example, we cannot process the content of some rich media files.

Somewhere between crawling and indexing, Google determines if a page is a duplicate or canonical of another page. If the page is considered a duplicate, it will be crawled much less frequently. Similar pages are grouped together into a *document,* which is a group of one or more pages that includes the canonical page (the most representative of the group) and any duplicates found (which might simply be alternate URLs to reach the same page, or might be alternate mobile or desktop versions of the same page).

Note that Google doesn't index pages with a noindex directive (header or tag). However, it must be able to see the directive; if the page is blocked by a robots.txt file, a login page, or other device, it is possible that the page might be indexed even if Google didn't visit it!

Improve your indexing

There are many techniques to improve Google's ability to understand the content of your page:

- Prevent Google from crawling or finding pages that you want to hide using the noindex tag. Don't "noindex" a page that is blocked by robots.txt; if you do so, the noindex tag won't be seen and the page might still be indexed.
- Use structured data.
- Follow the Google Webmaster Guidelines.
- Read our SEO starter guide and advanced user guide for more tips.

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 URL Inspection 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.

Here is a summary of terms, and how they are used in 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).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.

- **URL:** The URL used to reach a given piece of content on a site.
- **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).
- **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.
- **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.
- **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).
- **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).

Serving results

When a user enters a query, our machines search the index for matching pages and return the results we believe are the most relevant to the user. Relevancy is determined by hundreds of factors, and we always work on improving our algorithm. Google considers the user experience in choosing and ranking results, so be sure that your page loads fast and is mobile-friendly.

Improve your serving

There are many ways to improve how Google serves the content of your page:

- If your results are aimed at users in specific locations or languages, you can tell Google your preferences.
- Be sure that your page loads fast and is mobile-friendly.
- Follow the Webmaster Guidelines to avoid common pitfalls and improve your site's ranking.

- Consider implementing Search result features for your site, such as recipe cards or article cards.
- Implement AMP for faster loading pages on mobile devices. Some AMP pages are also eligible for additional search features, such as the top stories carousel.
- Google's algorithm is constantly being improved; rather than trying to guess the algorithm and design your page for that, work on creating good, fresh content that users want, and following our guidelines.