

Webmaster Questions

1. What is cloaking?

The term "cloaking" is used to describe a website that returns altered webpages to search engines crawling the site. In other words, the webserver is programmed to return different content to Google than it returns to regular users, usually in an attempt to distort search engine rankings. This can mislead users about what they'll find when they click on a search result. To preserve the accuracy and quality of our search results, Google may permanently ban from our index any sites or site authors that engage in cloaking to distort their search rankings.

2. Do I need to submit updated and/or outdated links and pages to Google?

Google updates its index frequently, so there's no need to submit updated or outdated links. We should pick up any changes to your site during our next crawl.

3. How do I submit multiple pages?

Please visit our [Add URL](#) page to input your URLs. There's no need to submit each individual page; the domain's top-level page will suffice. Our crawler, Googlebot, will find the rest.

4. Why doesn't Google index any of my pages?

If your pages haven't been indexed yet, it's probably because there aren't enough other pages on the web that link to them. Google looks at the link interconnectedness among pages, relying on the vastness and openness of the Internet to yield the most relevant search results. If other pages don't link to yours, we can't assign your pages a PageRank (our proprietary measure of a page's importance) in a reasonable way. Once other pages point to them, we'll pick your pages up.

5. How long does the Google robot take to index a URL once it's been submitted?

Depending on the timing of the submission and of our crawl, the entire process can take between six and eight weeks.

6. Where is my page's title?

Unlike many search engines, Googlebot can return results for pages that are known but haven't been crawled yet. Since we haven't looked at those pages yet, their titles aren't shown; the Google results page displays the URL instead.

7. How should I request that Google not return cached material from my site?

Google stores many web pages in its cache to retrieve for users as a back-up in case the server where the page resides temporarily fails. Users can view the cached version by choosing the "Cached" link on the search results page. If you don't want your content to be accessible through Google's cache, use a `<META>` tag with a `CONTENT="NOARCHIVE"` attribute. To do so, place the following line in the `<HEAD>` section of your documents:

```
<META NAME="ROBOTS" CONTENT="NOARCHIVE" >
```

This tag tells robots not to archive the page. Google will continue to index and follow links from the page, but will not present cached material to users. If you want to allow other robots to cache your content, but prevent Google's robots from doing so, use the following tag:

```
<META NAME="GOOGLEBOT" CONTENT="NOARCHIVE" >
```

Please note that the change will take effect the next time Google crawls the page containing the `NOARCHIVE` directive in a `<META>` tag. If you want this change to take effect sooner, the site owner must contact us and request immediate removal of archived content. Note also that the `NOARCHIVE` directive only controls whether a cached version of the page is made available. To control whether the page is indexed, use `CONTENT="NOINDEX"`. To control whether links are followed, use `CONTENT="NOFOLLOW"`. For more information, see the [Robots Exclusion page](#).

Googlebot Technology Questions

1. How should I request that Google not crawl part or all of my site?

The standard for robot exclusion given at <http://www.robotstxt.org/wc/norobots.html> provides for a file called `robots.txt` that you can put on your server to exclude Googlebot and other web crawlers. (Googlebot has a user-agent of "Googlebot".)

Googlebot also understands some extensions to the `robots.txt` standard. Disallow patterns may include `*` to match any sequence of characters, and patterns may end in `$` to indicate the end of a name. For example, to prevent Googlebot from crawling files that end in `.gif`, you may use the following `robots.txt` entry:

```
User-Agent: Googlebot
Disallow: /*.gif$
```

Please note that Googlebot does *not* interpret a 401/403 response ("Unauthorized"/"Forbidden") to a robots-txt fetch as a request not to crawl any pages on the site. To prevent Googlebot and other web crawlers from crawling any page on your site, you may use the following `robots.txt` entry:

```
User-Agent: *
Disallow: /
```

Please note also that each port must have its own `robots.txt` file. In particular, if you serve content via both `http` and `https`, you'll need a separate `robots.txt` file for each of these protocols. For example, if you wanted to allow all filetypes to be served via `http` but only `.html` pages to be served via `https`, the `robots.txt` file for the `http` protocol (<http://yourserver.com/robots.txt>) would be:

```
User-Agent: *
Allow: /
```

The `robots.txt` file for the `https` protocol (<https://yourserver.com/robots.txt>) would be:

```
User-Agent: *
Disallow: /
Allow: /*.html$
```

Another standard which is more convenient for page-by-page use involves adding a `<META>` tag to an HTML page to tell robots not to index the page or not to follow the links it contains. This standard is described at <http://www.robotstxt.org/wc/exclusion.html>. You may also want to read [what the HTML standard has to say about these tags](#). Remember that changing your server's `robots.txt` file or changing the `<META>` tags on its pages will not cause an immediate change in the results that Google returns, since your changes must propagate to Google's next index of the web before being reflected in Google search results.

2. Why is Googlebot asking for a file called `robots.txt` that isn't on my server?

`robots.txt` is a standard document that can tell Googlebot not to download some or all information from your web server. For information on how to create a `robots.txt` file, see [The Robot Exclusion Standard](#).

3. Why is Googlebot trying to download incorrect links from my server? Or from a server that doesn't exist?

It's a fact of life on the web that many links will be broken or outdated at any given time. Whenever someone publishes an incorrect link that points to your site (perhaps through a typo or a spelling error) or fails to update their pages to reflect changes on your server, Googlebot will try to download an incorrect link from your site. This is also why you may get hits on a machine that isn't a web server at all.

4. Why is Googlebot downloading information from our "secret" web server?

It is almost impossible to keep a web server secret by not publishing any links to it. As soon as someone follows a link from your "secret" server to another web server, it is likely that your "secret" URL is in the referer tag, and it can be stored and possibly published by the other web server in its referer log. So, if there is a link to your "secret" web server or page on the web anywhere, it is likely that Googlebot and other "web crawlers" will find it.

5. Why isn't Googlebot obeying my `robots.txt` file?

To save bandwidth, Googlebot only downloads the `robots.txt` file once a day or whenever we have fetched many pages from the server. So, it may take a while for Googlebot to learn of any changes that might have been made to your `robots.txt` file. Also, Googlebot is distributed on several machines. Each of these keeps its own record of your `robots.txt` file. Also, check that your syntax is correct against the standard at: <http://www.robotstxt.org/wc/norobots.html>. If there still seems to be a problem, please let us know and we'll correct it.

Please note that there is a small difference between the way Googlebot handles the `robots.txt` file and the way the `robots.txt` standard says we should (keeping in mind the distinction between "should" and "must"). The standard says we should obey the *first* applicable rule, whereas Googlebot obeys the *longest* (that is, the most specific) applicable rule. This more intuitive practice matches what people actually do, and what they expect us to do. For example, consider the following `robots.txt` file:

```
User-Agent: *  
Allow: /
```

```
Disallow: /cgi-bin
```

It's obvious that the webmaster's intent here is to allow robots to crawl everything except the `/cgi-bin` directory. Consequently, that's what we do.

6. How do I register my site with Googlebot so it will be indexed?

Please visit the [Add URL form](#).

7. How do I remove a site from Google?

Google updates its entire index automatically on a regular basis. When we crawl the web, we find new pages, discard dead links, and update links automatically. Links that are outdated now will most likely "fade out" of our index during our next crawl. For detailed information on how to remove or uncache a page in Google, [click here](#).

8. Help! Googlebot is crawling my site too fast. What can I do?

Please send an email to googlebot@google.com with the name of your site and a detailed description of the problem. Please also include a portion of the weblog that shows Google accesses, so we can track down the problem more quickly on our end.

9. Why are there hits from multiple machines at Google.com, all with user-agent Googlebot?

Googlebot was designed to be distributed on several machines to improve performance and to scale as the web grows.

10. Can you tell me the IP addresses from which Googlebot crawls so that I can filter my logs?

The IP addresses used by Googlebot change from time to time. The best way to identify accesses by Googlebot is to use the user-agent (Googlebot).

11. How do I block all crawlers except Googlebot from my site?

The following robots.txt file will achieve this for all well-behaved crawlers.

```
User-agent: *  
Disallow: /  
  
User-agent: Googlebot  
Allow: /
```

12. How do I tell Googlebot not to crawl dynamically generated pages on my site?

The following robots.txt file will achieve this.

```
User-agent: Googlebot  
Disallow: /*?
```

13. My question is not answered here. Where can I send it?

Please visit our [Contact Us](#) page to find the appropriate place to send your question.

For more answers, see the [Robots FAQ](#).