How to Remove Content from Google's Search Index

On occasion, something may end up on your site that you later decide shouldn’t be there. Even if you delete or replace that asset (whether it be an image or an entire page), a copy of it may live on in Google's index for a long time.

If you want to make sure that an image or text from your site isn’t visible in Google, here are the steps you need to take to get that asset or snippet cleared from the search engine’s cache: 1) remove the asset from the site and 2) submit a URL removal request in Google Webmaster Tools .

1. Remove the asset from your site

You can’t simply tell Google to take anything you don’t like out of the index. One of the things Google needs to know before it acts on a request to delete something from the index is that the owner of the site really doesn’t want it to be visible to search engines and that the site’s owner has taken steps to prevent it from surfacing.

There are three different steps you can take to meet this requirement:

- Delete the asset and make sure the associated URL returns a 404 or 410 status code
- Block the content using a robots.txt file
- Block the content using a meta noindex tag (only works for HTML pages, not images)

Only the first option also ensures that no human beings will see it either. The second two options are requests to robots not to include the files in their index – but not all robots honor those requests – and they don’t prevent any human visitors from seeing the content if they know the URL. Also, anyone can look inside a robots.txt file to see all the directories and files you don’t want to expose, so that’s not a great way to hide something you really don’t want anyone to see.

If you don’t want users or search engines to be able to see or find a particular file on your site, then the best thing is to delete it outright and make sure that the server tells the browser the asset no longer exists.

2. Submit a Removal Request to Google

Once you’ve deleted the asset from your site, you can wait for Googlebot to recrawl your site and discover the asset is gone. Eventually it’ll disappear from the cache.

But what if you can’t wait that long?

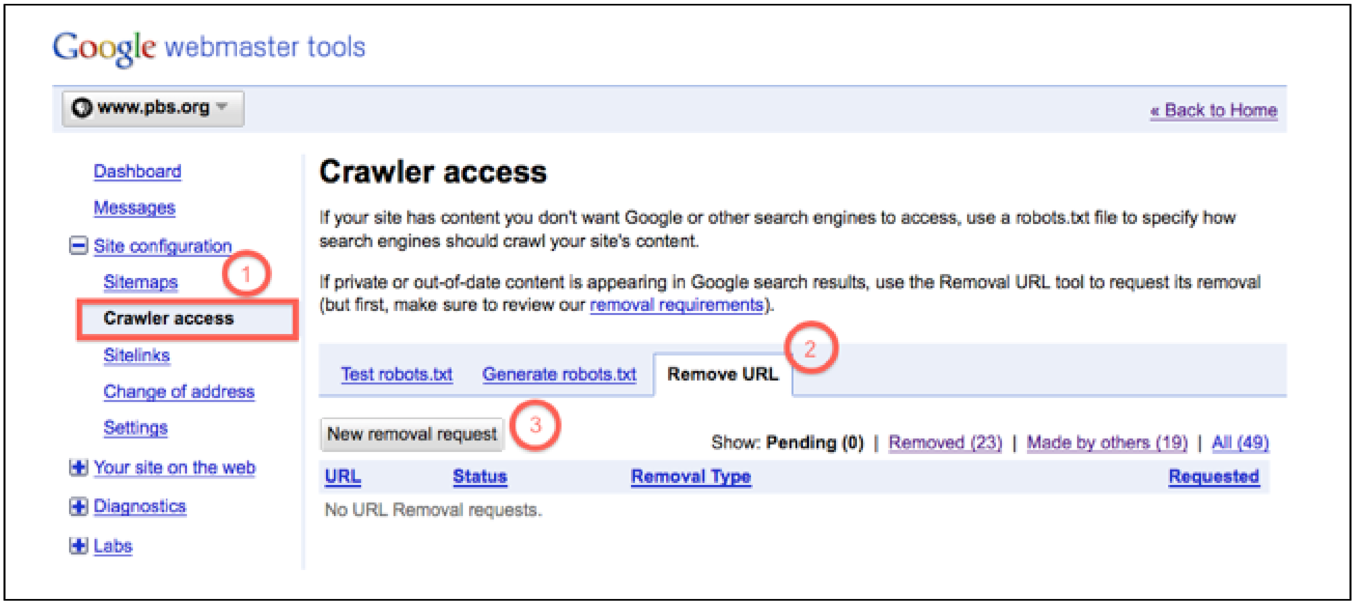
Luckily you can speed things along by logging into your Google Webmaster Tools account and filing a removal request. In our experience, you can get pages removed from Google’s cache in a matter of hours, rather than the days or weeks it could take if you simply waited. This tool is just one of the many reasons why I highly recommend every site owner register a GWT account.

To initiate a removal request:

1. Log into your Google Webmaster Tools Account
2. the main dashboard, click on the site associated with the asset you want to remove (This would be the domain that the asset resides on); if your site isn’t already listed as verified in your account, you must first follow the instructions to verify your site.
3. From the site dashboard, expand “Site Configuration” in the left nav
4. Click on “Crawler access”
5. In the main content area next to the “test robots.txt” and “generate robots.txt” tabs, click on the link to “Remove URL”
6. Click the “New removal request” button
7. Enter the URL of the asset you want to remove from Google.
8. The next page will give you three options:
    - Remove page from search results and cache
    - Remove page from cache only
    - Remove directory

If you’re trying to get all references to a single undesirable asset removed, pick the first option.

Your URL removal request will now show up in the “pending” list until the request has been processed. Check back in a couple hours to see if the request was approved and then go to Google to make sure the page or image is indeed gone. (It’s important to understand that this is an automated process--not a manual review --so be sure to select the appropriate options and follow the instructions).



Crawler access

If your site has content you don't want Google or other search engines to access, use a robots.txt file to specify how search engines should crawl your site's content.

If private or out-of-date content is appearing in Google search results, use the Removal URL tool to request its removal (but first, make sure to review our removal requirements).

No URL Removal requests.

Diagram: Google's Webmaster Tools console
(1) Site Configuration Tools: Crawler Access
(2) Remove URL tab
(3) "New removal request" button

But wait; A small hitch...

As I mentioned before, this process is not for requesting that Google remove something that's actually *on someone else's site* (unless that asset already meets the requirement about returning a 404/410 or is blocked from crawlers).

If you're still seeing your images in Google's search index after you've successfully followed all the steps above, you may actually be seeing a copy of it on a different site or you didn't submit all the right URLs. With web pages, it's very easy to see where Google is pulling the text from (just check the URL below the snippet in the main search results); make sure that what you're looking at is on your site. To see the originating location for *an image* that's cached in Google Image search, click the thumbnail to open up the cached image. Then on the right side of the pane, click "full size image" to see the full URL/file name of that asset. If that's your file, the process outlined above should work.

If the URL of the image is not hosted on your site but someplace else, then you'll need to work with the site owner to have them remove it.

One common example of when this might happen is when your site generates a feed that goes directly into Merlin. When you publish assets for consumption by Merlin, the system does not create a copy of the entire page, but it does create a copy of your images and stores them on a server in order to be able to resize them and make them available in places like the PBS homepage and topics pages.

If you need to change/remove an image from the Google cache that is coming from a Merlin-fed page, here are the steps you take:

- Make sure that the image is no longer in any feeds pulled into Merlin (or it'll just get repopulated with the next feed refresh).
- Work with your PBSi program manager to request that the image be deleted from the Merlin server and that the URL of that asset returns a 404 /410. You will need to provide the full URL location of that image. The URL could look something like this: http://www-tc.pbs.org/s3/pbs.merlin.cdn.prod/webobjects/name-of-image.640x360.jpg (Tip: Just because the image is no longer visible anywhere on the front-end doesn't mean that the image has been deleted from the server. Check that the file returns an error to verify that it meets Google's requirements for removal.)
- Once the Merlin-hosted image is deleted and returns a 404/410, PBSi can put in a removal request to expedite deletion from Google's cache.

Getting Back into the Index After a Removal Request

If you change your mind later, you can always republish the URL or image and then click the "reinclude" link in GWT to request a recrawl and reinclusion of that specific asset.

Let us know if you have any questions about these instructions or they don't work for you.