Link Checking with Google Console

Introduction

Google Search Console is a free service that lets you learn a great deal of information about your website and the people who visit it. You can use it to find out things like how many people are visiting your site and how they are finding it, whether more people are visiting your site on a mobile device or desktop computer, and which pages on your site are the most popular. It can also help you find and fix website errors,

I have divided the document into a brief overview of checking broken links with Google Search Console and then a more in-depth look into the information available. Content in this document has been compiled from Google help documents.

Brief Overview of Link Checker with Google Search Console

Finding broken links on your site can be tedious (although it's valuable to run a broken links checker on your site in any case to ensure you're providing the best user experience possible) and it would be nice to be able to fix broken external links at the source, rather than implementing all those redirects. Now fixing the source has gotten a lot easier. Beside each broken link listed in the 404 report will be the source URL for that link. You can download the report into Excel and sort it by the source URLs to get a list of all the internal broken links so you can easily fix them. You'll then also have a list of all the external sites with broken links to your pages. You can contact the site owners and ask for the links to be fixed (which will help the user experience for their visitors as well). You won't be able to get all external links fixed, so for the rest, you can continue implementing redirects.

Search Console is divided into two main sections: Site Errors and URL Errors.

Categorizing errors in this way is pretty helpful because there's a distinct difference between errors at the site level and errors at the page level.

1. Site-level issues can be more catastrophic, with the potential to damage your site's overall usability.
2. URL errors, on the other hand, are specific to individual pages, and are therefore less urgent.

At the extreme minimum, you should check at least every 90 days to look for previous errors so you can keep an eye out for them in the future — but frequent, regular checks are best.

- DNS Errors, DNS (Domain Name System) errors are the first and most prominent error because if the Googlebot is having DNS issues, it means it can't connect with your domain via a DNS timeout issue or DNS lookup issue.

- A server error most often means that your server is taking too long to respond, and the request times out. The Googlebot that's trying to crawl your site can only wait a certain amount of time to load your website before it gives up. If it takes too long, the Googlebot will stop trying.

- Server errors are different than DNS errors. A DNS error means the Googlebot can't even lookup your URL because of DNS issues, while server errors mean that although the Googlebot can connect to your site, it can't load the page because of server errors.

- A Robots failure means that the Googlebot cannot retrieve your robots.txt file, located at [yourdomain.com]/robots.txt.

Server Error	When you see this kind of error for your URLs, it means that Googlebot couldn't access your URL, the request timed out, or your site was busy. As a result, Googlebot was forced to abandon the request.
Soft 404	A soft 404 occurs when your server returns a real page for a URL that doesn't actually exist on your site. This usually happens when your server handles faulty or non-existent URLs as "OK," and redirects the user to a valid page like the home page or a "custom" 404 page.

Access Denied	Googlebot couldn't access a URL on your site because your site requires users to log in to view all or some of your content.
	Your server requires users to authenticate using a proxy, or your hosting provider may be blocking Google from accessing your site.
Not Found	Most 404 errors don't affect your site's ranking in Google, so you can safely ignore them. Typically, they are caused by typos, site misconfigurations, or by Google's increased efforts to recognize and crawl links in embedded content such as JavaScript.
Other	Google was unable to crawl this URL due to an undetermined issue.

Google Search Console Error Check In Detail

The report has two main sections:

Site errors: This section of the report shows the main issues for the past 90 days that prevented Googlebot from accessing your entire site. (Click any box to display its chart.)

URL errors: This section lists specific errors Google encountered when trying to crawl specific desktop or phone pages. Each main section in the URL Errors reports corresponds to the different crawling mechanisms Google uses to access your pages, and the errors listed are specific to those kinds of pages.

Site errors overview

In a well-operating site, the Site errors section of the Crawl Errors report should show no errors (this is true for the large majority of the sites we crawl). If Google detects any appreciable number of site errors, we'll try to notify you in the form of a message, regardless of the size of your site.

When you first view the Crawl Errors page, the Site errors section shows a quick status code next to the each of the three error types: DNS, Server connectivity, and robots.txt fetch. If the codes are anything other than a green check mark, you can click the box to see a graph of crawling details for the last 90 days.

High error rates

If your site shows a 100% error rate any of the three categories, it likely means that your site is either down or misconfigured in some way. This could be due to a number of possibilities that you can investigate:

Check that a site reorganization hasn't changed permissions for a section of your site.
If your site has been reorganized, check that external links still work.
Review any new scripts to ensure they are not malfunctioning repeatedly.
Make sure all directories are present and haven't been accidentally moved or deleted.

If none of these situations apply to your site, the error rate might just be a transient spike, or due to external causes (someone has linked to non-existent pages), so there might not even be a problem. In any case, when we see an unusually large number of errors for your site, we'll let you know so you can investigate.

Low error rates

If your site has an error rate less than 100% in any of the categories, it could just indicate a transient condition, but it could also mean that your site is overloaded or improperly configured. You might want to investigate these issues further, or ask about them on our forum. We might alert you even if the overall error rate is very low — in our experience, a well configured site shouldn't have any errors in these categories.

Site error types

The following errors are exposed in the Site section of the report:

DNS Errors

Server errors

What is a server error?

When you see this kind of error for your URLs, it means that Googlebot couldn't access your URL, the request timed out, or your site was busy. As a result, Googlebot was forced to abandon the request.

Fixing server connectivity errors

Reduce excessive page loading for dynamic page requests.
A site that delivers the same content for multiple URLs is considered to deliver content dynamically (e.g. www.example.com/shoes.php?color=red&size=7 serves the same content as www.example.com/shoes.php?size=7&color=red). Dynamic pages can take too long to respond, resulting in timeout issues. Or, the server might return an overloaded status to ask Googlebot to crawl the site more slowly. In general, we recommend keeping parameters short and using them sparingly. If you're confident about how parameters work for your site, you can tell Google how we should handle these parameters.
Make sure your site's hosting server is not down, overloaded, or misconfigured.
If connection, timeout or response problems persists, check with your web hoster and consider increasing your site's ability to handle traffic.
Check that you are not inadvertently blocking Google.
You might be blocking Google due to a system level issue, such as a DNS configuration issue, a misconfigured firewall or DoS protection system, or a content management system configuration. Protection systems are an important part of good hosting and are often configured to automatically block unusually high levels of server requests. However, because Googlebot often makes more requests than a human user, it can trigger these protection systems, causing them to block Googlebot and prevent it from crawling your website. To fix such issues, identify which part of your website's infrastructure is blocking Googlebot and remove the block. The firewall may not be under your control, so you may need to discuss this with your hosting provider.
Control search engine site crawling and indexing wisely.
Some webmasters intentionally prevent Googlebot from reaching their websites, perhaps using a firewall as described above. In these cases, usually the intent is not to entirely block Googlebot, but to control how the site is crawled and indexed. If this applies to you, check the following:
> To control Googlebot's crawling of your content, use the robots exclusion protocol, including using a robots.txt file and configuring URL parameters.
> If you're worried about rogue bots using the Googlebot user-agent, you can verify whether a crawler is actually Googlebot.
If you would like to change how frequently Googlebot crawls your site, you can request a change in Googlebot's crawl rate. Hosting providers can verify ownership of their IP addresses too.

Server connectivity errors

Error Type	Description
Timeout	The server timed out waiting for the request. Use Fetch as Google to check if Googlebot can currently crawl your site. If Fetch as Google returns the content of your homepage without problems, you can assume that Google is generally able to access your site properly. It's possible that your server is overloaded or misconfigured. If the problem persists, check with your hosting provider.
Truncated headers	Google was able to connect to your server, but it closed the connection before full headers were sent. Please check back later. Use Fetch as Google to check if Googlebot can currently crawl your site. If Fetch as Google returns the content of your homepage without problems, you can assume that Google is generally able to access your site properly. It's possible that your server is overloaded or misconfigured. If the problem persists, check with your hosting provider.
Connection reset	Your server successfully processed Google's request, but isn't returning any content because the connection with the server was reset. Please check back later. Use Fetch as Google to check if Googlebot can currently crawl your site. If Fetch as Google returns the content of your homepage without problems, you can assume that Google is generally able to access your site properly. It's possible that your server is overloaded or misconfigured. If the problem persists, check with your hosting provider.
Truncated response	Your server closed the connection before we could receive a full response, and the body of the response appears to be truncated. Use Fetch as Google to check if Googlebot can currently crawl your site. If Fetch as Google returns the content of your homepage without problems, you can assume that Google is generally able to access your site properly. It's possible that your server is overloaded or misconfigured. If the problem persists, check with your hosting provider.
Connection refused	Google couldn't access your site because your server refused the connection. Your hosting provider may be blocking Googlebot, or there may be a problem with the configuration of their firewall. Use Fetch as Google to check if Googlebot can currently crawl your site. If Fetch as Google returns the content of your homepage without problems, you can assume that Google is generally able to access your site properly. It's possible that your server is overloaded or misconfigured. If the problem persists, check with your hosting provider.
Connect failed	Google wasn't able to connect to your server because the network is unreachable or down. It's possible that your server is overloaded or misconfigured. If the problem persists, check with your hosting provider. Use Fetch as Google to check if Googlebot can currently crawl your site. If Fetch as Google returns the content of your homepage without problems, you can assume that Google is generally able to access your site properly.

| Connect timeout | Google was unable to connect to your server.

Use Fetch as Google to check if Googlebot can currently crawl your site. If Fetch as Google returns the content of your homepage without problems, you can assume that Googlebot is generally able to access your site properly.

Check that your server is connected to the Internet. It's also possible that your server is overloaded or misconfigured. If the problem persists, check with your hosting provider. |
|---|---|
| No response | Google was able to connect to your server, but the connection was closed before the server sent any data.

Use Fetch as Google to check if Googlebot can currently crawl your site. If Fetch as Google returns the content of your homepage without problems, you can assume that Googlebot is generally able to access your site properly.

It's possible that your server is overloaded or misconfigured. If the problem persists, check with your hosting provider. |

Robots failure

URL errors overview

The URL errors section of the report is divided into categories that show the top 1,000 URL errors specific to that category. Not every error that you see in this section requires attention on your part, but it's important that you monitor this section for errors that can have a negative impact on your users and on Google crawlers. We've made this easier for you by ranking the most important issues at the top, based on factors such as the number of errors and pages that reference the URL. Specifically, you'll want to consider the following:

> Fix Not Found errors for important URLs with 301 redirects. While it's normal to have Not Found (404) errors, you'll want to address errors for important pages linked to by other sites, older URLs you had in your sitemap and have since deleted, misspelled URLs for important pages, or URLs of popular pages that no longer exist on your site. This way, the information that you care about can be easily accessed by Google and your visitors.
> Update your sitemaps. Prune old URLs from your sitemaps, and if you add newer sitemaps that you intend to replace older ones, be sure to delete the old site map (not redirect it to the newer one).
> Keep redirects clean and short. If you have a number of URLs that redirect in a sequence (e.g. pageA > pageB > pageC > pageD), it can be challenging for Googlebot to follow and interpret the sequence. Try to keep the "hops" to a low number. Read more about Not followed.

Viewing URL error details

You can view URL errors in a variety of ways:

> Click Download to retrieve a list of the top 1,000 errors for that crawler type (e.g. desktop, smartphone).
> Use the filter above the table to locate specific URLs.
> See error details by following the link from individual URLs or Application URIs.

The Desktop and Smartphone tabs list URLs that produce crawl errors, as well as the status of the error, a list of pages that reference the URL, and a link to Fetch as Google so you can troubleshoot problems with that URL.

Mark URL errors as fixed

Once you've addressed the issue causing an error for a specific item, you can hide it from the list. You can do this singly or in bulk. Select the checkbox next to the URL, and click Mark as fixed. The URL will be removed from the list. However, this marking is just a convenience method for you; if Google's crawler encounters the error on the next crawl, the URL will reappear in the list the next time your URL is crawled.

URL error types

Common URL errors

Error Type	Description

Server error	When you see this kind of error for your URLs, it means that Googlebot couldn't access your URL, the request timed out, or your site was busy. As a result, Googlebot was forced to abandon the request. Read more about server connectivity errors.
Soft 404	Usually, when a visitor requests a page on your site that doesn't exist, a web server returns a 404 (not found) error. This HTTP response code clearly tells both browsers and search engines that the page doesn't exist. As a result, the content of the page (if any) won't be crawled or indexed by search engines. A soft 404 occurs when your server returns a real page for a URL that doesn't actually exist on your site. This usually happens when your server handles faulty or non-existent URLs as "OK," and redirects the user to a valid page like the home page or a "custom" 404 page. This is a problem because search engines might spend much of their time crawling and indexing non-existent, often duplicative URLs on your site. This can negatively impact your site's crawl coverage because your real, unique URLs might not be discovered as quickly or visited as frequently due to the time Googlebot spends on non-existent pages. If your page is truly gone and has no replacement, we recommend that you configure your server to always return either a 404 (Not found) or a 410 (Gone) response code in response to a request for a non-existing page. You can improve your visitors' experience by setting up a custom 404 page when returning a 404 response code. For example, you could create a page containing a list of your most popular pages, or a link to your home page, or a feedback link. But it's important to remember that it's not enough to just create a page that displays a 404 message. You also need to return the correct 404 or 410 HTTP response code.

404	Googlebot requested a URL that doesn't exist on your site.
	
	Fixing 404 errors
	
	Most 404 errors don't affect your site's ranking in Google, so you can safely ignore them. Typically, they are caused by typos, site misconfigurations, or by Google's increased efforts to recognize and crawl links in embedded content such as JavaScript. Here are some pointers to help you investigate and fix 404 errors:
	
	1. Decide if it's worth fixing. Many (most?) 404 errors are not worth fixing. Here's why:Sort your 404s by priority and fix the ones that need to be fixed. You can ignore the other ones, because 404s don't harm your site's indexing or ranking.
	  • If it is a deleted page that has no replacement or equivalent, returning a 404 is the right thing to do.
	  • If it is a bad URL generated by a script, or that never have existed on your site, it's probably not a problem you need to worry about. It might bother you to see it on your report, but you don't need to fix it, unless the URL is a commonly misspelled link (see below).
	2. See where the invalid links live. Click a URL to see Linked from these pages information. Your fix will depend on whether the link is coming from your own or from another site:
	  • Fix links from your own site to missing pages, or delete them if appropriate.
	    If the content has moved, add a redirect.
	    If you have permanently deleted content without intending to replace it with newer, related content, let the old URL return a 404 or 410. Currently Google treats 410s (Gone) the same as 404s (Not found). Returning a code other than 404 or 410 for a non-existent page (or redirecting users to another page, such as the homepage, instead of returning a 404) can be problematic. Such pages are called soft 404s, and can be confusing to both users and search engines.
	    If the URL is unknown: You might occasionally see 404 errors for URLs that never existed on your site. These unexpected URLs might be generated by Googlebot trying to follow links found in JavaScript, Flash files, or other embedded content, or possibly that exist only in a sitemap. For example, your site may use code like this to track file downloads in Google Analytics:
	    <a href="helloworld.pdf"
	    onClick="_gaq.push(['_trackPageview','/download-helloworld']);">
	    Hello World PDF
	    When Googlebot sees this code, it might try to crawl the URL http://www.example.com/download-helloworld, even though it's not a real page. In this case, the link may appear as a 404 (Not Found) error in the Crawl Errors report. Google is working to prevent this type of crawl error. This error has no effect on the crawling or ranking of your site.
	  • Fix misspelled links from other sites with 301 redirects. For example, a misspelling of a legitimate URL (www.example.com/redshoos instead of www.example.com/redshoes) probably happened when someone linking to your site simply made a typo. In this case, you can capture that misspelled URL by creating a 301 redirect to the correct URL. You can also contact the webmaster of a site with an incorrect link, and ask for the link to be updated or removed.
	3. Ignore the rest of the errors. Don't create fake content, redirect to your homepage, or use robots.txt to block those URLs—all of these things make it harder for us to recognize your site's structure and process it properly. We call these soft 404 errors. Note that clicking This issue is fixed in the Crawl Errors report only temporarily hides the 404 error; the error will reappear the next time Google tries to crawl that URL. (Once Google has successfully crawled a URL, it can try to crawl that URL forever. Issuing a 300-level redirect will delay the recrawl attempt, possibly for a very long time.) Note that submitting a URL removal request using the URL removal tool will not remove the error from this report.
	
	If you don't recognize a URL on your site, you can ignore it. These errors occur when someone browses to a non-existent URL on your site - perhaps someone mistyped a URL in the browser, or someone mistyped a link URL. However, you might want to catch some of these mistyped URLs as described in the list above.

Access denied	In general, Google discovers content by following links from one page to another. To crawl a page, Googlebot must be able to access it. If you're seeing unexpected Access Denied errors, it may be for the following reasons: Googlebot couldn't access a URL on your site because your site requires users to log in to view all or some of your content. Your server requires users to authenticate using a proxy, or your hosting provider may be blocking Google from accessing your site. To fix: Test that your robots.txt is working as expected and does not block Google. The Test robots.txt tool lets you see exactly how Googlebot will interpret the contents of your robots.txt file. The Google user-agent is Googlebot. Use Fetch as Google to understand exactly how your site appears to Googlebot. This can be very useful when troubleshooting problems with your site's content or discoverability in search results.
Not followed	Not followed errors lists URLs that Google could not completely follow, along with some information as to why. Here are some reasons why Googlebot may not have been able to follow URLs on your site: Flash, JavaScript, active content Some features such as JavaScript, cookies, session IDs, frames, DHTML, or Flash can make it difficult for search engines to crawl your site. Check the following: Use a text browser such as Lynx to examine your site, since many search engines see your site much as Lynx would. If features such as JavaScript, cookies, session IDs, frames, DHTML, or Flash keep you from seeing all of your site in a text browser, then search engine spiders may have trouble crawling your site. Use Fetch as Google to see exactly how your site appears to Google. If you use dynamic pages (for instance, if your URL contains a ? character), be aware that not all search engine spiders crawl dynamic and static pages. In general, we recommend keeping parameters short and using them sparingly. If you're confident about how parameters work for your site, you can tell Google how we should handle them. Redirects If you are permanently redirecting from one page to another, make sure you're returning the right HTTP status code (301 Moved Permanently). Where possible, use absolute rather than relative links. (For instance, when linking to another page in your site, link to www.example.com/mypage.html rather than simply mypage.html). Try to make every page on your site reachable from at least one static text link. In general, minimize the number of redirects needed to follow a link from one page to another. Check your redirects point to the right pages! Sometimes we discover redirects that point to themselves (resulting in a loop error) or to invalid URLs. Don't include redirected URLs in your Sitemaps. Keep your URLs as short as possible. Make sure you aren't automatically appending information (such as session IDs) to your redirect URLs. Make sure your site allows search bots to crawl your site without session IDs or arguments that track their path through the site.
DNS error	When you see this error for URLs, it means that Googlebot could either not communicate with the DNS server, or your server had no entry for your site. Read more about DNS errors.