knowledgeowl

Broken link checker & report

Last Modified on 02/18/2025 2:31 pm EST

The Broken Links Report is located in the **Tools** menu. You can use this report to clean up your broken links and rerun it to be sure they are fixed.

Our default **Editor** and **Writer** roles have permission to generate the Broken Links Report. If you're using a custom author role, that role must have the Tools Permission to **Run Broken Links Report**.

This report will scan all articles with Published (or Needs Review) publishing statuses in your knowledge base to check for broken hyperlinks and produce a Broken Links Report CSV file of all detected broken links. It includes most hyperlinks across your knowledge base, though there are some links it won't check.

You can refine the report to include additional content, such as:

- "Draft" articles
- "Archived" articles
- Article Versions (this will include all non-active versions in the checks, including historical versions)

The report can only check links that are publicly available. If you have hyperlinks to resources behind a company firewall, VPN, or other login, those links will show up in the report even though you might have no trouble accessing them. You may want to ignore 401 or 403 codes if you're seeing a lot of these.

By default, the report will ignore links that trigger a 301 or 302 code, since these codes indicate a redirection and a successful ultimate load. Not familiar with HTTP status codes? See Exclude codes for a quick primer!

Once the report has been generated, it remains available in **Tools > Broken Link Checker** until you generate a new one, so you can also start this report, go do other things, and then come back and download it later.

The checker will only allow one report per knowledge base to be generated at a time, so if you have multiple authors and one person starts it, everyone else will be prevented from running a new report until that report has finished generating.

See the links below for more information on generating a report, how to read it, and tips and tricks for handling odd situations!

Generating a Broken Links Report

Our default **Editor** and **Writer** roles have permission to generate the Broken Links Report. If you're using a custom author role, that role must have the Tools Permission to **Run Broken Links Report**.

To generate a Broken Links Report:

1. Go to **Tools > Broken Link Checker**.
2. If you'd like to include draft or archived article statuses or article versions, check the appropriate boxes in the Additional content options section.
3. If you'd like to ignore specific status codes (or include 301 and 302 response codes), check the appropriate boxes in the Exclude codes section.
4. Once you've finished making selections, select **Generate Report**.
5. The checker will display a detailed progress bar as it's running, letting you know as it checks content.
6. Once the checker has completed, it will display a success message and a **Download** button will appear next to the **Generate Report** button. The timestamp for the "Last report ran on..." statement will also update.
7. Select **Download** to download the Broken Links Report in CSV format.

You can then peruse the report and update content as you see fit. See How to read the Broken Links Report for more information.

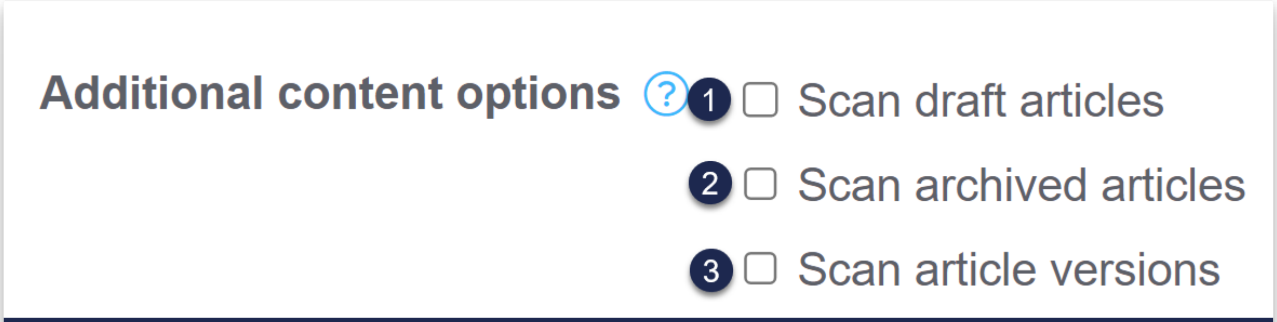
Once the report has been generated, it remains available in **Tools > Broken Link Checker** until you generate a new one, so you can also start this report, go do other things, and then come back and download it later.

The checker will only allow one report per knowledge base to be generated at a time, so if you have multiple authors and one person starts it, everyone else will be prevented from running a new report until that report has finished generating.

Additional content options

By default, the Broken Links Report will check all articles with either a Published or Needs Review status.

Use the checkboxes in the **Additional content options** section to include additional content:


Additional content options
?
1 □ Scan draft articles
2 □ Scan archived articles
3 □ Scan article versions

Additional content options

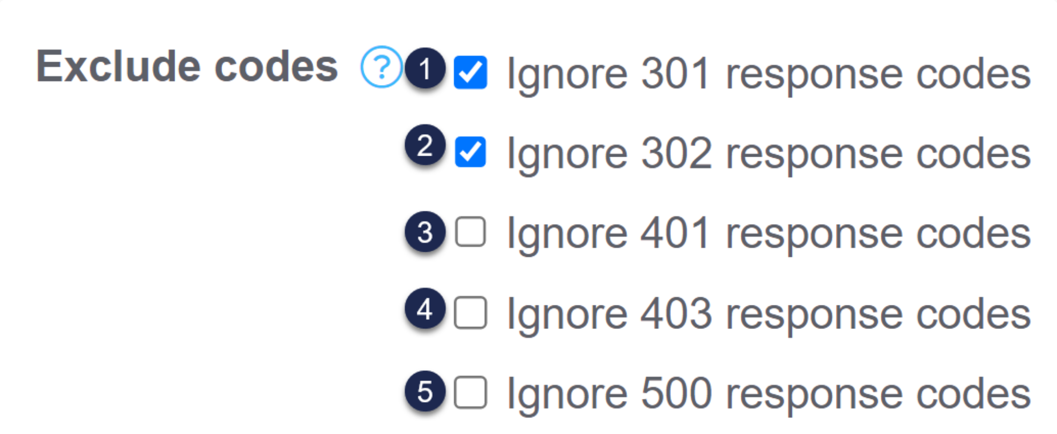
1. **Scan draft articles:** Include articles with any of the three draft statuses (Draft, Ready to Publish, Rejected Draft) in the report

2. **Scan archived articles:** Include articles with the Archived status in the report
3. **Scan article versions:** Include non-active versions of articles with the designated statuses in the report
    - This will include historical versions

Exclude codes

The Broken Link Checker will check for HTTP status codes at the 3xx, 4xx, and 5xx level.

You have the option to ignore certain codes:



Exclude codes ?

1 ☑ Ignore 301 response codes

2 ☑ Ignore 302 response codes

3 □ Ignore 401 response codes

4 □ Ignore 403 response codes

5 □ Ignore 500 response codes

Exclude codes checkboxes

Not sure which codes you need to worry about? We recommend using the default settings in this section to ignore 301 and 302 response codes and include all the rest.

1. **Ignore 301 response codes:** 301 status code means a resource has a permanent redirect in place.
    - 301 codes are generally safe to ignore, since they indicate a redirect but ultimately a successful resource load.
    - The box to ignore 301 codes is checked by default when you open this page.
    - See the 3xx section below for more detail.
2. **Ignore 302 response codes:** Like 301, a 302 status code means a resource has a redirect in place, though it's a temporary one.
    - 302 status codes are generally safe to ignore, since they indicate a redirect but ultimately a successful resource load.
    - The box to ignore 302 codes is checked by default when you open this page.

   - If you've used Old Links, you may see 302 redirects appear if you have hyperlinks to the old link. While you don't have to resolve these, it doesn't hurt to tidy them up. :)
   - See the 3xx section below for more detail.

3. **Ignore 401 response codes:** 401 status code means "Unauthorized". It's specifically used when a link or resource requires authentication and authentication has either failed or hasn't been provided.
   - We recommend including 401 response codes the first time you run a report. We often end up ignoring 403 codes in subsequent runs.
   - If you're referencing protected resources that you have access to but KnowledgeOwl doesn't (such as on a company intranet or behind a VPN), those resources might throw a 401 error code.
   - As long as you've confirmed that you and your readers can access these resources, you may choose to ignore this code when you rerun the report.
   - See the 4xx section below for more detail.
4. **Ignore 403 response codes:** 403 status code means "Forbidden." It's used when a resource requires authentication but the authentication provided does not grant the user access (such as lacking appropriate permissions).
   - We recommend including 403 response codes the first time you run a report. We often end up ignoring 403 codes in subsequent runs.
   - If you're referencing protected resources that you have access to but KnowledgeOwl doesn't (such as on a company intranet or behind a VPN), those resources might throw a 403 error code.
   - As long as you've confirmed that you and your readers can access these resources, you may choose to ignore this code when you rerun the report.
   - See the 4xx section below for more detail.
5. **Ignore 500 response codes:** 500 status code means "Internal Server Error". It's a catch-all error when a resource can't be accessed but there isn't a more specific reason.
   - We recommend including 500 codes the first time you run the report and only ignoring them if they feel noisy.

> If you're seeing a particular code show up that you'd like to ignore, contact us and let us know which code it is. We started with some we ran into while testing but imagine there may be other "noisy" codes people would like to remove!

More detail on HTTP status codes

If you're not very familiar with HTTP status codes, here's a general breakdown on what they cover:

- **3xx: redirection**: These links are not generally broken but must hop through at least one redirection to be opened.
    - By default, we choose to ignore 301 and 302 codes since those are functioning, but you can opt to include them.
    - We include these codes in the report in case you are referencing external resources that have moved and you'd like to identify them and grab the new URL to bypass that redirect step. If your own resources throw one of these errors, you can generally ignore them.
- **4xx: client errors**: These are links that are definitely broken, and you generally want to resolve them.
    - The only error codes you might want to ignore here are 401 and 403 codes, since they don't necessarily require resolution.
    - Both of these codes are generally triggered when a server recognizes a request as valid but the server won't return the resource. These are usually due to authentication issues. This can be caused by the resource being hidden behind a firewall or login that our systems can't access. If you or your readers can successfully access the resource, then you don't need to resolve these error codes, but you may want to ignore these codes when you run the report in the future.
- **5xx: server errors**: These are links that are also definitely broken. These are generally all errors you'd want to resolve.
    - 5xx errors include some old favorites like: 500 Internal Server Error (a generic error to let you know something went wrong); 502 Bad Gateway; 503 Service Unavailable; 504 Gateway Timeout. These errors generally indicate that a given resource can't be found or the overall server itself has issues.

For more information on HTTP status codes, you can check out the formal HTTP spec or the more readable Wikipedia list of HTTP status codes.

Which codes should I worry about?

Broken Links Reports are a lot like SEO audits--you get a large list of things you could fix, but the reality is that a lot of them are mostly functioning and you have to make some decisions about what you'll prioritize fixing. Some sites will *always* throw a particular code (particularly the 300-level codes), and working with your report means figuring out what those are so you can ignore that noise. Here's some general guidance:

In general, 400- and 500-level codes signify broken links (while most 300-levels are redirects that load successfully). But that's not always true.

While you might want to spot-check a few of them, ignoring 401 or 403 codes is likely a safe choice. Most of the 403 errors we investigated seemed to be perfectly functioning pages, and we now use the ignore checkboxes for these codes when we run our own reports.

You may also learn that certain errors seem tied to certain sites and aren't "real" broken links. We've noticed in our own testing that certain domains throw some errors on all URLs, even valid ones. In most cases, this is because the domains or sites block the kinds of automated requests we make to check the status code.

For example: We reference Zapier documentation a lot. Basically every link to Zapier documentation throws a 405 error (405: Method Not Allowed). If we view the pages in our browser, they load fine. But if we make the cURL request that our broken link checker uses (curl -I https://url-to-check.com), the 405 error is again returned. Zapier doesn't support this method of checking URLs. 😕

We now generally ignore 405 codes on Zapier hyperlinks since we know this is an issue.

You'll likely develop your own set of quirks like this as you work with your own report. If there's a status code you'd like us to add to the checkbox options of codes to ignore, contact us and let us know which code and some examples of sites that are throwing it. These reports can be noisy and we're happy to add some more filters to make them a bit quieter as you need!

How to read the Broken Links Report

The Broken Links Report includes eight columns:

- **Object ID**: ID of the object, relevant for individual objects like articles, snippets, and categories. Left blank for KB-wide things like Theme or Homepage.
- **Object Type**: One of 6 options:
    - Article
    - Article Version (only included if article versions are included as content)
    - Category
    - Homepage
    - Snippet
    - Theme
- **App Edit Link**: Editor link to the page that contains the broken link, so you can open it to resolve the issue.
    - For theme sections, this is just the URL to **Settings > Style**.
- **View Link**: Link to the live KB page that contains the broken link, so you can view it and try to open the link to see what happens. Some objects (such as snippets, theme, etc.) do not have a live View Link.
- **Object Field**: Indicates which "field" on the object contains the link. This list varies based on the object; for Theme, this field includes which portion of the Custom HTML is triggering it.
    - Most options are detailed in What content is checked.
- **Link URL**: the URL that's considered broken
- **Link Text**: for URLs included in body or HTML content, the link text is shown. This can help you identify where in the body you need to update the link.
    - Shows N/A for areas where you just assign the URL directly (such as URL redirects, icons, banners, thumbnails).
- **Link Type**: Contains either "Internal" or "External"
    - **Internal** for links that refer to resources within KnowledgeOwl (files, other articles, etc.)
    - **External** for links that refer to websites and resources outside of KnowledgeOwl
- **Status Code**: the code that we're returning that's prompting this link to be included in the report. See Exclude codes for a description of some common status codes.
- **Author**: For objects that have an author (such as articles or some category types), this will display the name of the current designated author.
    - **N/A** is displayed if this information is not tracked (such as for theme)
    - A blank is shown if the author has been deleted since they saved a change.
- **Last Modified Author**: For objects where we track the date and author when modifications are made, this displays the name of the author who last saved changes to the object.
    - **N/A** is displayed if this information is not tracked (such as for theme)
    - A blank is shown if the author has been deleted since they saved a change.

You can work through this report in any number of ways. We've found it useful to open the CSV in Excel or a similar program, filter by the **Object Type** column, and then work through each Object Type in more detail. See Tips & tricks for broken links for more suggestions on working through the report.

What content is checked

Here are all the objects and fields the Broken Link Checker will examine as it creates the Broken Links Report:

- **Article**
  - **Article Content:** all text entered in the editor of the article.
  - **Redirect URL:** If an article has the URL Redirect box checked, we will check that URL.
  - **Thumbnail URL** and **Banner URL:** If an article has a designated thumbnail and banner, we will check those URLs.
- **Category**
  - **Category description:** We'll check the category description field.
  - **Redirect URL:** If a category is set up as a URL Redirect category type, we will check that URL.
  - **Icon URL:** If a category has a designated icon, we'll check that URL.
  - **Thumbnail URL** and **Banner URL:** Topic Display categories and Custom Content categories can also have thumbnails and banners. If these categories have those fields, we'll also check those URLs.
  - **Category Content:** If it's a Custom Content category, we will check all text entered in the editor, just like an article.
- **Snippet**
  - **Snippet Content:** all text entered in the editor of the snippet.
- **Homepage**
  - **Homepage Content:** all text entered in the **Customize > Homepage custom content** editor.
- **Theme**
  - **Custom HTML:** all HTML entered into any of the Custom HTML templates in **Customize > Style (HTML & CSS)**.
  - The report will identify which Custom HTML template is the issue in the Object Field column (404 Error, Article, Body, Homepage, Login, Manage Reader Subscriptions, Restricted Access Page, Right Column, or Top Navigation)
- **Article Versions (if versions are included):**
  - **Version Content:** all text entered in the editor of the article version.

> **No checks for embedded images, videos, or iframes**
> For all **Content** checks, the report WILL NOT check img, video, and iframe sources (src), just straight hyperlinks. So if you've used the editor options to Insert Image or Insert Video, those URLs will not be checked. Refer to What is not checked for more information.

What is not checked

The Broken Link Checker has three known limitations:

1. **Displayed images, videos, and iframes:** For all Content checks, the report will not check the source value; it only checks explicit hyperlinks. So if you've used the editor options to Insert Image or Insert Video, or you've inserted your own iframe, those src URLs will not be checked. If you'd like to see image, video, or iframe

sources included in these reports, please contact us and let us know!

2. **Anchors:** The Broken Links Checker will not check the validity of anchor or other hash portions of a hyperlink. Issues with anchors don't surface as independent HTTP status codes, so we had no way to check these.
   - The Checker will ignore same-page anchors altogether.
   - For external page URLs that include an anchor, it will check the base URL of the page only.
3. **Private content:** If you are linking to resources that require a login of some kind to view (such as within your company intranet, or behind a VPN, or documentation elsewhere that is behind a login page), those links will generally show up in the report with some 400-level status code. While we can validate resources that are stored within KnowledgeOwl, we have no way to pass authorization to check private resources stored elsewhere.
4. **mailto: links:** Since these links aren't a true URL, they cannot be verified through the automated process we use. The report ignores these links.
5. **javascript links:** Since these links usually perform an action rather than hitting a URL, they cannot be verified through the automated process we use. The report ignores these links.
6. **tel: links:** As with the mailto links, these links cannot be verified through the automated process we use. The report ignores these links.

Tips & tricks for broken links

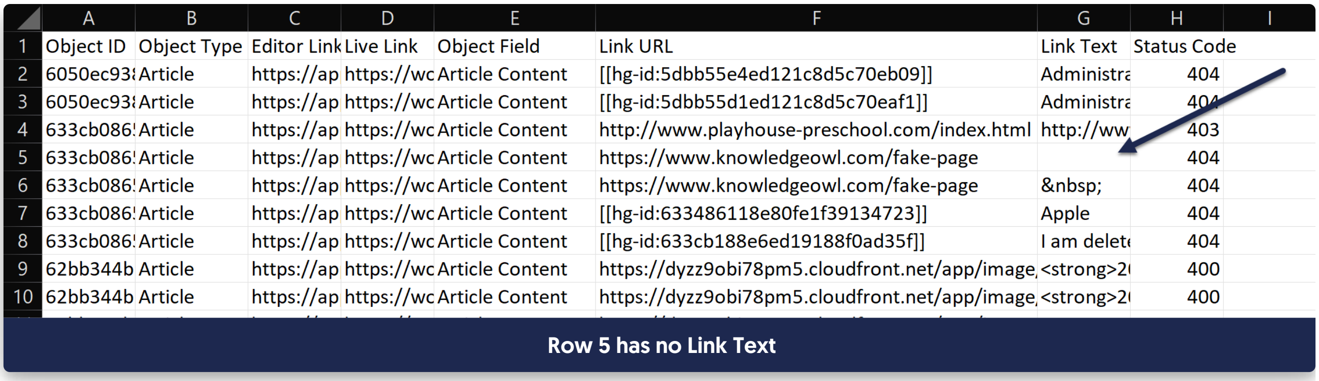
The Broken Links Report can generate a lot of links the first time you run it, and some links won't seem broken. Here are some tips and tricks we've learned while working with these reports.

Use the link text

The Broken Links Report includes the Link Text for all hyperlinks. If you're looking at a broken link that has "Article Content" in the Object Field, the Link Text is your best way to find that link in your article. (We usually use either the Editor Link or Live Link and then use browser search to find the text.)

Article content link has no link text

Some broken links have "Article Content" in the Object Field, but no Link Text, like this:



	A	B	C	D	E	F	G	H	I
1	Object ID	Object Type	Editor Link	Live Link	Object Field	Link URL	Link Text	Status Code	
2	6050ec93	Article	https://ap	https://wc	Article Content	[[hg-id:5dbb55e4ed121c8d5c70eb09]]	Administra	404	
3	6050ec93	Article	https://ap	https://wc	Article Content	[[hg-id:5dbb55d1ed121c8d5c70eaf1]]	Administra	404	
4	633cb086	Article	https://ap	https://wc	Article Content	http://www.playhouse-preschool.com/index.html	http://ww	403	
5	633cb086	Article	https://ap	https://wc	Article Content	https://www.knowledgeowl.com/fake-page		404	
6	633cb086	Article	https://ap	https://wc	Article Content	https://www.knowledgeowl.com/fake-page		404	
7	633cb086	Article	https://ap	https://wc	Article Content	[[hg-id:633486118e80fe1f39134723]]	Apple	404	
8	633cb086	Article	https://ap	https://wc	Article Content	[[hg-id:633cb188e6ed19188f0ad35f]]	I am delet	404	
9	62bb344b	Article	https://ap	https://wc	Article Content	https://dyzz9obi78pm5.cloudfront.net/app/image,	<strong>2	400	
10	62bb344b	Article	https://ap	https://wc	Article Content	https://dyzz9obi78pm5.cloudfront.net/app/image,	<strong>2	400	

Row 5 has no Link Text

The problem with these links is that you have nothing to search for. In fact, since these links have no text, they're not displayed in the live article or in the article WYSIWYG editor.

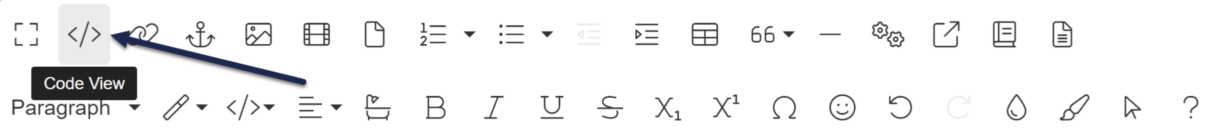
This situation is usually caused by a specific workflow:

- At some point, the link was inserted
- When someone went to remove it, they did not use the **Unlink** option; they just deleted the text itself

With this set of interactions, the WYSIWYG editor does something unusual: it keeps the hyperlink in the code, but just deletes the text from it.

To find these hyperlinks:

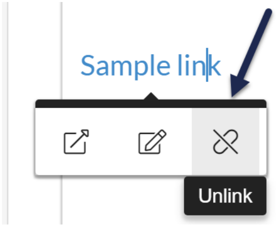
1. Open the article in the editor using the Editor Link provided in the report.
2. In the upper left of the editor, toggle to **Code View** by selecting the </> button in Modern Editor or the **Source** button in Legacy Editor.



3. You should be able to use a browser search for the hyperlink from here, though for really long articles we'll sometimes copy that code view into a text editor and use that search functionality.
4. Generally, these hyperlinks should be deleted (they haven't been visible/used since that text was deleted), though this might also prompt you to revisit if there should be a link here and where it should point.
5. Be sure to **Save** your changes.

To avoid having these types of links show up in your report in the future, we recommend this workflow whenever you remove a hyperlink:

1. Find the text for the link and click on it to open the Hyperlink menu.
2. Select the **Unlink** option.



3. Once the hyperlink has been unlinked, you can delete the text as usual.

Article content link has &nbsp;

&nbsp; is the HTML encoded version of a space, so this is another situation where the hyperlink won't display in the WYSIWYG editor or the live article. Follow the same instructions mentioned above for Article content link has no link text.

LinkedIn link has 999 code

999 is not a "real" error code, and you likely won't see these error codes for pages outside of LinkedIn.

They appear here because LinkedIn does not like automated URL checks, and has set up their pages to return a 999 code when these kinds of link checks are made, regardless of whether the URL is valid or invalid.

For these links, it's probably good to review them the first time you run the report to be sure they seem to work, but we generally filter them out after that since they're not informative at all.

If you're seeing these a lot and would like us to add a checkbox so you can ignore 999 codes, contact us and let us know!

Demio link has 502 code

Links to Demio resources will generally show up with a 502 error in the reports. We contacted Demio and discovered that they don't support the automated checks our broken link checker runs, so these links will always generate a 502. 😕You can choose to skip these in the report or manually check them.

Codes you can probably ignore

In working with our own reports, we have found most pages with these status codes to load fine:

- 307 (a type of redirect, similar to the 301 code we ignore by default)
- 308 (a type of redirect, similar to the 302 code we ignore by default)
- 405

For now, we are including these codes in all reports. If you find that these--or any other error codes--consistently are things you want to ignore, contact us and let us know--we'd be happy to add some more ignore options!

External link has 404 code but loads fine for me

Most of us are trained that a 404 page is an error, so these are often the status codes we prioritize first when reviewing a report.

But there are cases where some external links throw a 404 code in the report but seem to load fine for you when you check them. What's going on?

This can happen for a few different reasons:

- Our Broken Link Checker will only wait 5 seconds when it's checking a URL. If it doesn't get a response within 5 seconds, it will log the URL as a 404 and move on. We do this to try to keep the reports from taking a long time to generate. So if the page loads for you--but it loads slowly--that can be the issue.
- Try testing the URL in an Incognito or Private window. Sometimes, you have accepted a security risk (such as an expired SSL certificate) or otherwise saved a cookie that makes a resource available. Incognito or Private browser windows don't allow cookies, so they should give you the cleanest slate to test with.
- External links are very much at the mercy of availability at the exact time the report was run. If the server the page or resource is hosted on had a small blip, it might trigger a 404 in our report but open just fine.
- Like LinkedIn, the provider who hosts that site might not like automated link checkers and may throw 404s on perfectly valid resources just because they can. We've noticed this with a lot of the Google documentation we link to, all of which loads perfectly fine to a human browser, for example.

Who can run the Broken Links Report?

Our default **Editor** and **Writer** roles have permission to generate the Broken Links Report. If you're using a custom author role, that role must have the Tools Permission to **Run Broken Links Report**.