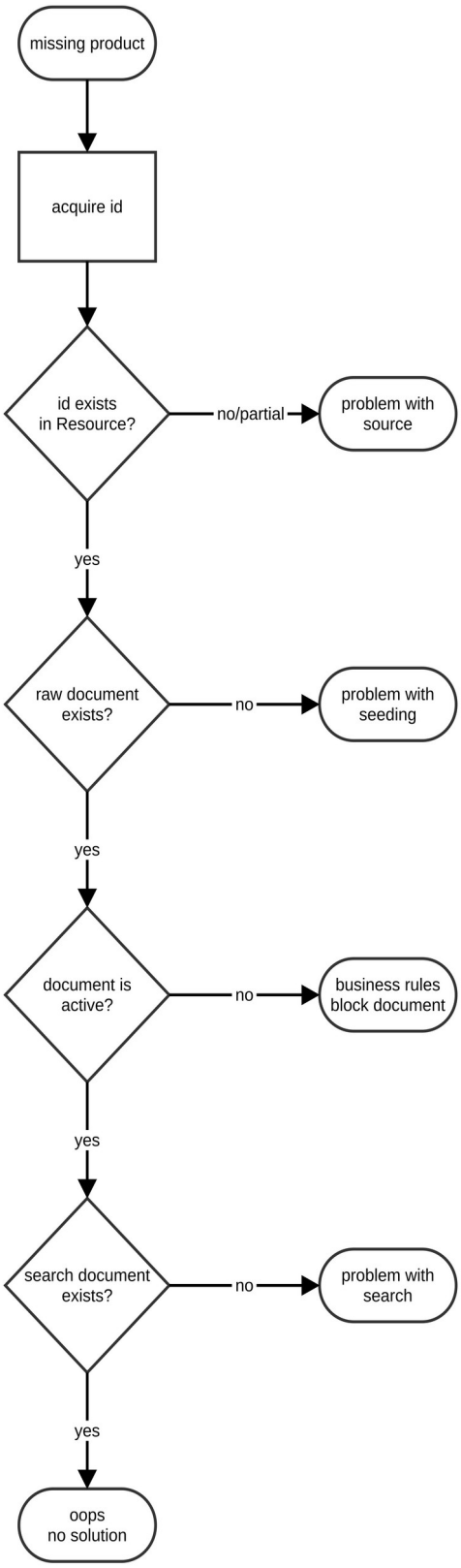
Document troubleshooting

With some regularity we get the question. Why can’t I find document X, Y or Z. Where a document is a research product or learning material. This guide helps to determine what is going on and provides solution directions to remedy the issue.

The flowchart below is a quick reference to what is described in this guide.



1. Aquire the id of a missing document

It's hard to find data based on "I feel like there should be more". It helps a lot to have ids of research products or learning materials that we expected to find, but are not in the system. A piece of text from the title or description also comes a long way. For a missing file the URL can be sufficient. Anything that helps to find concrete examples of what is missing. Now proceed to step 2.

2. Check if id exists in source responses

If we have a way to find the missing piece of information we should start looking for it in the responses that we received from the source systems like: Sharekit, Pure or Edurep. These responses are stored as something that is called a Resource. Most resources can be found in the sources Django app, but Sharekit is an exception with its own app (something we may want to change in the future). Here's a list of Resources in the admin:

- Sharekit (Edusources and Publinova)
- Pure v2 (Hva)
- Hanze
- Pure v1 (BUAS)
- Greeni (VHL)
- HKU
- Edurep
- AnatomyTOOL
- Saxion
- Publinova (is a source as well as a consumer of the output data)

On all the admin pages there is a search bar. Copy your id into the search bar and see if there are any results.

There are no results
The remote system did not return the id in its output to the harvester. Therefor it's not found in the harvester. Check with the remote system what went wrong.

There is one result
Make sure that the result actually contains metadata with the id. It's also possible that the id is only referenced in the metadata of other documents. For instance if Sharekit deletes a publication it keeps around id's of this publication in parent publications (deletion won't delete references in Sharekit). You need to absolutely make sure that the result actually found something that should be added by the harvester instead of finding something that's related to the missing document. Try to pinpoint the exact document that you're looking for by opening the Resource and searching for your id inside the "body" field (Ctrl+F does the trick here). If your id is a piece of text this is a good time to find which id actually belongs to that piece of text. There is always only one id for a document which should be near the text (or other non-unique id) you're searching for. If you found the id proceed to step 4.

There is more than one result
This can happen if your id is not unique (it's a piece of text?) or the document you search for is referenced by other documents (the document is a parent/child to another document?). Follow the text under "there is one result" and try to end up with an exact id that belongs to the missing document. In addition make sure that there is not a mixture of responses where one response adds the document while another removes it again.

3. Check if raw data for document is available

Next up you can search for the id in Products or Files depending on what you're looking for. If you're looking for research output or learning materials then search in Products.

There are no results
Something is wrong with the seeding mechanism. This is a technical issue that can't be resolved by a non-engineer.

There are one or more results
Make sure that at least one SURF Resource Name (SRN) partially matches with the id you're looking for. If at this time you're still looking for a piece of text you should go back to step 2 and try to distill the system id for the document from the external responses in Resources. After making sure proceed to step 4.

4. Check if document is active

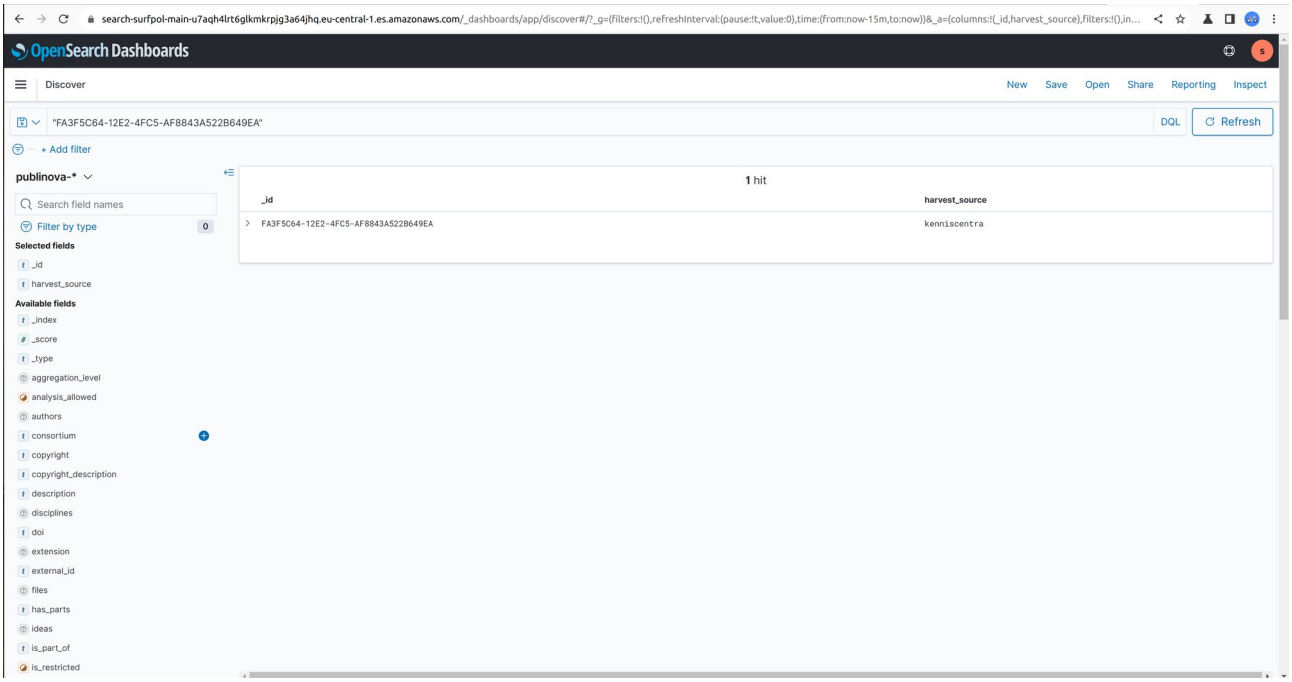
Using the admin pages of step 3. In the results listing there is a column named "state". Note what is written there.

The state is something else than active
There are some business rules that prevent the document from showing up in search. A document can be deleted by the source or by the harvester, because it went missing between harvests. It should be possible to re-create what happened by looking closer at the responses in step 2. The document can also become inactive if it doesn't meet quality standards. In that case the solution is to figure out what in the properties field of the found Document is triggering the "inactive" state. We need better documentation about "business rules" and "quality", but this is largely a non-technical debate and therefor should be in non-technical documentation.

The state is active
Proceed to step 5

5. Check if document exists in search

This step is a bit more tricky to execute. Currently it's only possible to do this for Products, because only Products are added to the search indices. You need to first login to the OpenSearch dashboard. For this you'll need an EduVPN connection and an OpenSearch dashboard user. Then you need to go to "discover" and there you need to select the proper index (publinova-* or edusources-*) instead of the default harvest-logs*. After selecting the correct index you can search for the id between double quotes. I'm adding a screenshot to show how using the dashboard should look:



In this example a single document is found. Based on the amount of results you can troubleshoot the issue.

There are no results
Somehow an active document didn't make it into the search index. This is a technical issue that can't be resolved by a non-engineer.

There is one result
There is nothing wrong with the harvester or search. The problem of missing documents must exist within the frontend application.

There is more than one result
Go back to step 2 and try to find the system id that uniquely identifies a document. It's impossible that multiple results are showing for a truly unique identifier. However these results are likely to be a problem with the steps taken to troubleshoot the problem. The real problem of a missing document is unlikely to be caused anywhere in the harvester or search. The frontend application is probably at fault.