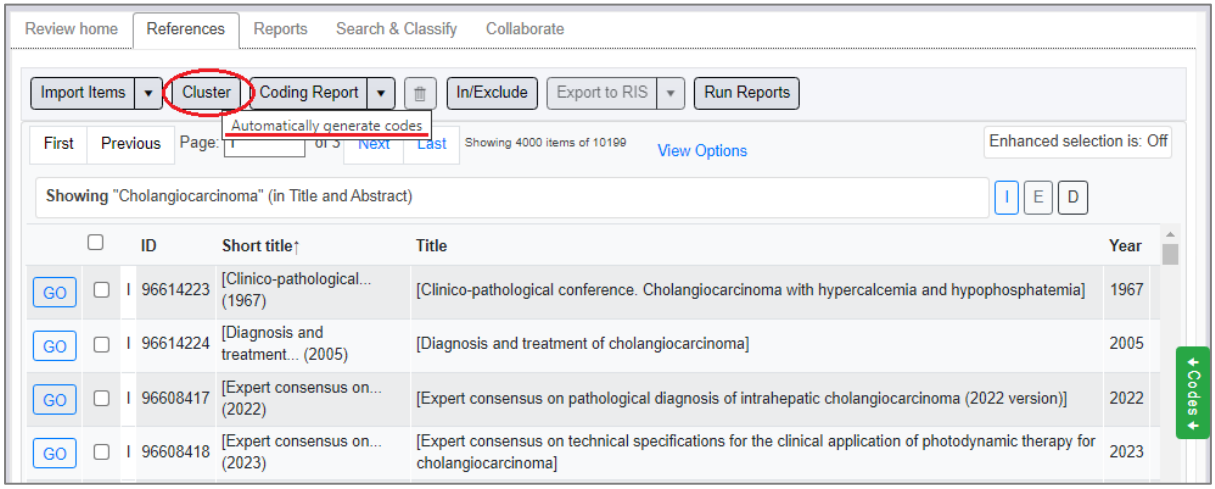
Clustering

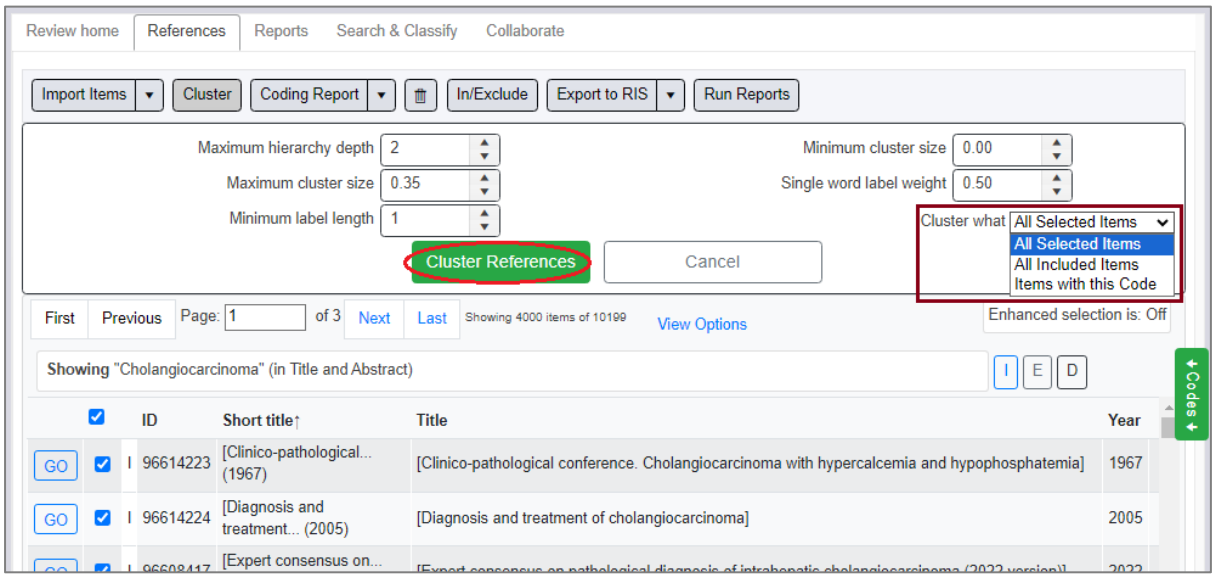
Clustering or Auto-coding in EPPI-Reviewer can categorise your references into clusters, based on the text in their abstracts. This automatic document clustering, using text mining, is one way of describing the range of studies you have identified at the click of a button.

Text mining can assist with searching by identifying significant terms in the documents you have already included. Clustering takes this a step further by taking the extracted terms and coding the items into a tree structure, consisting of the extracted terms that can be searched and manipulated. This is done using the Lingo3G clustering engine.

To auto-code references click on the ***Cluster*** button on the ***References*** tab.



A window will appear where you can set the parameters of the clustering and what items are clustered.

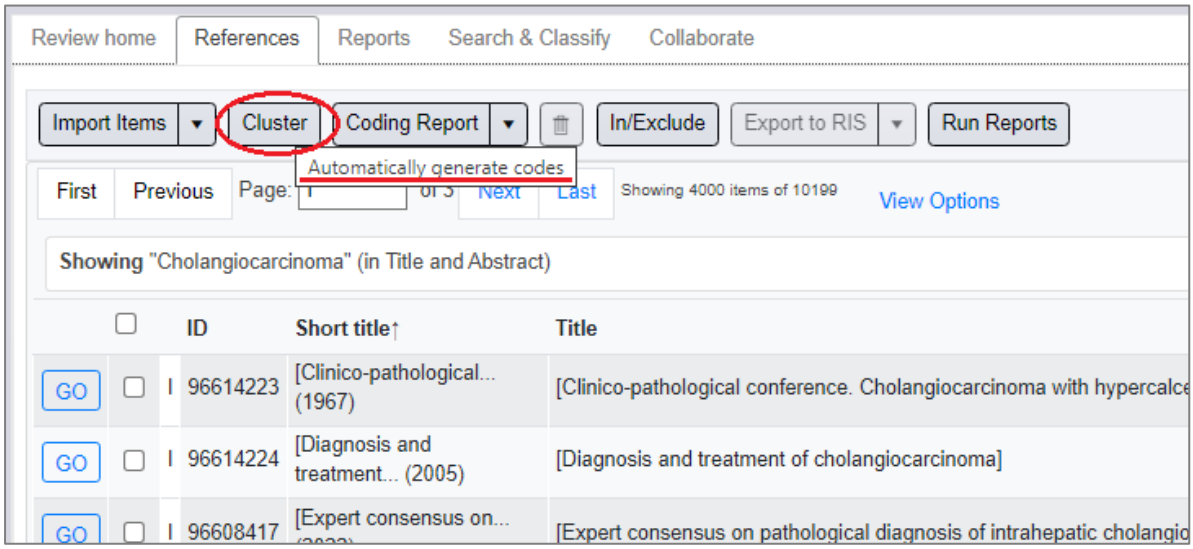


The parameters are -:

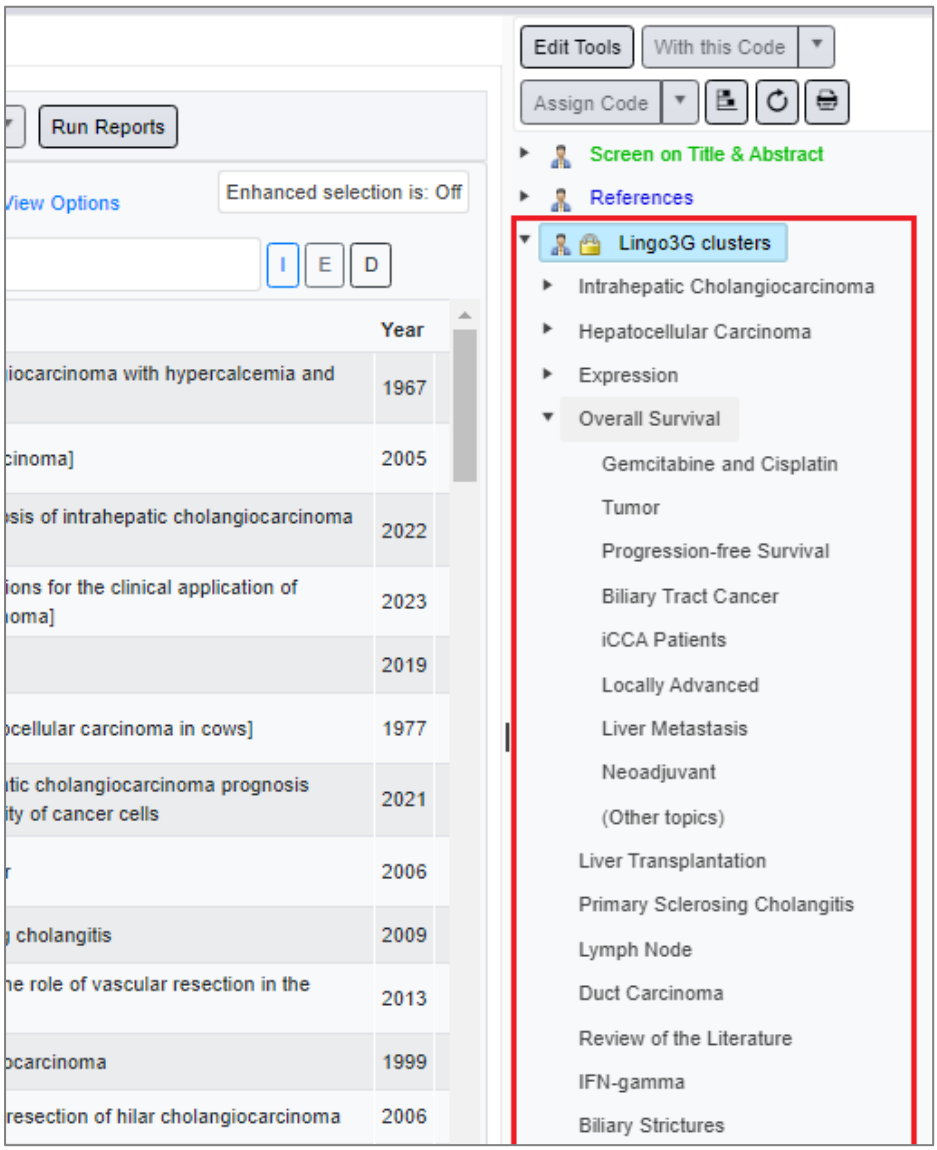
- Maximum hierarchy depth
    - o Default 2.0
    - o Determines how deep in the code tree to go
- Minimum cluster size
    - o Default 0.0 (values 0.0 to 1.0)
- Maximum cluster size
    - o Default 0.35 (values 0.0 1.0)
- Single word label weight
    - o Default 0.5
    - o How the minimum label length is weighted.
- Minimum label length
    - o Default 1.0
    - o The minimum number of words in a cluster

You can either select the items to cluster ahead of time, run it on all included items, or run it on items with a particular code.

Once the options have been set as you wish, click on ***Cluster References*** button.



The newly created codes will appear in the Code Tree on the right of the screen, under a new coding tool called '*Lingo3G clusters*'.

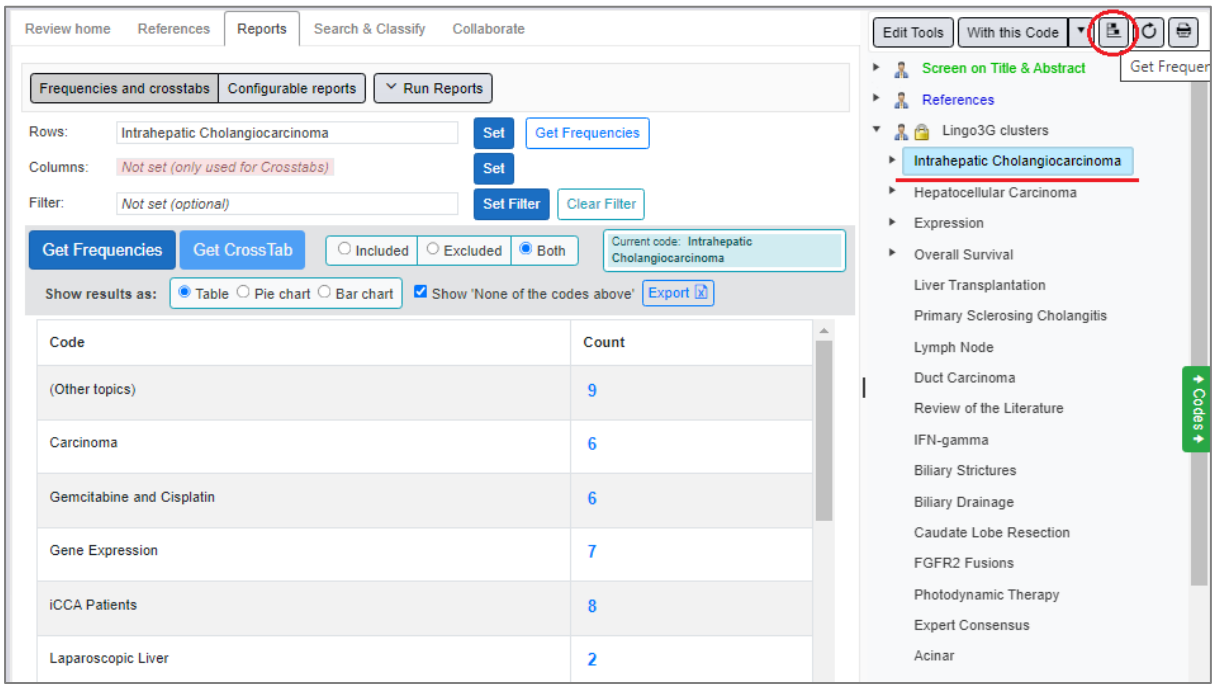


(If you run the clustering tool again, a new set of codes will be generated under the same default coding tool name, so we suggest renaming these if you intend to try multiple uses of the tool.)

Now that the items are coded with the extracted terms the items can be searched and manipulated similar to any other coded item.

This clustering functionality becomes useful when dealing with search results that are in the range of many thousands. It allows you to quickly categorise, or 'Auto-code' the terms contained in the documents and identify the most relevant documents to your subject.

What you may also find useful is running ***Frequency Reports*** on the clustering codesets to see how common the terms found are across your references e.g.



Crosstab reports could also be run to see where items have codes from multiple codesets within the Lingo 3G coding tool.