Corpus Loader User Guide

Author: Hamish Croser
Version: 1.8.3

Contents

Overview

The Corpus Loader is a tool used to build a corpus from a wide range of file types. The tool is designed to be used in conjunction with corpus analysis tools from the Australian Text Analytics Platform.
The Corpus Loader features a file loader, where users can upload files to the notebook environment and load them as a corpus. There is also an Oni loader feature, which enables loading a corpus directly from a platform running the data archiving service Oni.

File loader

Preparing your data

The Corpus Loader accepts the following file types:
TXT, DOCX, ODT, CSV, TSV, XLSX, ODS, XML
Files of the above file types can also be archived into a ZIP file and loaded.

When loading files, the Corpus Loader will load textual and tabular files differently.

- Textual files (TXT, DOCX, ODT, XML) will have their text content read in as a document with no consideration for their internal structure. This means that no metadata can be extracted from the contents of the file, just the file name and path.
- Tabular files (CSV, TSV, XLSX, ODS) will have their table-like structure preserved when loading. Once loaded, you can include/exclude metadata columns from the final built corpus, and select a column to be used as the document column

The Corpus Loader allows you to load a set of textual files as a corpus with no metadata (e.g. a collection of TXT files) and separately a tabular file for metadata (e.g. a CSV). Crucially, the metadata file must contain a linking column, which is a metadata column used to link each metadata row to a corpus document.
For textual files, the Corpus Loader constructs two metadata columns for textual files: filename and filepath. The filepath for a document is the path displayed in the file selector window of the Corpus Loader. The filename for a document is the name of the file without the file type extension, e.g. a file 'example.txt' will have a filename metadata of 'example'.

Notes on file type specifics

All files are expected to be UTF-8 encoded.

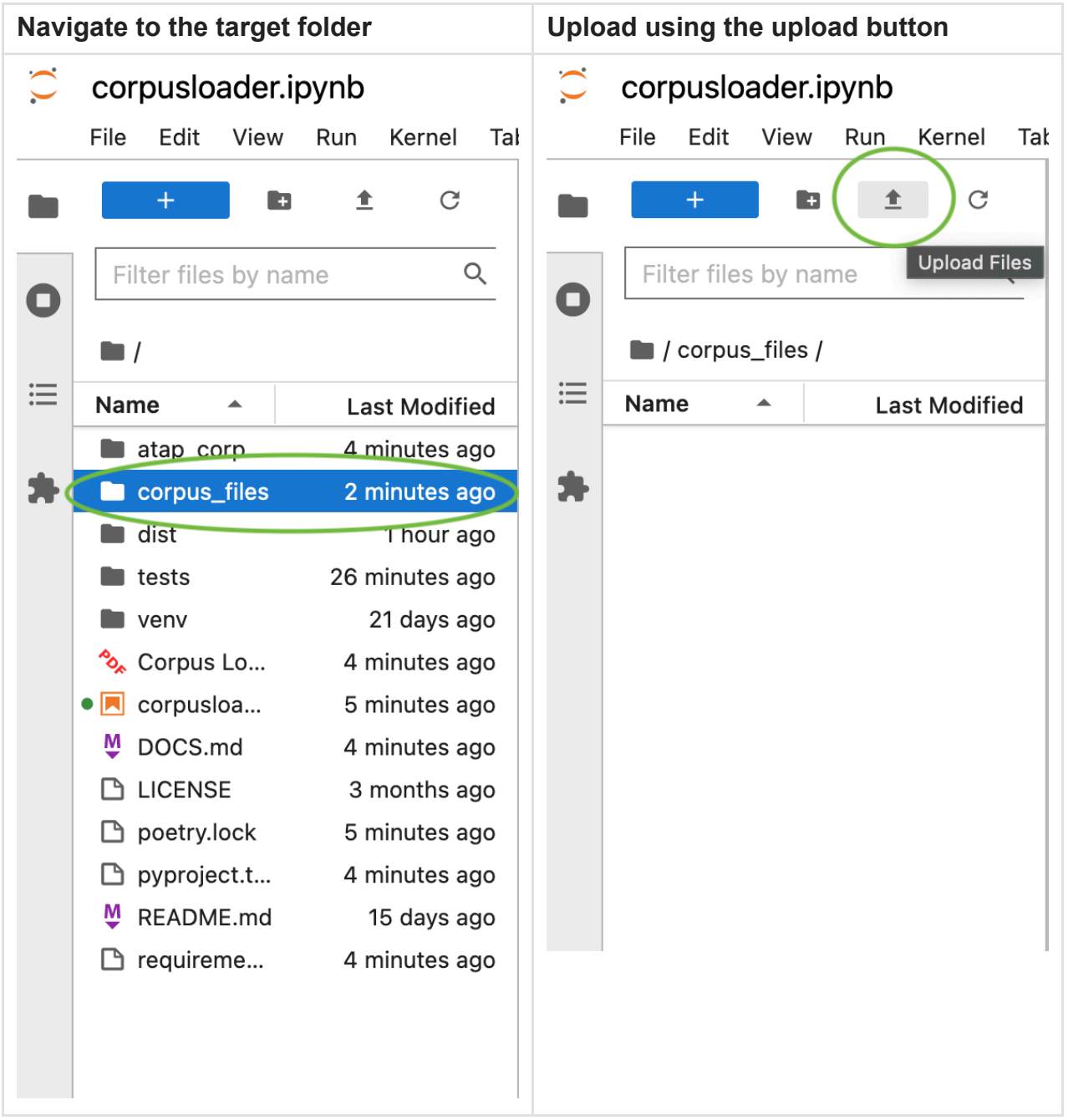
TXT
This file type has almost no pre-processing applied when being loaded into the corpus, i.e. the text is simply read into the corpus as-is.
This means that if you would like to load another file type verbatim (e.g. you wish to load a CSV as a text document rather than a tabular document), you should rename that file's extension to TXT.

DOCX, ODT
The Corpus Loader will not include tables, images, text boxes, and their contents in the loaded corpus. Only top-level text will be included in the corpus.

XML
XML files are loaded in the same way as TXT files, which is with no pre-processing.
In order to interpret the XML tags in a specific way, a custom pre-processing function should be passed to the CorpusLoader.

Upload

The Corpus Loader will display files from a specific folder, defined in the cell that starts the Corpus Loader in the notebook.
Use the notebook file browser to upload corpus and metadata files to the specified folder. This can be done by dragging the files from your computer's file browser into the notebook file browser. You can also use the 'Upload files' button near the top left of the notebook file browser.



Load a corpus

Use the file selector to select one or more files, and then load these files as corpus or metadata.

- Loading a file as corpus indicates that it contains 'document' data (the primary data to be analysed). 'Corpus' files can also contain metadata, but may not be exclusively metadata
- Loading a file as metadata indicates that it does not contain 'document' data, and only contains metadata

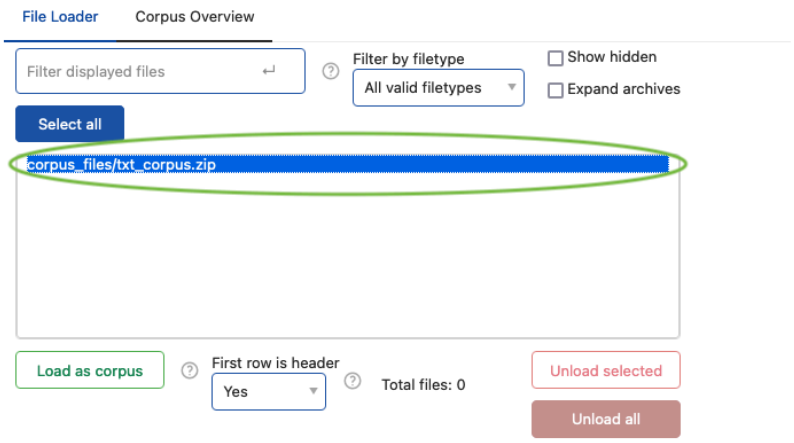
The file selector contains several features to filter and display the files to be loaded:

- The 'Select all' button: this will select all files displayed in the selector (including those files that are off-screen and must be scrolled to view). Using 'Select all' can be slow if

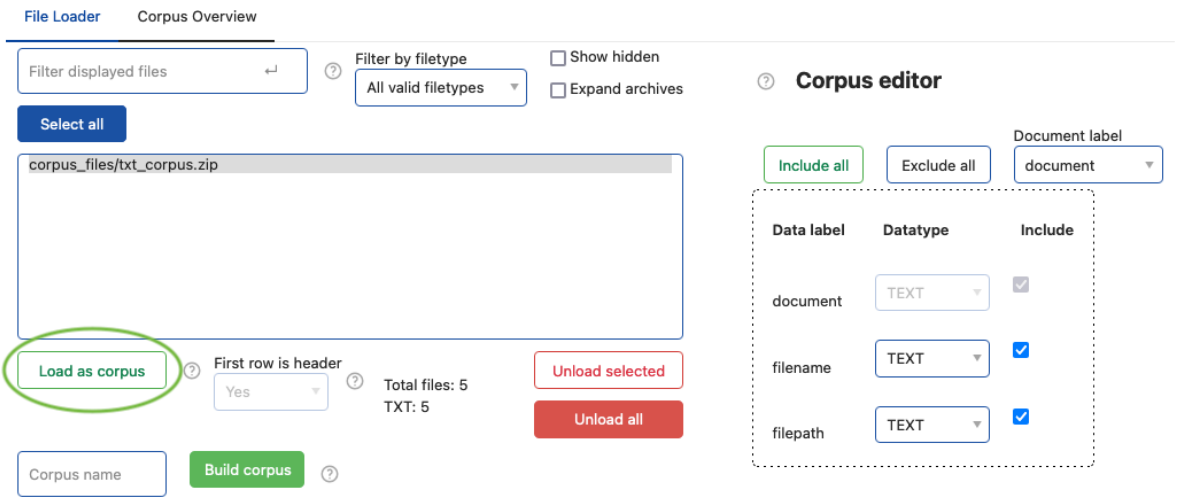
there are many files (~ >1000) in the selector.

- The filter text field: typing text in this field will filter files whose file path matches the text, e.g. typing 'example' in the field will display the files 'example.txt' and 'example_file.txt' but not 'another_file.txt'.
  This can be used to filter for subfolders. The * character is a 'wildcard', and represents any text, e.g. typing '197*record' in the field will display the files '1971-record.txt' and '1975-record.txt' but not '1980-record.txt' or '1907-record.txt'
  The filter will be applied when you press 'enter' or if you deselect the field. To clear the filter, delete all text within the field.
- The file type dropdown filter: this allows filtering of a specific file type. It defaults to 'All valid filetypes'. This dropdown will filter internal files in a ZIP file.
  The file selector will not display files that are not one of the valid file types.
- The 'Show hidden' checkbox: when checked, this will display files whose filename begins with '.', e.g. '.example.txt'. These files are unlikely to appear and so can be ignored by most users
- The 'Expand archives' checkbox: when checked, all ZIP files displayed in the file selector will have every internal file displayed in the selector, allowing files to be selected individually. When unchecked, selecting and loading a ZIP file will load all internal files.
- Keyboard keys:
  - Click and drag will select/deselect multiple files
  - Cmd+click (Mac) / Ctrl+click (other) will select/deselect multiple files that are not adjacent
  - Up/down arrow keys can be used to navigate the file selector. Holding shift while navigating will select all files visited

Select the files



Load the files



First row is header (optional)

When loading tabular data it is assumed that the data contains a header row (a row at the start that names each column). If your data does not contain a header row, you can change this setting to 'No' in the dropdown labeled 'First row is header'. If you are unsure whether the data contains a header, you can change the setting to 'Infer', which will apply a simple heuristic to guess whether the first row is a header.
Note: only one option can be selected when loading data.

Build a corpus

If both corpus and separate metadata files are loaded, then the metadata and corpus files must be linked. To link the two, use the linking dropdown selectors between the corpus editor and the metadata editor choose metadata labels to join on. Once the metadata label is chosen, a 'link' symbol will be seen next to the selected label.

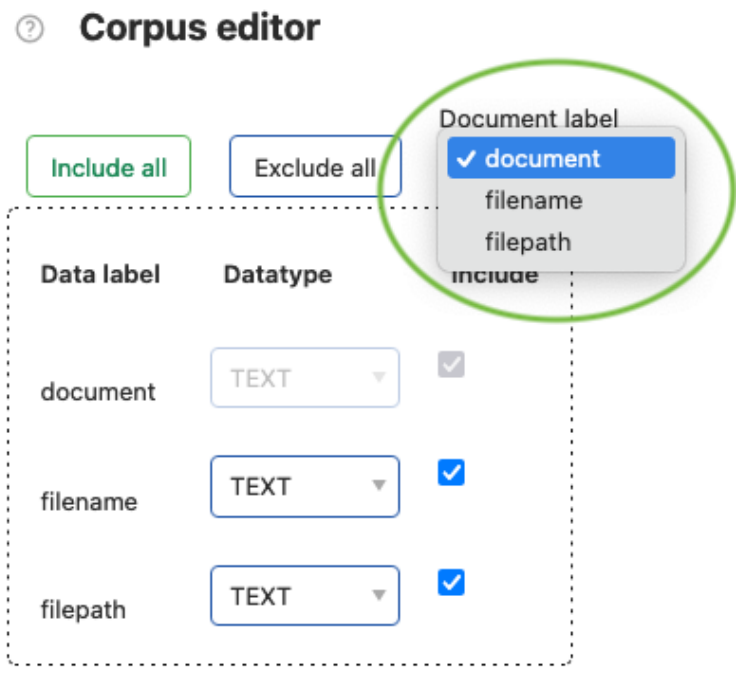
Once metadata linking has been done, or if not required, prepare the metadata for loading as follows:

1. A document label should be selected in the corpus editor. This is the data to be used as the primary data to be analysed. For textual file types, this will default to the content of the file. For tabular file types, this will default to the left-most column.
2. All metadata is included by default, but can be excluded from the final corpus by unchecking the checkbox under the 'Include' heading. Document data and linking metadata cannot be excluded from the corpus.
3. Data types for metadata are inferred but can be modified using the dropdown selectors under the 'Datatype' heading. The TEXT data type is unconstrained and allows the most freedom for the metadata (inconsistencies, missing values, etc.). The other data types have the following constraints:
    - INTEGER - each entry must be a whole number
    - DECIMAL - each entry must be a number
    - BOOLEAN - each entry must be either 'True' or 'False'

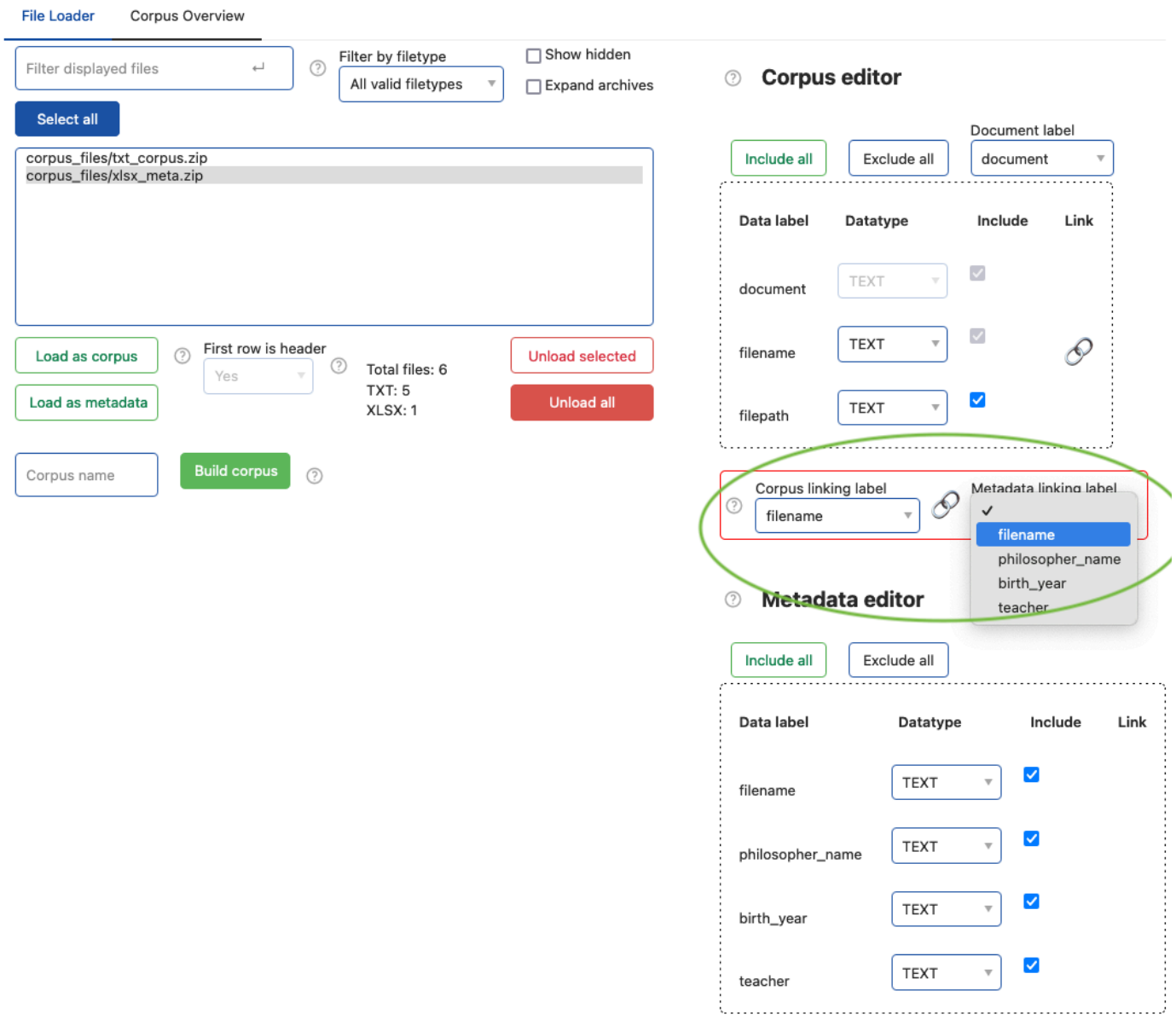
- DATETIME - each entry must be of the format 'yyyy-mm-dd hh:mm:ss'
- CATEGORY - similar to text but some analyses will treat equal values as part of the same category

Once the metadata has been prepared, the corpus is ready to be built. In the bottom left of the tool, type in a name for the corpus and then click the 'Build corpus' button. A progress bar will appear while the corpus is being built, and a green notification in the bottom right of the window will appear when the corpus is complete.
If there is an error during a corpus build, a red notification in the bottom right of the window will appear with information on why the build failed.

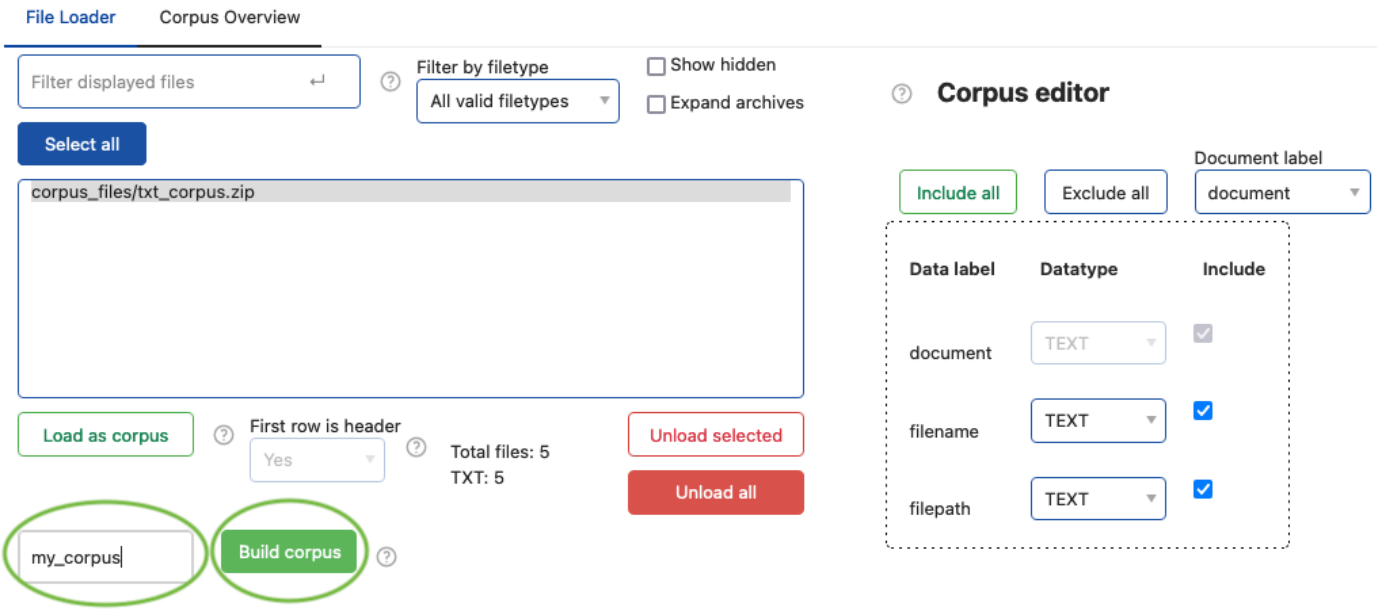
Select the document label



Link metadata labels (optional)



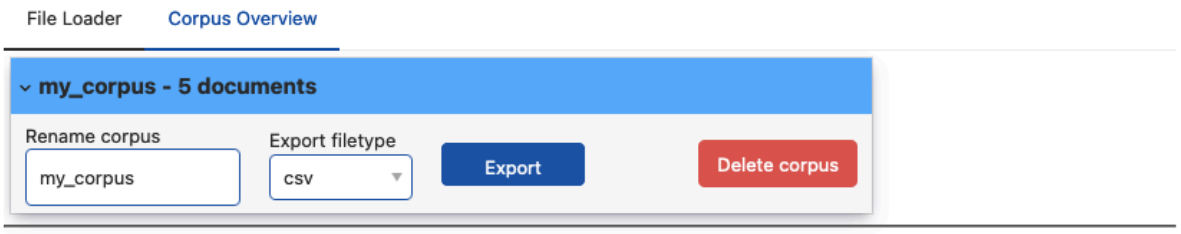
Name and build the corpus



View the corpus

The 'Corpus Overview' tab lists all corpuses that have been built in the current session. It contains information about each corpus (name, number of documents, data labels, and data types).
To view information about a corpus, click on the corpus name to expand the panel. From here, the corpus can be renamed by typing in the 'Rename corpus' text field and pressing enter.
The corpus can be exported by selecting an export filetype using the 'Export filetype' dropdown, and then clicking the 'Export' button. This will trigger a download of the corpus as a single file.
The corpus can be deleted by pressing the 'Delete' button.

View the corpus summary

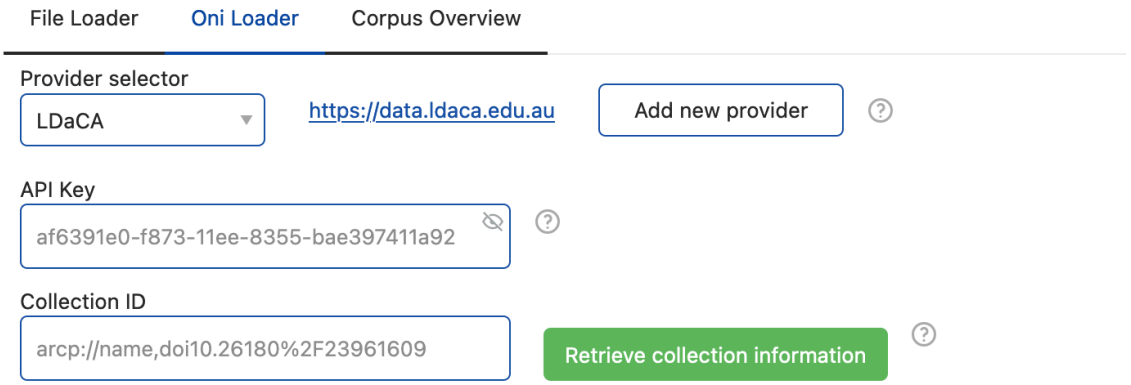


my_corpus

Data label	document_	filename	filepath
Datatype	TEXT	TEXT	TEXT
First document	A pupil of Plato, Aristotle became one of history'	aristotle	corpus_files/txt_corpus.zip/txt_corpus/aristotle.t

Oni Loader

The Oni loader enables loading a corpus directly from a platform running the data archiving service Oni. To use the Oni loader, click the tab at the top of the Corpus Loader labelled "Oni Loader".



Select a provider

Select an Oni provider using the dropdown to set the current Oni provider.
The link next to the selector displays the current provider address.

Add a provider

You can add an Oni provider if your data is stored somewhere running an Oni implementation that isn't listed in the dropdown.

To add a provider, click the "Add new provider" button and fill out the fields. The provider name can be anything but not cannot be left empty. The Provider address should be the base URL of the provider, e.g. https://data.ldaca.edu.au

Set the API key

An API key from the selected Oni platform must be provided to access a collection.

To obtain your API key, do the following:
4. Visit the provider portal by clicking the link next to the provider selector
5. If not logged in, login using the button in the top right
6. If logged in, navigate to the 'User Information' page
7. Under 'User Details' > 'API Key', click "Generate"
8. Copy the API Key shown. Navigate back to this tool and paste in the field provided

Retrieve a collection

To access a collection at the Oni provider, enter the collection ID in the provided field and click the "Retrieve collection information" button.

To find the collection ID visit the provider portal, view the page of the collection you want to access, and the collection ID will be the text labelled "@id". Ensure you copy the *text* of the ID and not the link of the ID.

Building the corpus

Once the collection is retrieved, the corpus can be built by following the steps outlined in the File loader instructions above.