

Create a Document Corpus

This section contains a short guide on how to create a document corpus in PoolParty.

Creating a document corpus is crucial if you want to enrich your thesaurus with terms from relevant documents.

You can choose to create a document corpus based on documents (PDF, DOC, Powerpoint, TXT, etc.) that are related to your project's domain or harvest RSS feeds, web sites and DBpedia resources linked to the concepts in your thesaurus. Supported file types are those supported by the Apache Tika library.

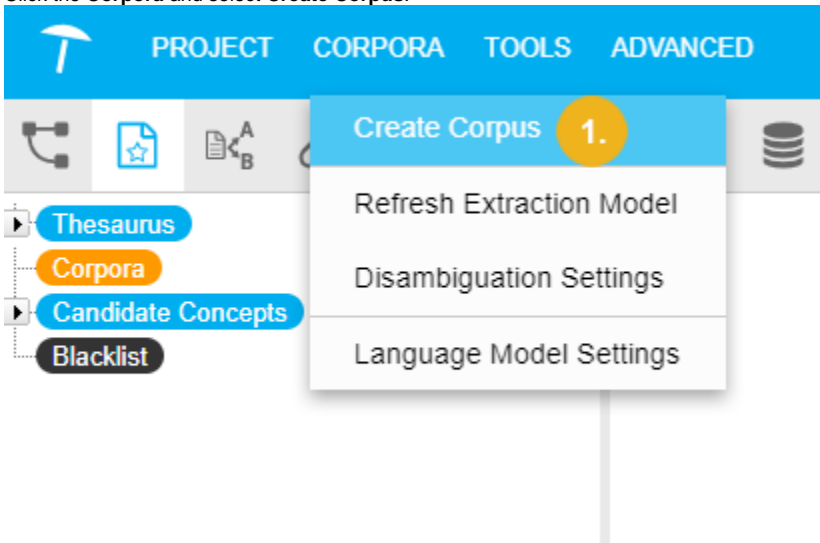
Prerequisites

- An opened PoolParty project with an existing [thesaurus](#).

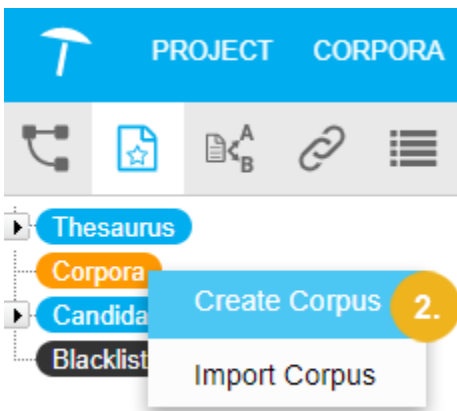
How to Create a Document Corpus in PoolParty

You can create a corpus by doing the following:

1. Click the **Corpora** and select **Create Corpus**.



2. Or, click the **Corpus Management** and right-click the **Corpora**, select **Create Corpus**.



3. Define in **Create Corpus**:

- **Title:** Enter the title of your choice. It will be the name for this corpus throughout the project, appearing in the **Hierarchy Tree** for example.
- **Language:** Choose the language of the corpus' documents from the drop down. Languages available are determined by the projects [Language Settings](#).

- **Repository:** Select the repository where your corpus is going to be saved. The primary **local store** is set to **Embedded GraphDB**. GraphDB is shipped as an add-on module. Other options depend on the PoolParty setup using the [Semantic Middleware Configurator](#).

The screenshot shows a 'Create Corpus' dialog box. The title bar is blue and contains a back arrow, the text 'Create Corpus', a yellow circle with the number '3.', and a close 'X' button. The main content area is white and contains three input fields: 'Title' with the text 'Cocktails', 'Language' with a dropdown menu showing 'en', and 'Repository' with a dropdown menu showing 'Embedded GraphDB'. At the bottom right, there is a yellow circle with the number '4.', a blue 'Create' button, and a blue 'Cancel' button.

4. Click Create Corpus.

Now, you can start [uploading documents to your corpus](#).



PoolParty Advanced Server allows one corpus per project. Multiple corpora are available for PoolParty Enterprise Server and PoolParty Semantic Integrator.