Corpus Management - Overview

This section is dedicated to the corpus management functionality in PoolParty.

The corpus management functionality in PoolParty supports you in extending thesauri with relevant terms derived from documents matching the domain of your thesauri. In addition corpora are used to improve entity extraction providing improved scoring of terms and concepts and offering shadow concept suggestions based on co-occurrences.

You can also start to create a new thesaurus from scratch based on a corpus.

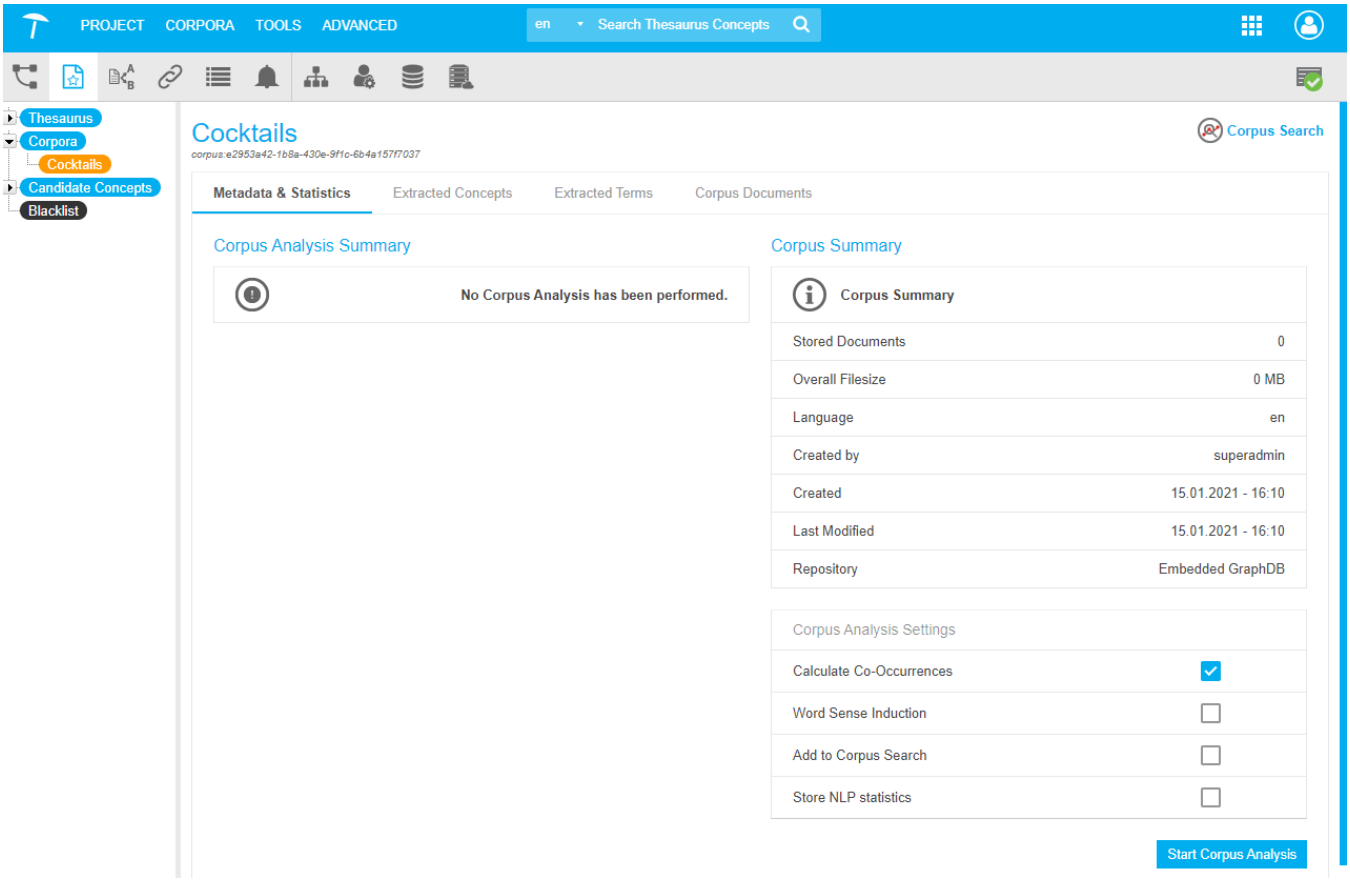
PoolParty's Corpus Management Functionality

In order to enrich your thesaurus with terms, using the corpus management function, you can process 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.

The PoolParty corpus management tightly integrates the PoolParty Extractor into the thesaurus management process. It uses the extractor's ability to analyse text and extract terms and phrases, which then are matched against the concepts in your thesaurus. You can then integrate extracted domain specific terms as new concepts or synonyms of existing concepts into your thesaurus.

The terms you decide to select and use for integration into your thesaurus from the extracted terms are called 'Candidate Concepts' in PoolParty. Find details about their handling and the possible workflow here: Candidate Concepts List

The following image shows an example **Corpus Management** view, where a corpus called 'Cocktails' already has been created:



To learn in detail how to use the **Corpus Management** feature, refer to the following topics:

- Create a Document Corpus — This section contains a short guide on how to create a document corpus in PoolParty.
    - Rename a Corpus — This section contains a short guide on how to rename a corpus you created in your PoolParty project.
    - Delete a Corpus — This section contains a short guide on how to easily delete an existing corpus from your PoolParty project.
- The Corpus Management Tree — This section contains a short guide on how to access the Corpus Management tree of your PoolParty project using the Corpus Management.
- Upload Documents to a Corpus — This section contains a short guide on how to use the document upload dialogue for your corpus.
    - Upload Documents From Your Local Drive — This section contains a short guide on how to upload documents from a local drive to your corpus.
    - Paste a Text to Add It to Your Documents — This section contains a short guide on how to add text to your corpus by copying and pasting it into the Upload Documents dialogue.
    - Grab Documents From HTML Pages — This section contains a short guide on how to use the Crawl Website function for your corpus.

    - Use an RSS Feed as Corpus Document Source — This section contains a short guide on how to upload documents to your corpus by crawling RSS feeds.
    - Use DBpedia as Corpus Document Source — This section contains a short guide on how to crawl DBpedia to add documents to your corpus.
- Analyse Documents in Your Document Corpus — How do you use PoolParty's Corpus Analysis?
    - Corpus Quality — This section contains a short guide on how to use the additional information on the corpus quality information available in PoolParty after the analysis has been performed.
    - Extracted Concepts List — This section contains a short guide on how to use the Extracted Concepts list of the corpus management.
    - Extracted Terms List — This section contains a short guide on how to use the Extracted Terms list of the corpus management.
    - Candidate Concepts List — This section contains a short guide on how to use the list of candidate concepts: they are terms that have been extracted from the corpus as possible new concepts and been added manually to this list.
    - Corpus Management Thesaurus Tree - Options — This section contains a short guide on how to use the Thesaurus Tree PoolParty's Corpus Management.
    - Blacklist Concepts and Terms — This section contains a short guide on how to blacklist concepts and terms from inside of one of the two lists.
- Viewing the Documents of a Corpus — This section contains a short guide on how to use the Corpus Documents tab you can use to view the documents in a corpus.
    - Documents Details Tab — This section contains a short guide on options and functions you have in the documents details dialogue.
- Export a Corpus — This section contains a short guide on how to export a corpus from an existing PoolParty project.
    - Export Corpus Documents — This section contains a short guide on how to export corpus documents.
- Import a Corpus — This section contains a short guide on how to import a corpus into an existing PoolParty project.
- Create a Corpus Search Interface Within PoolParty — This section contains a short guide on how to use the Corpus Search feature to create a simple interface.
- Co-occurrences in PoolParty - Usage and Function — This section contains a short overview of what the co-occurrences calculation in PoolParty does and how you can use its results to advantage.
    - Shadow Concepts in PoolParty — This section contains a short guide on how to use the Shadow Concepts functionality in PoolParty.
- Word Sense Induction - Usage and Function — This section contains a short guide about what Word Sense Induction in PoolParty means and how it helps disambiguating terms in your corpus.
    - How to Add Terms to Your Thesaurus From The Word Senses List — This section contains a short guide on how to use the Word Sense Induction results for your PoolParty thesaurus.

Multiple corpora are available for PoolParty Enterprise Server and PoolParty Semantic Integrator.

PoolParty Advanced Server allow one corpus per project.

You can manage your corpus or corpora programmatically as well, or automated remotely by using the PoolParty Corpus API services, such as: Web Service Method: Create a New Corpus, Web Service Method: Upload a Document to a Corpus, Method: analyse corpus, Web Service Method: Request Concept Matches of a Corpus, etc.

In addition you can significantly improve extraction results of free terms by using a corpus. Details find here: Free Terms Extraction Based on a Text Corpus

PoolParty Academy Tutorial

(Duration: 15m07s)