Corpora

Page Contents

Overview of Corpora

A Corpus is a collection of read-only textual items, such as documents, excerpts, websites, etc.—along with their associated metadata. The original items are always imported from external sources, such as content management systems or web sites and are **never originally created nor edited** within EDG. Thus, a version of the Tagger interface allows viewing of Corpus content, without editing. The textual content of Corpus items provides the foundation for manual or automated tagging and annotation with Content Tag Sets (serving as the content graph).

Corpora Home

Selecting the **Corpora** link in the left-navigation pane of TopBraid EDG lists all of the Corpus collections currently available to the user and, it allows authorized users to create new ones.

Prerequisites: Licensing and Enablement

The availability of any collection type (including Corpora and *customer-defined* types) is determined by what is (a) licensed and (b) configured under *Server Administration*. To install a license or to view the currently licensed features, see Setup > Product Registration. To configure which licensed collection types are currently enabled or disabled, see EDG Configuration Parameters > Configure Asset Collection Types. For general licensing information, see the TopQuadrant website, which describes the TopBraid products and the *data governance packages* that determine the available collection types.

Listing of Corpora

This home view shows a table with all Corpora that you can access in some way. For each collection some brief metadata is available in columns of the table. Columns are sortable and you can filter content of the table by typing search strings in the **Refine** field at the upper right of the table. To access an asset collection, click on its link.

To create an asset collection click on **Create New Corpus** button.

You can also select an asset collection in the table and start a workflow for it.

This page provides a focused view on Corpora. To see a view of all asset collections, irrespective of their type, that you have a governance role for, click on your **User Name** in the upper right corner of the page. To see all asset collections you have access to organized by their subject areas, click on the **Governance Areas** link in the left hand-side vertical Navigation Bar.

If a Corpus is either missing or it is lacking expected features in your views, you or your security role(s) may lack proper permissions for the Corpu s. A manager of the Corpus can give you the needed permissions via its utilities' **Users** settings. For background information, see Asset Collection Permissions: Viewer, Editor, and Manager.

Another possible cause of a missing feature is that it requires administrative setup to become active. See EDG Administration for relevant within-application settings and/or see other EDG Administrator Guide documents for relevant external installation and integration setup.

Create New Corpus

The **Corpora > Create New Corpus** link opens a form with fields used to define the new Corpus. Note that you can also create a Corpus by using a Create link in the Governance Areas page.

Nobody will have a link for creating any asset collection until an administrator configures EDG's persistence technology as documented in *Server Administration: Teamwork Platform Parameters: Application data storage* . Additionally, each user will not have a create link unless the user or their role has a *Create* permission for the EDG *Repositories* project as documented in *EDG Rights Management* .

Whenever a new Corpus is created, EDG requires the user to select its *data source*, as Corpora can be configured to connect to an external source. EDG will then harvest content from that external source and store it in the project graph. Harvesting can be repeated later on demand, and changes to the external source's documents will be picked up. (Harvesting needs to be triggered manually in the Corpora management UI. So a Corpus does not synchronize automatically with its external source, but only when requested.)

Four types of connectors to such sources are currently offered, their respective creation wizard pages showing different forms depending on parameters that must me indicated for the connector to operate. These parameters can be adjusted later on by accessing **Manage > Corpus-CONNECTOR-TYPE Configuration** .

- **No connector** if content documents are available as RDF already, these can be imported into the Corpus with the usual RDF import function. Similarly, raw documents can be imported singly from local files as described in Import > Import Single Document. No external source will be configured with this connector type.
- **sitemap.xml** If a website supports the *sitemaps* protocol, a configured sitemap.xml connector will harvest its content accordingly.
- **URL list** This connector will simply fetch content from all of the URLs listed in its configuration.
- **CMIS** If a website is an interface to a Content Management System and offers a Content Management Interoperability Services (CMIS) service endpoint as defined by the standard, a configured CMIS connector will harvest its content accordingly.

Once you have finished configuring your new Corpus, it will appear as a link on the Corpora tab.