- **Taxonomy Division Open House**
  Join the Taxonomy Division for our 10th anniversary celebration at our annual open house.

Submitted by Janice Keeler, Taxonomy Division Treasurer and 2019 Program Planning Co-Chair. Edee Edwards served as co-chair for planning, with additional session planning by Margaret Nunez.

Bridging Vocabulary Silos with Auto-Generated Crosswalks – Nuno Lopes

A common issue when working with multiple standard vocabularies is determining which terms are equivalent between the different vocabularies, a task commonly referred to as entity matching or, more generally, as record linkage. The task of linking records often arises from the need to share information between different applications and/or different user communities. It is a time-consuming activity that quite often requires specific domain expertise. Thus, there is a high desire for automation of the linking process, even if partial.

In this article, I describe the solution adopted by TopQuadrant that tackles this problem following a client's specific requirements. A major Pharma company receives datasets with results of drug trials from different partners, where each dataset describes medical conditions, diagnoses, and outcomes using terminologies local to the party that created the dataset. The company needs to 1) align, normalize, and cleanse this information prior to importing it into their internal systems; and 2) enhance its internal ontologies with terms that are included in the supplied datasets.

The solution is based on TopBraid Enterprise Data Governance (EDG), TopQuadrant's data governance solution based on W3C standard RDF graphs. EDG is a flexible solution composed of

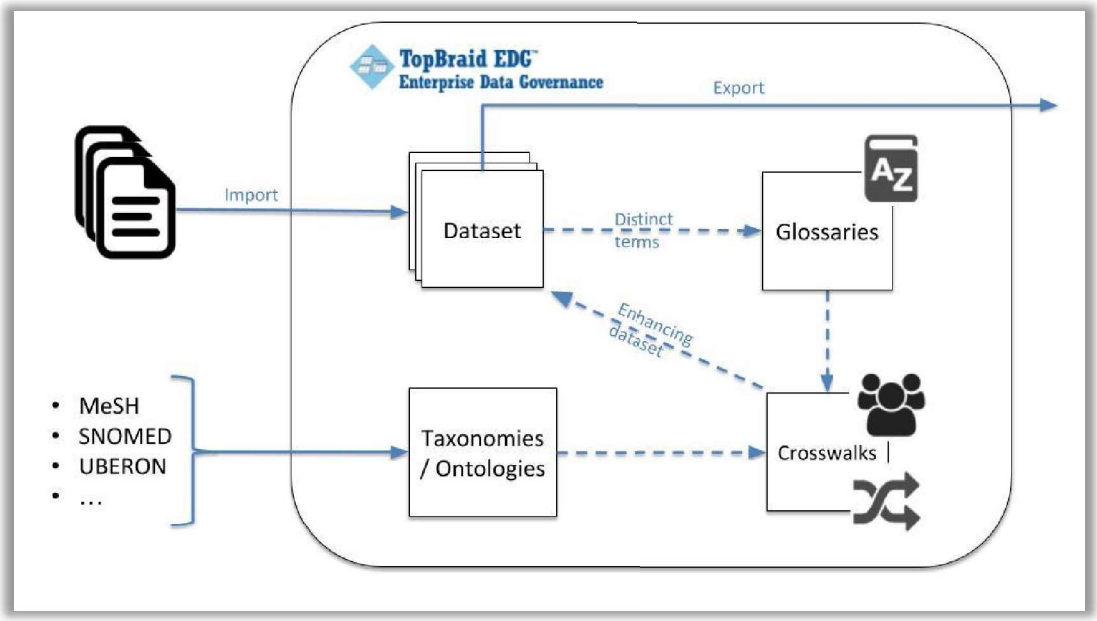
different pre-defined asset types and allows for new asset types to be defined on a per-installation basis. For this client, we relied on a combination of pre-defined asset types:

- *Ontologies* and taxonomies are one of the core EDG asset types that allow users to create and manage their taxonomies and ontologies.
- *Glossaries* are a pre-defined asset type that allow users to store a list of terms of specific types (Glossary, Business, Industry, or Technical terms) along with their definitions.
- *Crosswalks* allow users to create, edit, and automatically discover connections between terms in existing EDG assets. Crosswalks are used to find suitable terms from standard vocabularies such as MESH, SNOMED, UBERON, Entrez Gene, NCit, and the company's own controlled vocabularies of relevant terms.

And other asset types custom-built for this project:

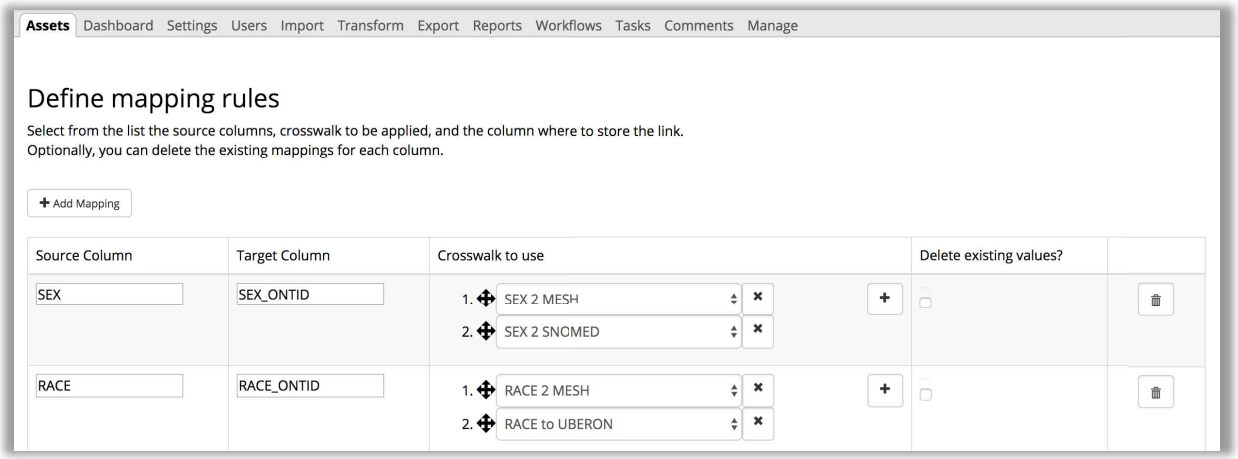
- *Mappings* allow the company users to create glossaries and define crosswalks between them and other vocabularies.
- *Datasets* are used to load the input data and act as input for the glossaries.

The following diagram represents a flow of data between the different EDG asset types:



Importing Datasets

Input study data is imported from CSV files and other structured formats, either automatically from other internal systems or users can directly upload data into EDG. Upon import, TopBraid EDG automatically analyses the input data and determines which terms should be added to the glossaries. Users specify which columns in the input data should be analyzed via the Mappings custom asset:



These mapping specifications include:

- Source Column: The column containing the values users want to map to other vocabularies, in the example 'SEX' and 'RACE' are columns expected to exist in the dataset;
- Target Column: The column that will be populated with the matches to the external vocabularies, in the example 'SEX_ONTID' and 'RACE_ONTID'; and
- Crosswalk to Use: The crosswalk to use to determine which Glossaries shall be populated and to which Ontologies they shall be linked.

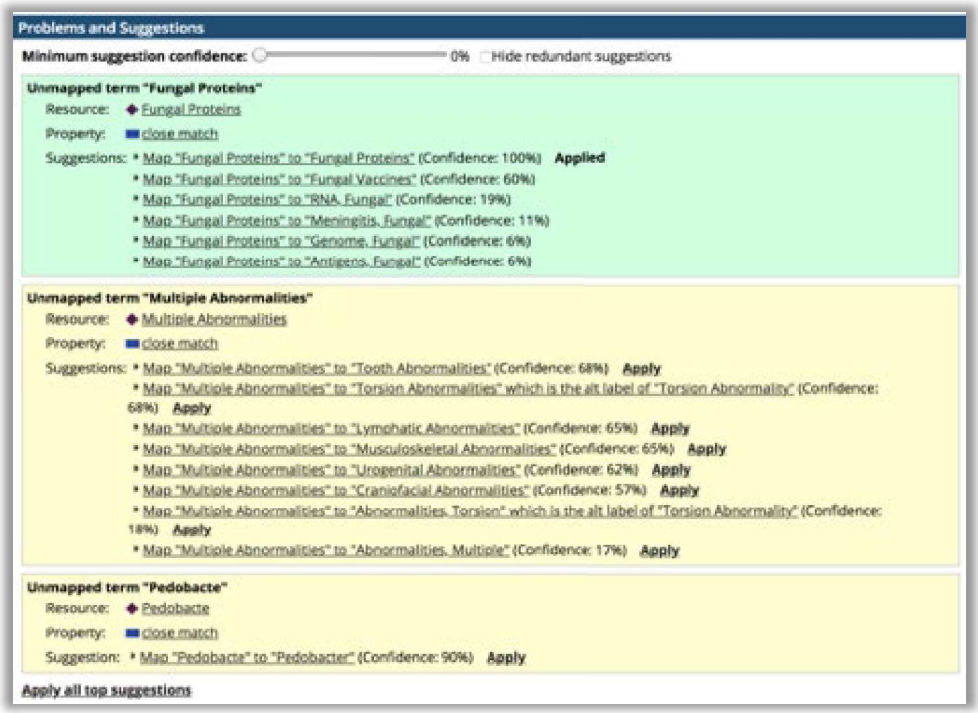
For each mapping, users can select multiple crosswalks, which will be considered as an ordered list in the process to enhance the dataset: EDG will use the match found in the source value in the highest ranked Crosswalk. The mapping specification also defines if existing values should be deleted when reading links from the crosswalks in order to populate the dataset.

Using the information about each mapping, the import process can be performed in a fully automated fashion: for each mapping whose "SOURCE" column is present in the input data, the values of those columns are loaded into the respective glossaries (defined as source of the crosswalks). If the import happens in an automatic way, the users responsible for the curation are presented with a list of datasets that they need to curate.

Auto-Generating Candidate Matches

Each crosswalk is designed to match terms in the glossaries to different external vocabularies (one crosswalk per vocabulary) and also contains information regarding the classes in the external vocabulary to which it will match. This information is configured when creating the crosswalk and allows users to restrict the term matching to subsets of the vocabularies: only instances of the specified class will be candidates for matching.

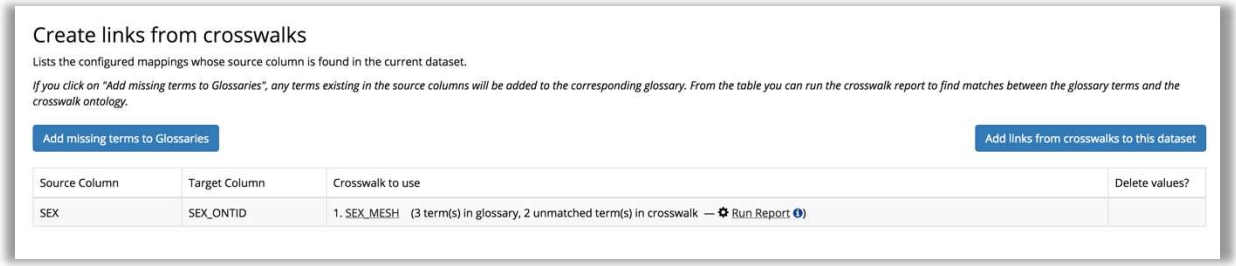
Crosswalks are then capable of automatically providing suggestions for mappings based on the labels in the glossary terms and the instances of the selected class in the target vocabulary. The example below shows the suggestions of matches that are presented to the user where glossary terms can be listed along with possible matches from the target vocabulary and the confidence EDG assigns to that specific match.



Depending on specific crosswalk configuration, suggestions that have 100% confidence are automatically applied. For suggestions with confidence below 100%, users can then select the most adequate match from the list of possibilities. At a later stage users can also override the system-suggested matches and add matches of their own. All decisions made by the users are annotated and the history of changes is kept.

Enhancing the Original Datasets and Internal Vocabularies

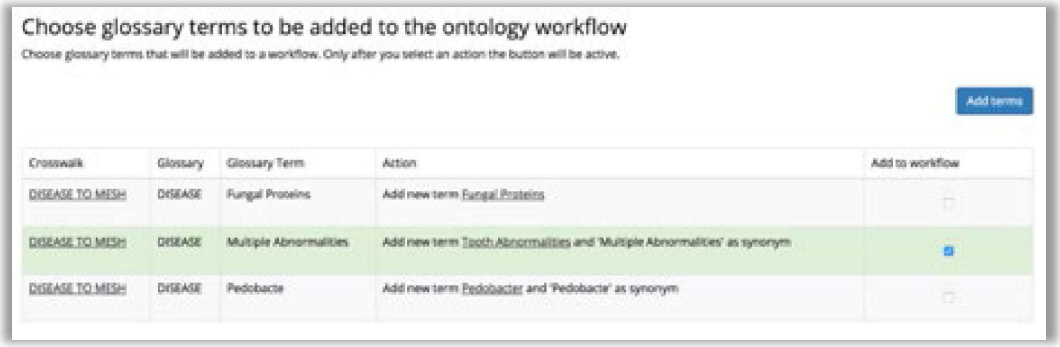
Once all the terms in the glossaries are matched to external vocabularies, users can request TopBraid EDG to enrich the original datasets by adding or replacing data as deemed necessary. Users are presented with an overview of the curation status of a dataset, where they can also access the crosswalk's auto-generation feature if any unmatched terms still exist:



In the previous example, users are presented with the information that the Glossary contains three terms, and the Crosswalk does not include matches for two of them. Clicking on the "Run Report" link allows users to complete the matches in the Crosswalk, and they will be shown the suggestions page detailed in the previous section.

In case users need to curate more columns from a dataset that was already imported, they can define the new mappings and then use the option "Add missing terms to Glossaries," which will ensure all terms are present in the Glossaries and populate any missing data.

Enhancing the internal vocabularies is the final step in the curation process. Any term in the dataset that is not included in the company's internal vocabulary can be selected for addition following a pre-defined set of rules. The result will be that new terms or new synonyms (alternative labels) to existing terms will be included in the company's internal vocabulary:



The rules used to determine the actions to be applied are as follows:

- If the label of the glossary term is an exact match to an external vocabulary term and is not present in the internal vocabulary, show an option to add a new term.
- If the glossary term is mapped to an external vocabulary term but the label is not a 100% match, show an option to add a new term using the external vocabulary term as preferred label and the study term as alternative label / synonym.
- If the glossary term is matched to a term in the internal vocabulary but the label is not an exact match, show an option to add the glossary term label as synonym of the internal vocabulary term.

All the changes to the internal vocabulary are performed using a workflow. In EDG, a workflow is a sandboxed copy of the data in any asset that allows users to perform changes and manipulate data without interfering with the main (production) data. Users also rely on workflows to process changes to the vocabulary in batches and, when they are finished with these changes, the workflow data can be merged into production causing an increment in the version of the internal vocabulary.

Throughout the curation process users can export any of the data they use (glossaries, crosswalks, etc.). The input datasets specifically allow users to retrieve its curated data back in CSV or Excel format or export the data directly into other of the company's internal applications. The extended versions of the company's internal vocabulary can also be reused by other downstream applications that can leverage the new information provided.

Nuno Lopes is a Senior Semantic Solutions Architect at TopQuadrant. His work focuses on applying semantic standards, deploying and customizing TopQuadrant products, and working in different client projects mostly in the biomedical and life sciences domains. Nuno has more than ten years of experience in Research and Development and, prior to joining TopQuadrant, he worked as a Research Engineer at the Smarter Cities Technology Centre, IBM Research, Dublin and as a Postdoctoral Researcher with the Digital Enterprise Research Institute (DERI).

New Members

Welcome to the members who joined the Taxonomy Division in the fourth quarter of 2018!

- Danielle Boulay
- Helen Challinor
- Helmi Fournier
- Yonah Levenson
- Sarah Lin
- Lorre Smith
- Ann Swearingen
- Judith Theodori