Using Graph Technologies and Artificial Intelligence to Power Terminology Management

Brad Farrell, Leonya Ivanov, Emily Lorenzen,
Yaw Nti-Addae, Felicia Sabatino, Oleg Stroganov

Rancho BioSciences

Abstract

For data to be findable, integrable and reusable, it first needs to be normalized (so that data from different sources can be aligned) and, most importantly, it needs to be cleaned up, so it is free from original human and machine errors.

For both tasks, it is a standard practice to align data to well-established standard ontologies and controlled vocabularies and to curate it, both manually and digitally. While there is no automated solution that can guarantee clean and well-aligned data, an efficient semi-automated solution can do the preliminary work, thus leaving curators with fewer, more complex cases.

Furthermore, resulting data dictionaries often need to be classified and tested for heterogeneity, so that they can be used in more structured, domain-specific, targeted and harmonized fashion.

The annotation and mapping services can be used for streamlining data that comes from public resources and is often presented in variety of formats and flavors. The goal if this project is to provide one-stop-shop for data collection, harmonization, alignment and mapping.

Methods

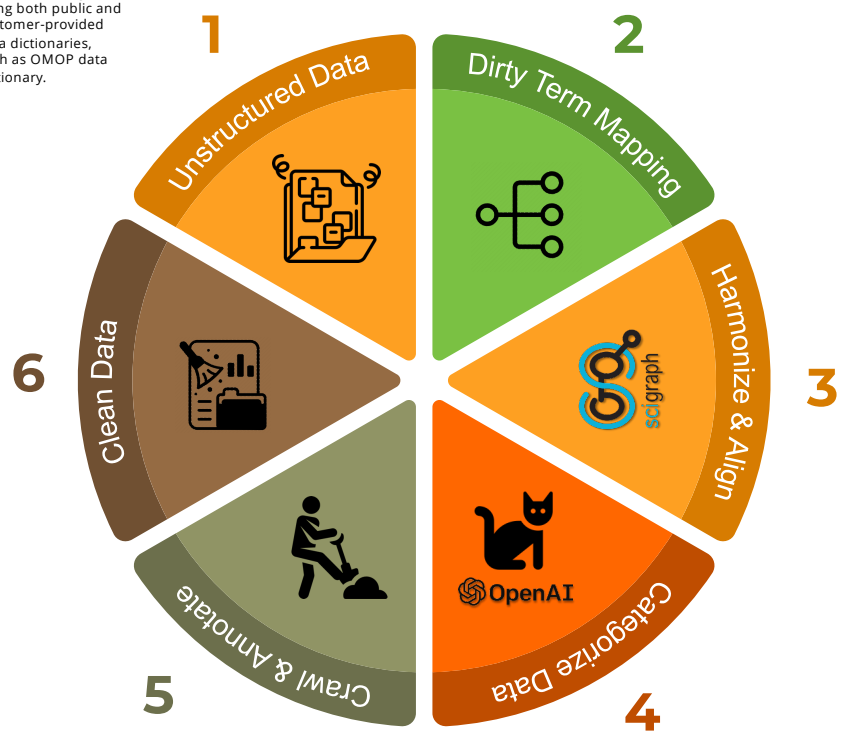
For Data Crawler, we use related data source APIs, which are then fed to our terminology management ecosystem for deeper analysis and annotation.

For Term Mapping, we use PostgreSQL database that is indexed using trigram extension. Trigram method calculates all combinations of 3 characters and measures matches between them. This method allows us to pre-index mappings so that term mapping can be performed live.
For hierarchical tasks and queries, we use SciGraph (and open-source ontology store based on Neo4j). This tool allows us to manage standard and custom ontologies, and perform major semantic tasks such as term level alignment, common ancestor searches and text annotation.

For Categorization Tool, we use OpenAI embeddings endpoint and DBSCAN (Density-based spatial clustering of applications with noise) clustering algorithm. Unlike other clustering methods, such as k-means, DBSCAN does not require predetermined number of clusters, and uses epsilon distance instead. That feature allows us to let users define epsilon via UI to fine-tune clustering results until they get groups they need. Once clustering is done, we use OpenAI completion endpoint to obtain names of suggested clusters.

Data

- Data Crawler is configured to crawl the following resources:
  - PubMed
  - ClinicalTrials.gov
  - GEO
  - SRA
  - EGA
  - ArrayExpress
  - DbGaP
  - We will be adding SCP, FigShare and Zenodo in the near future.
- Terminology Mapping uses public ontologies and vocabularies, such as OMOP, SNOMED, DOID, NCIT, MedDRA, etc.
- We tested Categorization tool using both public and customer-provided data dictionaries, such as OMOP data dictionary.



Categorization Tool

Categorization Tool is a light-weight product for QC'ing and heterogeneity testing of Data Models and Data Dictionaries

Given a list of terms, the tool suggests possible term groups ("domains" or "clusters") and "outliers" (terms that don't fit into the list).

It is used for suggesting best data organization structures and can be fine-tuned to fit your exact needs by adjusting distances and sensitivity. It has an option for specifying categories or "blind" clustering.

Terminology Mapping

IRancho Terminology Services are built on our Ontology store containing standard and custom ontologies . It includes Ontology management tools , a tool for suggesting best ontologies for your project , a free-text annotation , common ancestor, and term level alignment endpoints.

Ontologies and controlled vocabularies are mirrored in PostgreSQL Database that is pre-indexed for "dirty term mapping" using trigram method. The "fuzzy" mapping endpoint build on top of Postgres is designed to find best matches for the term with resulting similarity scores

The "fuzzy" mapping endpoint results are combined with Open AI embeddings and synonym suggestions to provide better results .

Public Data Crawling

The Rancho Data Crawler is a web application that allows input of a complex search string input to crawl study-level metadata in PubMed, ClinicalTrials.gov, GEO, SRA, EGA, and ArrayExpress. It outputs results to an XLSX (Excel) file within minutes. This application is meant to accelerate the data crawling process by putting our proprietary Rancho data crawler scripts at the fingertips of any project team that needs them, no scripting needed.

Our Deep Crawler goes beyond dataset-level data as it also crawls experiments, samples, etc. Furthermore, in combination with Rancho Annotation Service, it annotates free-text fields from both metadata and sample levels against standard ontologies (e.g. Uberon, DOID, BTO) to extract information about treatment, condition and tissue. It also has an AI-assisted scoring mechanism that allows users to rank results by significance. For sources like PUBMED, the Deep Crawler also calculates journal ranks, citation indices, provides cross-links to other repositories (GEO, SRA, DbGap), links to supplemental materials and extracts PMC availability information. Finally, Deep Crawler allows for regular incremental crawling so that the user can get regular updates when new data becomes available.

Results

Data Crawler
- allows users to efficiently collect and manage public datasets, on both dataset and sample level, and to stay informed about new data in their area of interest

Terminology Mapping
- helps users normalize and harmonize their data dictionaries. Depending on the original data quality, it provides between 50 and 75% successful mapping, greatly improving time and effort needed to harmonize large data streams.

Categorization Tool
- allows users to effectively QC and categorize their data models and data dictionaries.

Altogether, Rancho Terminology Solution is an all-in-one product that allows you to clean up, harmonize, align, categorize, QC, annotate and manage your ontologies, data dictionaries and controlled vocabularies.