Data Curation: The essential step for integrated data-driven research

Heterogeneity within and between data sets makes scientific analysis difficult and, therefore, cleaning and harmonizing the data, as well as standardizing data encoding are of great importance. However, the process of clinical and phenotypic data curation is a laborious and manual process that demands deep substance expertise and significant personnel resources.

The data curation process

Heterogeneity in data can derive from many sources: data is collected over long periods of time (changing practices), by multiple persons in the same or different institutions (lack of agreed standards) or it may originally have been collected for another purpose (e.g. EHR data). In addition, errors, such as typos, occur frequently. Irrespective of source, curation of the data is essential for any scientific analysis. This process needs to be efficient, accurate, reproducible and auditable. Additionally, in larger organizations and collaborations, consistency between curators is necessary.

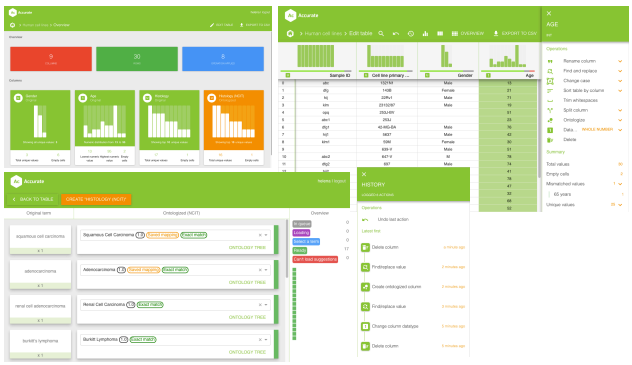

Figure 1. MediSapiens Accurate™ provides efficient and intuitive functions for e.g. data exploration (top left), data cleaning & enrichment (top right), ontology matching (bottom left) and auditing (bottom right).

Data curation can be defined as the process of identifying and correcting incomplete and incorrect data, as well as harmonising and integrating data from different sources. This process can be divided into the following steps:

- **Exploration:** In this step the focus is to understand not only the data that has been gathered but also the objectives of the analyses, which will result in a more accurate characterization of data fields, types and values.
- **Cleaning:** Raw clinical data from various sources often contains inaccurate or corrupt records. The objective of this step is to detect and correct these records, harmonizing the data across and within datasets.
- **Enrichment:** Certain data can be further curated by creating new variables from the original values. Depending on the final use of the data, for example, numerical data can be categorized for easy use or multiple primary data variables collapsed into a single descriptive term.
- **Standardization:** This can be done using ad hoc / internal standards, but that is inefficient, frequently not reusable and makes integration with other sources difficult. Ontologies provide natural common vocabularies for this purpose, with several added benefits.

Mapping data to ontologies

Multiple ontologies in the biomedical domain, such as Human Phenotype Ontology, SNOMED-CT and Disease Ontology, provide existing and continuously maintained standards for data representation. In addition, the hierarchical structure of an ontology can be used to analyze data at different levels of abstraction, such as recently published for UK Biobank data (Cortes A et al.).

However, matching original data to ontologies is a time consuming manual process, requiring deep substance expertise from curators. While multiple methods have been developed to automate mapping data to ontologies (Pang et al., Whetzel et al), significant curator time is required to validate their output, and the user experience provided is not optimal. MARDO, the ontology mapping algorithm developed by MediSapiens (figure 2), solves these problems by providing highly accurate automated ontology mapping for clinical and phenotypic data, along with multi-language support, remembering previous curator approved mappings and a user friendly mapping workflow in Accurate™.

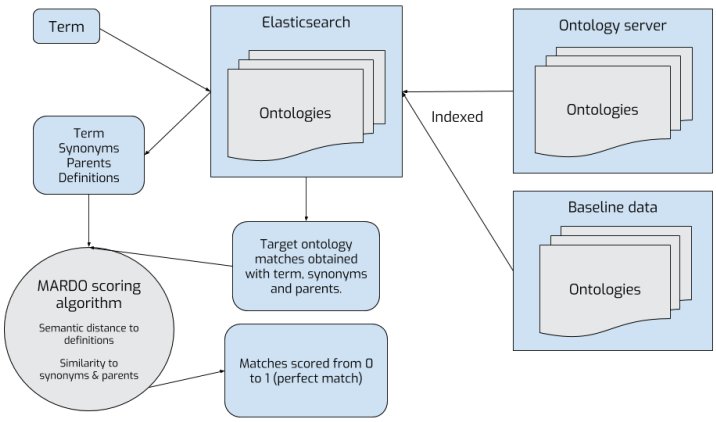

Figure 2. An overview of the MediSapiens MARDO ontologizing algorithm.

Future directions

While Accurate™ already provides significantly more efficient workflow for data curation, several functionalities are in development, including typing of column headers to provide stronger data validation features, machine learning algorithms for tagging of free-form text with ontology terms occurring in it as well as on-the-fly language translation in ontology mapping.

References

Cortes A et al., Bayesian analysis of genetic association across tree-structured routine healthcare data in the UK Biobank. Nat Genet. 2017 Sep;49(9):1311-1318.
Pang C et al., SORTA: a system for ontology-based re-coding and technical annotation of biomedical phenotype data. Database (Oxford). 2015 Sep 18;2015.
Whetzel PL et al., BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications.
Nucleic Acids Res. 2011 Jul;39(Web Server issue):W541-5.