Bringing Schemas to the Schemaless

Mark A. Miller [1], Rebecca C. Jackson [2] and Christopher J. Mungall [1]

[1] *Lawrence Berkeley National Laboratory, Berkeley, CA, USA*
[2] *Bend Informatics, Bend, OR, USA*

Abstract

Ad-hoc spreadsheets can be appealing for managing some kinds of scientific data, but they allow bad practices like inconsistent types (e.g., string, integer, decimal) within individual columns, as well as open-ended categorical values (e.g., cell shapes). We have developed tools that infer LinkML schemas from spreadsheets, asserting types and enumerated lists that can be used to validate the dataset as it grows. Our tools can also associate OBO Foundry term identifiers with an enumeration's permissible values in the spirit of interoperability.

Keywords

LinkML, spreadsheets, schemas, enumerations

1. Introduction

As with many other industries, small to medium groups of scientists are likely to use a spreadsheet application for data collection in the early phases of a project. Spreadsheets with a single tab or worksheet offer a ubiquitous, familiar interface. They also enable convenient column-wise and row-wise copy and paste, plus formula-based data transformations.

There are many well-characterized disadvantages to using spreadsheets for scientific data management [1]. The initial flexibility they offer can be rapidly eclipsed by the lack of formal typing of columns. Distinguishing between numeric, data and text columns is generally not that difficult. However, the determination of whether a text column is intended to represent categorical (or enumerable) entities or open-ended narrative is outside the scope of routine spreadsheet usage.

We have developed a suite of tools as part of the LinkML framework to bootstrap discovering a semantic schema for an ad-hoc spreadsheet. When our methods infer that all values in one column come from a small set of permissible values, that column is modeled as an enum, and the strings can be mapped to terms from the OBO foundry. We utilize tools and standards that are from and by the OBO foundry developers' community, like the Ontology Lookup Service API [2], SSSOM [3], and rdftab [4]. Since our methods can read and write LinkML [5] files in addition to spreadsheets and delimited text files, they open the door to LinkML's support for numerous semantically oriented serialization and validation formats.

Outcomes of processing spreadsheets with our methods can include:

- Discussions that lead to handshake agreements about how new data will be added to the spreadsheet, along with sanitization of the existing data.
- The decision to use LinkML as the team's primary data tools for data management
- The design of a database and web frontend for managing the data. LinkML can generate SQL DDL for defining the necessary tables. Alternatively, LinkML provides validation frameworks for data stored in noSQL formats like JSON and RDF.

International Conference on Biomedical Ontologies 2021, September 16–18, 2021, Bozen-Bolzano, Italy
EMAIL: MAM@lbl.gov (A. 1); rbca.jackson@gmail.com (A. 2); cjmungall@lbl.gov (A. 3)
ORCID: 0000-0001-9076-6066 (A. 1); 0000-0002-6601-2165 (A. 2); 0000-0002-4232-6880 (A. 3)

2. Background

We have developed a method for inferring a semantic model from the most common data serializations, including delimited text files. In the case of CSV or TSV files, it is assumed that each row represents one instance of a single class, each column header represents one predicate P, and the contents of each cell represent the object of a P relationship with that instance. Patterns observed in the columns can be used to specify the range of the predicates: all numerical columns would specify a numerical range, columns with high entropy would specify a very open type, like String, and columns with extensive re-use of a limited set of textual values would be interpreted as categorical objects, perhaps the labels of classes that are defined elsewhere.

Methods are provided for proposing term identifiers with the same semantics as the observed labels. Specifically, the one term that best accounts for the longest span of cell contents is injected back into the inferred model. We provide support for running term mappings against locally downloaded files or over REST APIs, like that provided by OLS. We use SQLite for storing ontology content locally, so installing a database engine is not required. While the results are almost guaranteed to include some false positives and false negatives, the inclusion of lexical similarity metrics enables an expert reviewer to rapidly curate the mappings.

Inferred models are serialized in YAML, following the LinkML framework. This serialization can be converted to a wide variety of formats familiar to ontologists, like RDF, OWL and JSON-LD, can also be used to generate Python Dataclasses, and can drive validation, when converted to SHEX or JSON schema. The mappings from categorical cell contents to term identifiers can also be saved as SSSOM tables.

3. Methods

Our methods can be found in the turbomam/schemas-for-schemaless GitHub repository. There, we illustrate the usefulness of this schema inference and mapping workflow with a table of microbial traits from bacteria-archaea-traits/bacteria-archaea-traits. This repository aggregates over twenty sources of knowledge about bacterial and archaeal traits. One use case for the tabular outputs from this repository, such as condensed_species_NCBI.csv, is the construction of a knowledge graph such as Knowledge-Graph-Hub/kg-microbe. If the knowledge graph is to succeed in sustainably linking even more data sources, string values like “flask” the “cell_shape” column should be normalized to identifiers like MICRO:0000406, which bears the label “pear-shaped cell” and the exact synonym “flask”.

We start with a Makefile that:

- clones the linkml/linkml-model-enrichment GitHub repository
- clones bacteria-archaea-traits/bacteria-archaea-traits

applies the infer_model.py script from linkml-model-enrichment/linkml_model_enrichment to the output/ condensed_traits_NCBI.csv file from bacteria-archaea-traits/output/ data file. An inferred schema in the LinkML format, serialized as a YAML file, results

- downloads the rdftab executable
- downloads ontologies that we expected to be applicable to the bacterial traits file: pato.owl and uberon.owl
- creates an SQLite database and populates it with statements from the downloaded OWL files, using rdftab

The remainder of the workflow is implemented in Python:

- term ID and label pairs are quickly retrieved from the rdftab SQLite database
- The inferred LinkML schema for condensed_traits_NCBI.csv is parsed, and the permissible values for the cell_shape column are extracted.
- The permissible values are sanitized with lowercasing, whitespace normalization and replacement of several punctuation characters.

- A first pass of mapping the cell_shape_enum permissible values to OBO foundry terms is performed by merging the sanitized values against the labels obtained from the rdftab SQLite database.
- Sanitized cell_shape_enum permissible values that don't match any labels from the rdftab SQLite database are submitted to the OLS search API
- Additional metadata for each matching class is retrieved with a. sperate API call in order to determine what kind of relationship justified the match (exact synonym from Genbank, close synonym, etc.)
- One best OLS hit for each Sanitized cell_shape_enum is retained, based on relevance and string similarity metrics.
- The term mappings from the rdftab SQLite and OLS paths are merged, and any Sanitized cell_shape_enum permissible values that failed to map are reported
- The mappings are injected back into the LinkML model

4. Results

condensed_traits_NCBI.csv consists of 172,325 taxon rows x 35 columns of identifiers, metadata and traits. infer_model.py automatically infers that the following columns take enumerable values:

- cell_shape
- data_source
- gram_stain
- metabolism
- motility
- rRNA16S_genes
- range_salinity
- range_tmp
- sporulation
- superkingdom

infer_model.py can be forced to consider other columns as enumerable with the --enum-columns flag. The cell_shape values in condensed_traits_NCBI.csv are already well normalized and the sanitization step does not shrink the number of strings that require mappings:

- bacillus
- branced
- coccobacillus
- coccus
- disc filament
- flask
- fusiform
- irregular
- NA
- pleomorphic
- ring
- spindle
- spiral

Four terms obtain mappings via merging with the rdftab database. (See Jupyter notebook for an accounting [6].) All of the remaining values obtain hits from the OLS search. However, seven values do not retrieve any acceptable mappings. For example, the cosine string similarity between the sanitized query and the best-matching annotation for the best matching term may be greater than our default cutoff of 0.05. Seven of the OLS matches come from MICRO and one each comes from OMP or NCBITAXON.

5. Discussion

Our method rapidly infers a schema from the bacterial and archaeal traits table and discovers columns that appear to take enumerable values. Two methods are provided and demonstrated for mapping the string representations of the enumerable values to terms from OBO Foundry ontologies. The selection of those two methods is informed by one's tolerance for downloading ontologies to local storage vs depending upon a REST API. Visual inspection of the mappings suggests that they are all sound. Mapping failures can be attributed to misspellings in the input (e.g., branced instead of branched) or limitations placed on the search space. For example, our OLS search was limited to NCBITAXON, MICRO and OMP. Had FMA been included, the cell morphology FMA:70989 would have been discovered for the sanitized input 'spindle'. Prioritizing quality over quantity, we were conservative in our selection of ontologies.

We have found that these methods work well for a variety of sources, including the INSDC BioSample metadata collection [7]. BioSample contains ~ 20 million records and ~ 500 entity types and attributes, although we have only mapped a few attributes from tens of thousands of records at any one time.

Qualitatively speaking, the sensitivity and specificity of mapping gene names and symbols to OBO Foundry terms has not been as good as it is for microbial traits, sequence features, and environmental features. We are considering the addition of searches against gene or protein specific databases outside of the OBO Foundry, like UniProt or Entrez Gene. Whether we would take a local download approach or a REST API approach needs to be determined.

6. Acknowledgements

This work is supported in part by the Genomic Science Program in the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research (BER) under contract number DE-AC02-05CH11231 (LBNL).

7. References

[1] Ghada AlTarawneh, Simon Thorne, A Pilot Study Exploring Spreadsheet Risk in Scientific Research, in: Proceedings of the EuSpRIG 2016 Conference "Spreadsheet Risk Management", pp. 49-69. ISBN: 978-1-905404-53-7. URL: https://arxiv.org/abs/1703.09785

[2] URL: https://www.ebi.ac.uk/ols/docs/api

[3] URL: https://github.com/mapping-commons/SSSOM

[4] URL: https://github.com/ontodev/rdftab.rs

[5] URL: https://github.com/linkml/linkml

[6] URL: https://github.com/turbomam/schemas-for-schemaless/blob/main/notebooks/schemas-for-schemaless.ipynb

[7] URL: https://www.ncbi.nlm.nih.gov/biosample/