The Linked Data Modeling Language (LinkML): A General-Purpose Data Modeling Framework Grounded in Machine-Readable Semantics

Sierra Moxon[1], Harold Solbrig[2], Deepak Unni[3], Dazhi Jiao[2], Richard Bruskiewich[4], James Balhoff[5], Gaurav Vaidya[5], William Duncan[1], Harshad Hegde[1], Mark Miller[1], Matthew Brush[6], Nomi Harris[1], Melissa Haendel[6], and Christopher Mungall[1]

[1] *Lawrence Berkeley National Laboratory, Berkeley, CA, USA*
[2] *Johns Hopkins University, Baltimore, MD, USA*
[3] *European Molecular Biology Laboratory, Heidelberg, Germany*
[4] *Star Informatics, Victoria, BC, Canada*
[5] *RENCI, Chapel Hill, NC, USA*
[6] *University of Colorado, Denver, CO, USA*

Abstract

Data integration is a major challenge in the life sciences, due to heterogeneity, complexity, the proliferation of ad-hoc formats and data structures, and poor compliance with FAIR guidelines. The Linked data Modeling Language (LinkML, https://linkml.github.io) is an object-oriented data modeling framework that aims to bring semantic web standards to the masses, simplifying the production of FAIR ontology-ready data. It can be used for schematizing a variety of kinds of data, ranging from simple flat checklist-style standards to complex interrelated normalized data utilizing polymorphism/inheritance. Although it is still a young and evolving standard, LinkML is already in use across a wide variety of projects with different applications including cancer data harmonization, environmental genomics, and knowledge graph integration.

Keywords

Ontology, semantic web, RDF, JSON-schema

1. Introduction

Data integration is a major challenge in the life sciences. In principle ontologies and semantic web formats can help address the problem of data integration, but these technologies are not sufficient in themselves. Having an ontology for a domain does not guarantee that data can be exchanged robustly, and semantic web standards are built on the open-world assumption, whereas for most database use cases closed-world constraints are required.

The Linked data Modeling Language (LinkML [1], https://linkml.github.io) is an object-oriented data modeling framework that aims to bring semantic web standards to the masses, simplifying the production of FAIR [2] ontology-ready data. It is intended to be used for schematizing a variety of kinds of data, ranging from simple flat checklist-style standards to complex interrelated normalized data utilizing polymorphism/inheritance. Although it is still a young and evolving standard, it is already in

International Conference on Biomedical Ontologies 2021, September 16–18, 2021, Bozen-Bolzano, Italy

EMAIL: smoxon@lbl.gov (A. 1); solbrig@jhu.edu (A. 2); deepak.unni3@gmail.com (A. 3); djiao@jhu.edu (A. 4), richard.bruskiewich@delphinai.com (A. 5); balhoff@renci.org (A. 6); gaurav@ggvaidya.com (A. 7); wdduncan@lbl.gov (A. 8); hhedge@lbl.gov (A. 9); mam@lbl.gov (A. 10); matt@tislab.org (A. 11); melissa@tislab.org (A. 12); cjmungall@lbl.gov (A. 13)

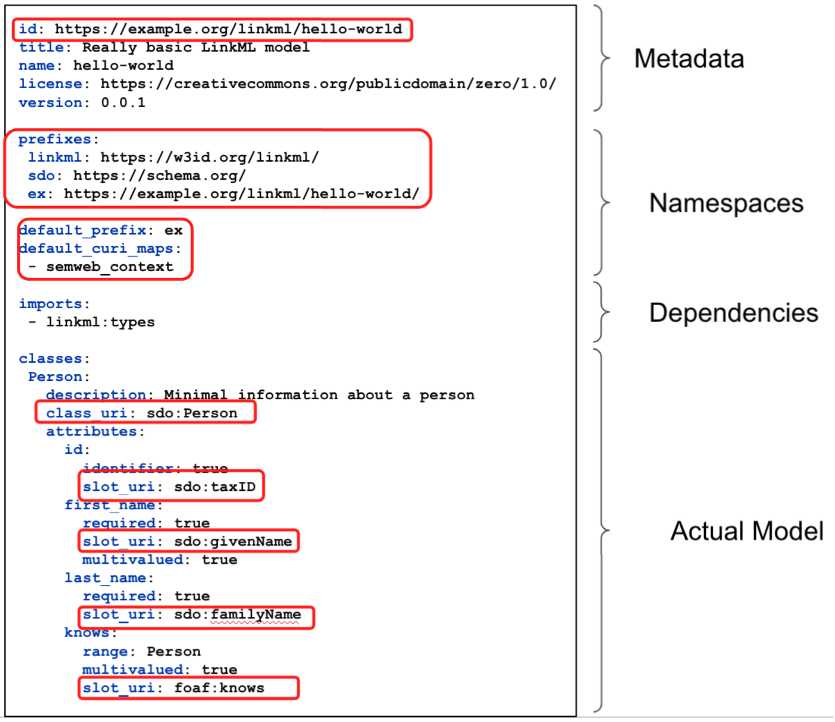
ORCID: 0000-0002-8719-7760 (A. 1); 0000-0002-5928-3071 (A. 2); 0000-0002-3583-7340 (A. 3); 0000-0001-5052-3836 (A. 4); 0000-0002-4447-5978 (A. 5); 0000-0002-8688-6599 (A. 6); 0000-0003-0587-0454 (A. 7); 0000-0001-9625-1899 (A. 8); 0000-0002-2411-565X (A. 9); 0000-0001-9076-6066 (A. 10); 0000-0002-1048-5019 (A. 11); 0000-0001-9114-8737 (A. 12); 0000-0002-6601-2165 (A. 13)

use across a wide variety of projects with different applications including cancer data harmonization, environmental genomics, and knowledge graph integration.

2. LinkML Structure

LinkML is designed to fit in well with frameworks familiar to most developers and database engineers -- JSON files, relational databases, document stores, Python object models -- and at the same time provide a solid semantic underpinning by mapping all elements to RDF URIs. LinkML's formal RDF-based framework allows semantics to hide in plain sight, while also making it easy for both domain and technical experts to design schemas in a shared platform.

An example of a simple schema represented in LinkML is shown in Figure 1.



```
id: https://example.org/linkml/hello-world
title: Really basic LinkML model
name: hello-world
license: https://creativecommons.org/publicdomain/zero/1.0/
version: 0.0.1

prefixes:
 linkml: https://w3id.org/linkml/
 sdo: https://schema.org/
 ex: https://example.org/linkml/hello-world/

default_prefix: ex
default_curi_maps:
 - semweb_context

imports:
 - linkml:types

classes:
 Person:
   description: Minimal information about a person
   class_uri: sdo:Person
   attributes:
     id:
       identifier: true
       slot_uri: sdo:taxID
     first_name:
       required: true
       slot_uri: sdo:givenName
       multivalued: true
     last_name:
       required: true
       slot_uri: sdo:familyName
     knows:
       range: Person
       multivalued: true
       slot_uri: foaf:knows
```

Figure 1: An example of LinkML syntax.

The basic structure is a schema plus associated metadata (including namespace to URI mapping), a set of classes, plus their attributes. Classes follow object-oriented semantics rather than OWL semantics, and classes can be metaclasses -- i.e., a LinkML schema can be used to model the design patterns in an ontology, with instances being OWL [3] classes. Each element in the schema can be assigned URIs from existing vocabularies, allowing for increased integration via semantic web standards.

LinkML favors ontologies over free text and gives information meaning by establishing identity via resolvable URIs. The framework allows the modeler to model both open and closed world assumptions, and when operating in a closed world, provides ways to validate and constrain schema instances and their relations (in a variety of different modeling paradigms like JSON-Schema, SQL-DDL, etc.). In addition, the LinkML language itself reuses existing semantic standards. For example, it provides modelers with a variety of mapping terms from the Simple Knowledge Organization System Namespace (SKOS) [4] (e.g., the *broad_mapping* relation, https://linkml.github.io/linkml-model/docs/broad_mappings.html, implements the *broadMatch* predicate, https://www.w3.org/2009/08/skos-reference/skos.html#broadMat). These formalisms allow the flexibility to extend or reuse existing object definitions while at the same time easily mapping data to existing standards where appropriate (e.g., a ‘gene’ object in one LinkML schema can be mapped directly to another LinkML schema’s representation of a ‘gene’ via ‘skos:exact_match’ predicates.).

LinkML tooling is another important piece of this framework. LinkML generators provide automatic translations from the schema YAML to a growing number of other formats, including:

- JSON-schema[5]
- JSON-LD/RDF[6]
- SQL DDL
- ShEx[7]
- GraphQL[8]
- Python data classes
- Markdown[9]
- UML diagrams[10]

This automated translation allows tooling from these frameworks to be easily reused and combined. For example, JSON-Schema provides robust validators, and these can be used for any LinkML schema.

The LinkML runtime provides loaders (https://github.com/linkml/linkml-runtime) and dumpers (https://github.com/linkml/linkml-runtime) to convert instances of the schema between these formats. And, because LinkML also generates (Python) class instances it provides a clear path to distributing data (via API or one of the formats native to LinkML like JSON, TSV, etc.) in the same well-defined format. LinkML tooling even auto-generates markdown documentation and UML diagrams from the schema YAML. The growing collection of LinkML schemas can be found at the LinkML schema registry (https://github.com/linkml/linkml-registry).

3. Use Cases

LinkML is already being used in a range of projects, including:

- National Microbiome Data Collaborative (https://microbiomedata.org/, https://github.com/microbiomedata/nmdc-schema), for storing environmental microbiome studies, associated samples, biogeochemical and environmental parameters, and associated omics datasets and function predictions
- Center for Cancer Data Harmonization (https://datascience.cancer.gov/data-commons/center-cancer-data-harmonization-ccdh, https://github.com/cancerDHC/ccdhmodel), for human patient and cancer sample data plus associated omics and imaging data
- The NCATS Biomedical Data Translator (https://ncats.nih.gov/translator, https://github.com/biolink/biolink-model), for integrating multiple knowledge graphs through the LinkML-authored Biolink schema
- The Alliance of Genome Resources (https://alliancegenome.org, https://github.com/alliance-genome/agr_curation_schema) for modeling complex model organism data for a persistent curation store
- The https://github.com/biodatamodels project, collecting schemas for core bioinformatics data formats, including GFF3

In summary, LinkML is a modeling framework that allows computers and people to work cooperatively: it is platform agnostic, compilable down to RDF, easy to use by both domain and technical experts, self-documenting and allows modelers to map common concepts to other well-defined resources and models. Most importantly, LinkML is a modeling framework that makes it easy to store, validate, and distribute data that is reusable and interoperable.

4. Acknowledgments

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). Additional support was provided by NIH OD R24 OD011883, NHGRI Center of Excellence in Genome Sciences RM1 HG010860, NHGRI 5U01HG009453-03, and NCI IAA #ACO21007-001-00000.

5. References

[1] URL: https://github.com/linkml/linkml
[2] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
[3] https://www.w3.org/TR/owl2-manchester-syntax/
[4] https://www.w3.org/2009/08/skos-reference/skos.html
[5] https://json-schema.org/
[6] https://shex.io/
[7] https://json-ld.org/
[8] https://graphql.org/
[9] https://www.markdownguide.org/
[10] https://www.uml-diagrams.org/