Brief Introduction to LinkML: Linked data Modeling Language

Sierra Moxon
Staff Software Developer - LBNL

OBO Summit - April 2023
LinkML is a modeling framework

THE STANDARD

A meta-datatype for structuring your data

TOOLS

Pragmatic developer and curator friendly tools for working with data

- Validators
- Data Converters
- Code Generation
- Data entry tooling
- Schema inference
Can I compare bacterial compositions of bodies of water?
Can I compare the bacterial compositions of samples taken from the epipelagic zones?
Can I compare bacterial compositions from salt water samples?
Common vocabularies are key

Pacific Ocean Sample Dataset

<table>
<thead>
<tr>
<th>id</th>
<th>depth</th>
<th>bacteria</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>PO1</td>
<td>22 cm</td>
<td>p</td>
<td>ENVO:00000015</td>
</tr>
</tbody>
</table>

Crater Lake Sample Dataset

<table>
<thead>
<tr>
<th>id</th>
<th>depth</th>
<th>bacteria</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL1</td>
<td>22 inches</td>
<td>x,p</td>
<td>ENVO:00000020</td>
</tr>
</tbody>
</table>
Existing frameworks not designed for interop

Pacific Ocean Sample Dataset

```
CREATE TABLE biosample (  
  acc varchar primary key,  
  depth float,  
  lat float,  
  long float,  
  ...  
)
```

```
CREATE TABLE lake_sample (  
  id varchar primary key,  
  depth foreign key,  
  location foreign key,  
  environment foreign key  
  ...  
)
```
We have many frameworks to structure data
LinkML for schemas

Biocurator
Data Scientist

dct:creator

OWL
ShEx, SHACL
JSON-LD
Contexts
{JSON}

JSON-Schema

Python
Dataclasses
SQL DDL
TSVs

PostgreSQL

pandas
Pacific Ocean Sample Database

<table>
<thead>
<tr>
<th>depth</th>
<th>salinity</th>
<th>bacteria</th>
<th>latitude</th>
<th>longitude</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>22 cm</td>
<td>35</td>
<td>x,p</td>
<td>44.808</td>
<td>24.0632° W</td>
<td>ENVO:00001999</td>
</tr>
</tbody>
</table>

**LinkML YAML specification**

```yaml
classes:
  BioSample:
    description: a sample of biological material.
    attributes:
      depth:
      salinity:
      bacteria:
        multivalued: true
      latitude:
        description: >-
          the angular distance of a place north or south of the earth's equator, or of a celestial object north or south of the celestial equator, usually expressed in degrees and minutes.
        type:
          required: true
      longitude:
      type:
```

LinkML
LinkML loves ontologies

classes:
    BioSample:
        description: a sample of biological material.
        attributes:
            depth:
                slot_uri: ENVO:3100031
            bacteria:
                multivalued: true
            salinity:
                exact_mappings:
                    - PATO:0085001
            longitude:
            latitude:
            environment:
LinkML provides validation tools

classes:
BioSample:
  description: a sample of biological material.
  attributes:
    depth:
      slot_uri: ENVO:3100031
    bacteria:
      multivalued: true
    salinity:
      exact_mappings:
        - PATO:0085001
    longitude:
    latitude:
  type:
    required: true
    range: EnvironmentEnum
ensems:
EnvironmentEnum
reachable_from:
  source_ontology: ENVO
Import LinkML models from other LinkML models

classes:
  BioSample:
    description: a sample of biological material.
    attributes:
      longitude:
      latitude:
      depth:
        slot_uri: ENVO:3100031

imports:
  - POSD:Biosample

LakeSample:
  is_a: BioSample
  attributes:
    salinity:
      exact_mappings:
        - PATO:00850001
LinkML Features

Modeling Language

- Import other LinkML models
- Inheritance and polymorphism
- Rule Language
  - Complex expressions
  - Boolean conditionals
  - Pattern Matching
- Enumeration support
- Mapping to other models
- Support for 3D array models

Additional Tools

- Validators
- Code generation (linkml-runtime)
  - Python, Typescript
  - Planned: Java, R
  - JSONSchema
  - Protobuf
  - GraphQL
  - SQL DDL
- Documentation Generation
- Spreadsheet->LinkML model automation: schemasheets
- Data curation tool generation: Data Harmonizer
- New project generator script: linkml-project-cookiecutter
  - metadata queries: schemaview
  - linkml-linter: best practices
Auto-generated
Hosted “for free” on GitHub
User friendly/human readable
Has schema diagrams
Navigate the model visually
Documentation is written in-place in the schema (write once)
Adoption

Tentative exploration:

https://linkml.io/linkml/examples.html
## Example: biosample datasets

<table>
<thead>
<tr>
<th>Lake Albert Sample Dataset</th>
<th>Pacific Ocean Sample Dataset</th>
<th>Crater Lake Sample Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>depth</strong></td>
<td><strong>bacteria</strong></td>
<td><strong>type</strong></td>
</tr>
<tr>
<td>22 cm</td>
<td>x</td>
<td>lake</td>
</tr>
<tr>
<td>23 cm</td>
<td>x, y, z</td>
<td>lake</td>
</tr>
</tbody>
</table>

Can I compare bacterial compositions of bodies of water?
Can I compare the bacterial compositions of samples taken from the epipelagic zones?
Can I compare bacterial compositions from salt water samples of epipelagic zones?
Common vocabularies are key
Common vocabularies are key

- **water body** (ENVO:00000063)
  - is-a **marine water body** (ENVO:00000015)
  - is-a **lake** (ENVO:00000020)
    - is-a **saline lake** (ENVO:00000019)
Example: biosample datasets

<table>
<thead>
<tr>
<th>Pacific Ocean Sample Dataset</th>
<th>Crater Lake Sample Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>depth</td>
<td>bacteria</td>
</tr>
<tr>
<td>22 cm</td>
<td>p</td>
</tr>
<tr>
<td>22 cm</td>
<td>p</td>
</tr>
</tbody>
</table>

wait…I need epipelagic zones… cm to meters to inches to feet?
Example: biosample datasets

Pacific Ocean Sample Dataset

<table>
<thead>
<tr>
<th>depth</th>
<th>bacteria</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>22 cm</td>
<td>p</td>
<td>ENVO:00001999</td>
</tr>
<tr>
<td>22 cm</td>
<td>p</td>
<td>ENVO:00001999</td>
</tr>
</tbody>
</table>

Crater Lake Sample Dataset

<table>
<thead>
<tr>
<th>depth</th>
<th>bacteria</th>
<th>type</th>
</tr>
</thead>
<tbody>
<tr>
<td>22 inches</td>
<td>x</td>
<td>ENVO:00000020</td>
</tr>
<tr>
<td>15 feet</td>
<td>x, p</td>
<td>ENVO:00000020</td>
</tr>
</tbody>
</table>

These are “standards” (and “models”), but they are not machine actionable, nor interpretable without a human.