

Brief Introduction to LinkML: Linked data Modeling Language



Sierra Moxon
Staff Software Developer - LBNL

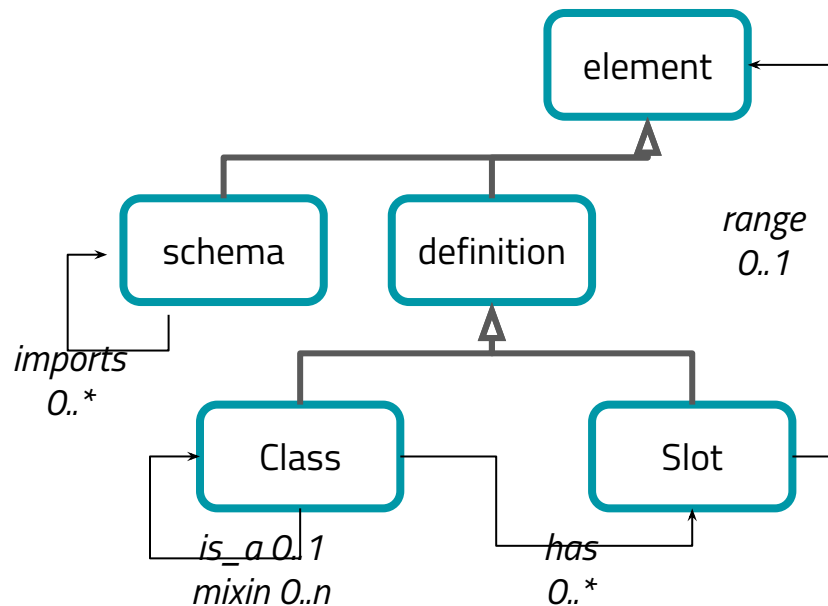
OBO Summit - April 2023



LinkML is a modeling framework

THE STANDARD

A **meta-datamodel** 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

Example: biosample datasets

Lake Albert Sample Dataset

depth	bacteria
22 cm	x
23 cm	x, y, z



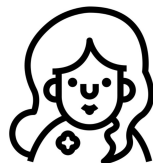
Pacific Ocean Sample Dataset

depth	bacteria
22 cm	p
22 cm	p



Crater Lake Sample Dataset

depth	bacteria
22 inches	x
15 feet	x, p



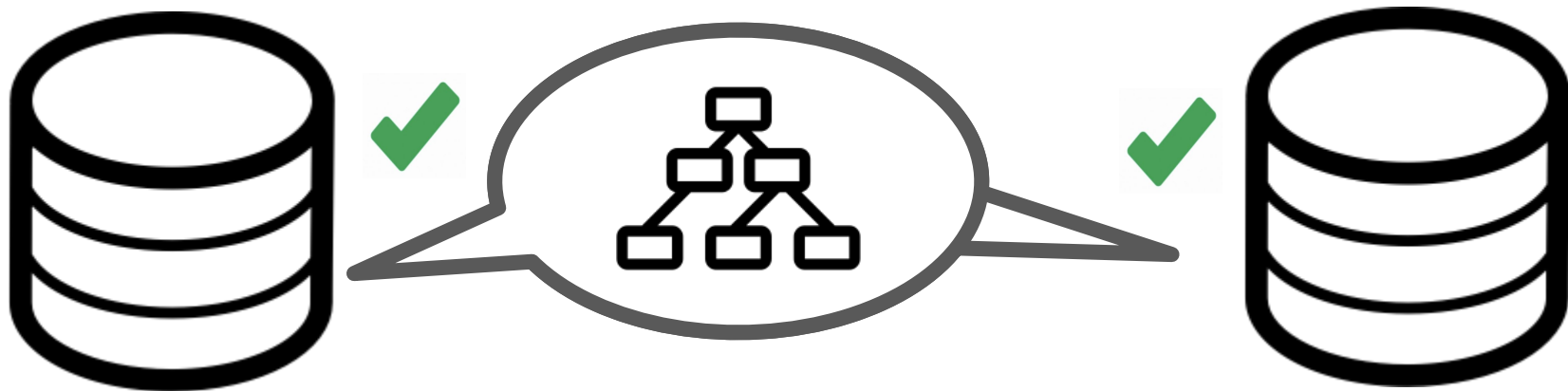
?

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



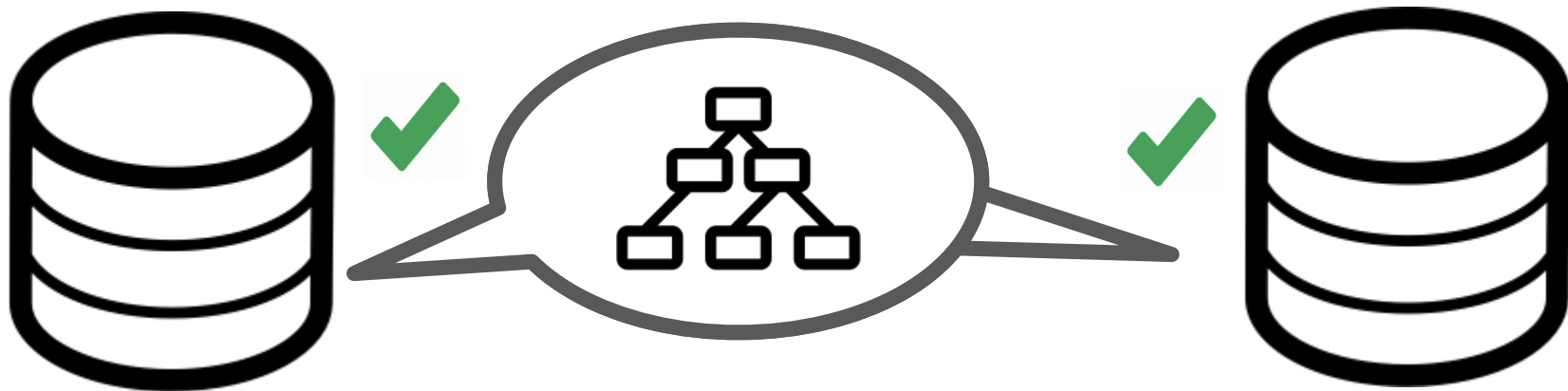
Pacific Ocean Sample Dataset

id	depth	bacteria	type
PO1	22 cm	p	ENVO:00000015

Crater Lake Sample Dataset

id	depth	bacteria	type
CL1	22 inches	x,p	ENVO:00000020

Existing frameworks not designed for interop



Pacific Ocean Sample Dataset

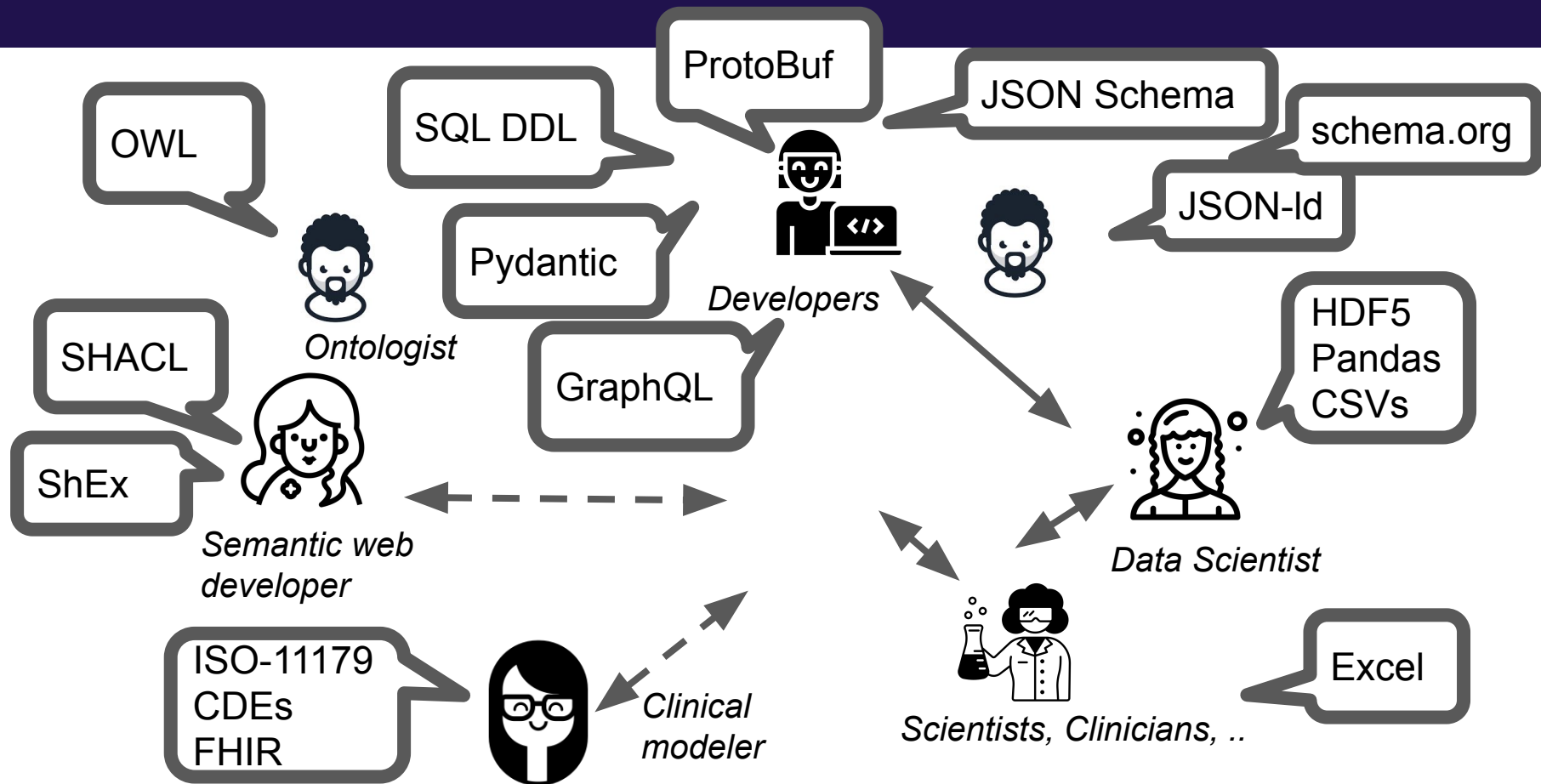
```
CREATE TABLE biosample (  
  acc varchar primary key,  
  depth float,  
  lat float,  
  long float,  
  ...  
)
```



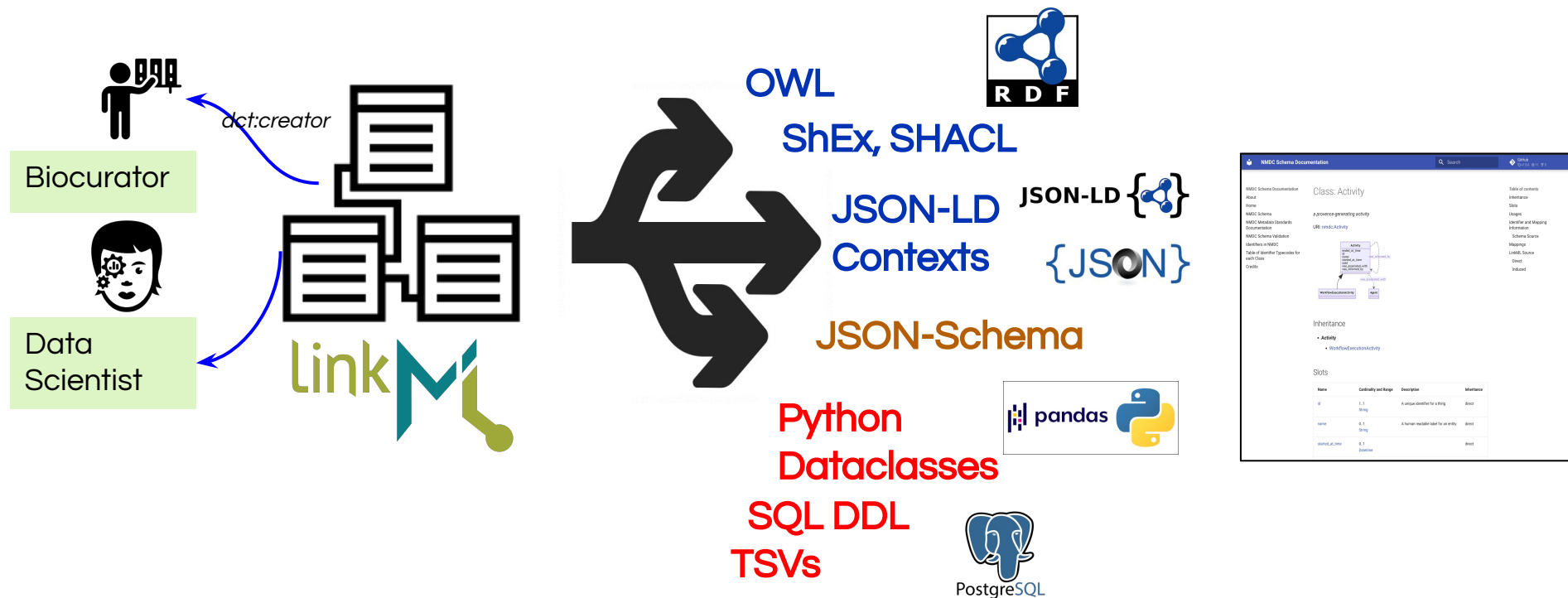
Crater Lake Sample Dataset

```
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



LinkML YAML specification

Pacific Ocean Sample Database

depth	salinity	bacteria	latitude	longitude	type
22 cm	35	x,p	44.808 4° N	24.0632° W	ENVO:0 0001999

LinkML

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.

longitude:

type:

required: true

LinkML loves ontologies

classes:

BioSample:

description: a sample of biological material.

attributes:

depth:

slot_uri: ENVO:3100031

bacteria:

multivalued: true

salinity:

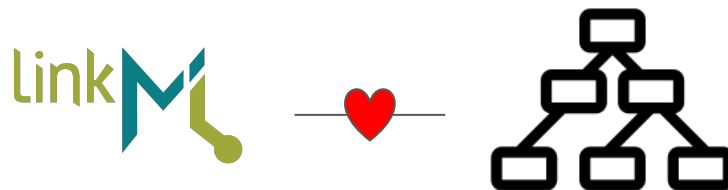
exact_mappings:

-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

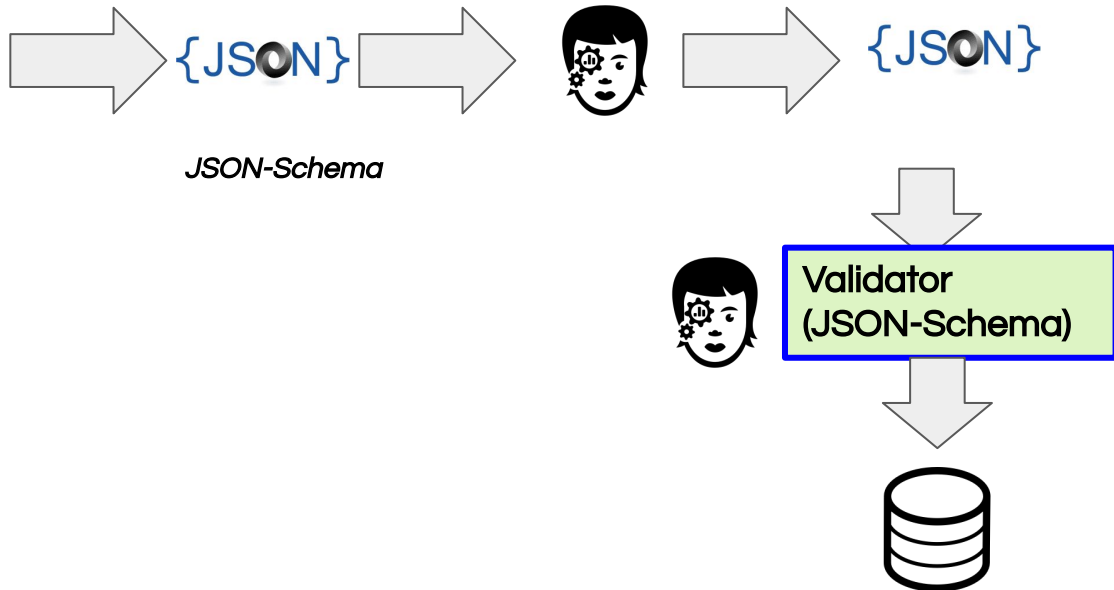
range: EnvironmentEnum

enums:

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:

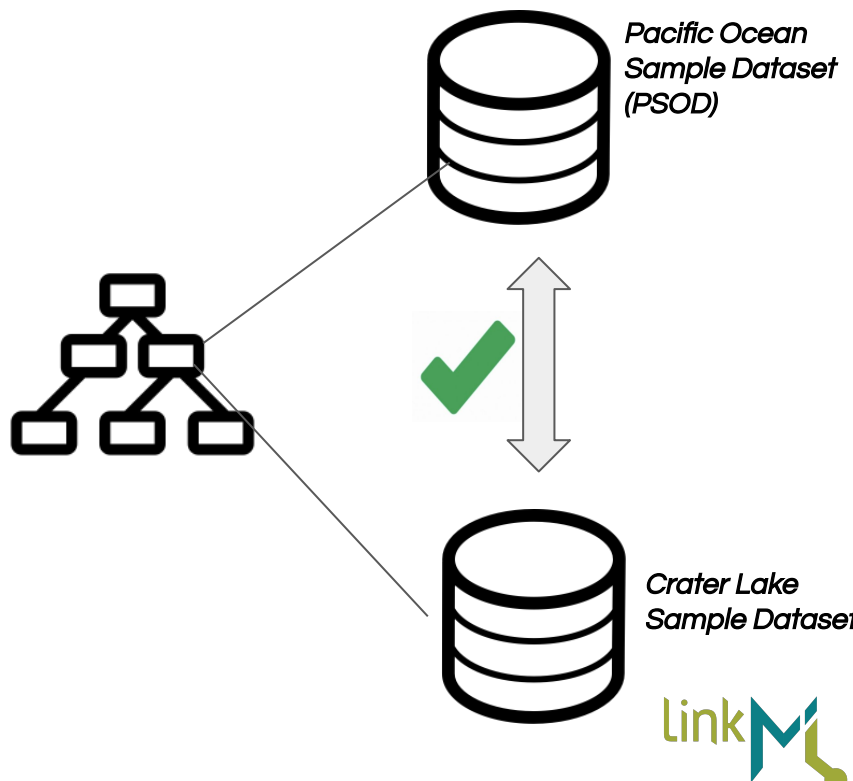
is_a: BioSample

attributes:

salinity:

exact_mappings:

-PATO:0085001



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

LinkML Documentation


[v7.5.0](#)
[☆ 11](#)
[¥ 5](#)

[Search](#)


[NMDC Schema Documentation](#)

[About](#)

[Home](#)

[NMDC Schema](#)

[NMDC Metadata Standards Documentation](#)

[NMDC Schema Validation](#)

[Identifiers in NMDC](#)

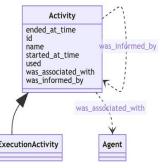
[Table of Identifier Typecodes for each Class](#)

[Credits](#)

Class: Activity

a *provenance-generating activity*

URI: `nmdc:Activity`



[Table of contents](#)

[Inheritance](#)

[Slots](#)

[Usages](#)

[Identifier and Mapping Information](#)

[Schema Source](#)

[Mappings](#)

[LinkML Source](#)

[Direct](#)

[Induced](#)

Inheritance

- Activity
 - WorkflowExecutionActivity

Slots

Name	Cardinality and Range	Description	Inheritance
<code>id</code>	1..1 String	A unique identifier for a thing	direct
<code>name</code>	0..1 String	A human readable label for an entity	direct
<code>started_at_time</code>	0..1 Datetime		direct

- 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)

Generate LinkML from Spreadsheets

Get intelligent assistance
from autoschema tools



refine

```
id: https://example.org/linkml/hello-world
title: Really basic LinkML model
name: hello-world
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
```

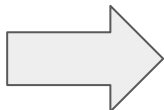
Metadata

Namespaces

Dependencies

Actual Datamodel

YAML conformant
to LinkML
standard



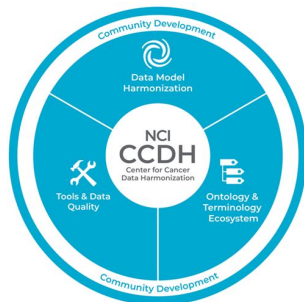
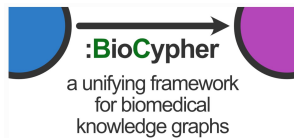
Schema-automator

Semi-structured
data sources

Adoption



iSamples
The internet of samples



Tentative exploration:



NEURODATA
WITHOUT BORDERS

<https://linkml.io/linkml/examples.html>

Example: biosample datasets

Lake Albert Sample Dataset

depth	bacteria	type
22 cm	x	lake
23 cm	x, y, z	lake

Pacific Ocean Sample Dataset

depth	bacteria	type
22 cm	p	ocean
22 cm	p	ocean

Crater Lake Sample Dataset

depth	bacteria	type
22 inches	x	lake
15 feet	x, p	lake



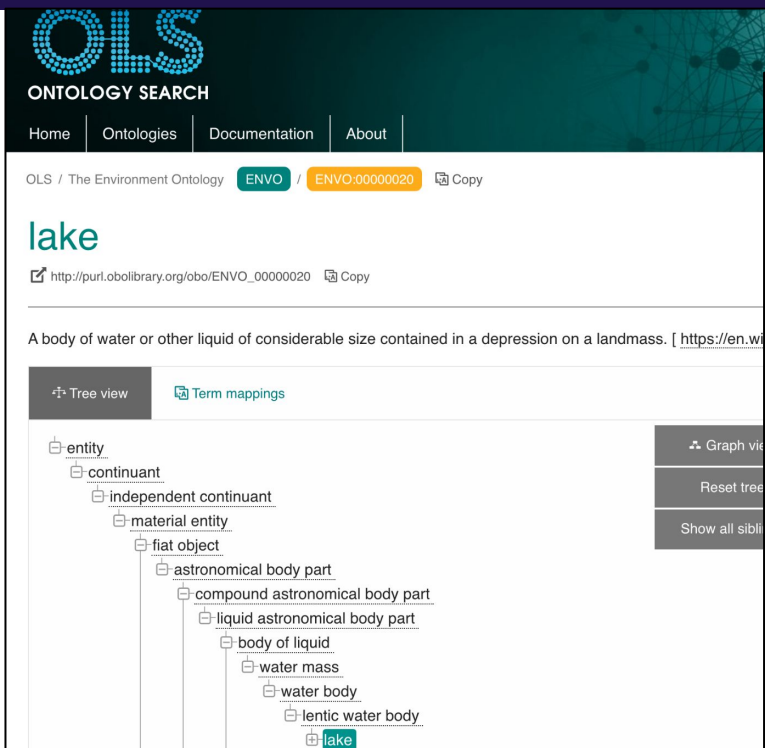
?

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



OLS
ONTOLOGY SEARCH

Home | Ontologies | Documentation | About

OLS / The Environment Ontology **ENVO** / **ENVO:00000020** 

lake

 http://purl.obolibrary.org/obo/ENVO_00000020 

A body of water or other liquid of considerable size contained in a depression on a landmass. [<https://en.wikipedia.org/wiki/Lake>]

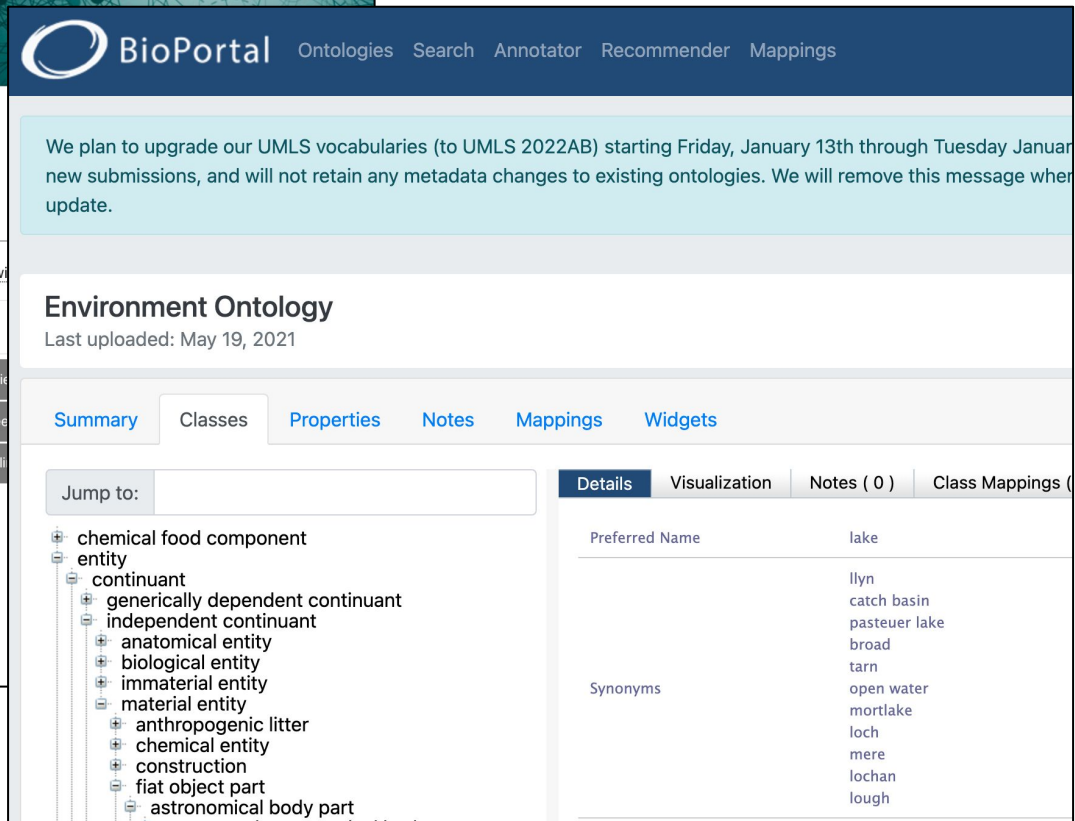
 Tree view  Term mappings

- entity
 - continuant
 - independent continuant
 - material entity
 - fiat object
 - astronomical body part
 - compound astronomical body part
 - liquid astronomical body part
 - body of liquid
 - water mass
 - water body
 - lentic water body
 - lake**

 Graph view

 Reset tree

 Show all siblings



BioPortal Ontologies Search Annotator Recommender Mappings

We plan to upgrade our UMLS vocabularies (to UMLS 2022AB) starting Friday, January 13th through Tuesday January 18th, 2022. During this time, we will accept new submissions, and will not retain any metadata changes to existing ontologies. We will remove this message when we complete the upgrade.

Environment Ontology

Last updated: May 19, 2021

Summary Classes Properties Notes Mappings Widgets

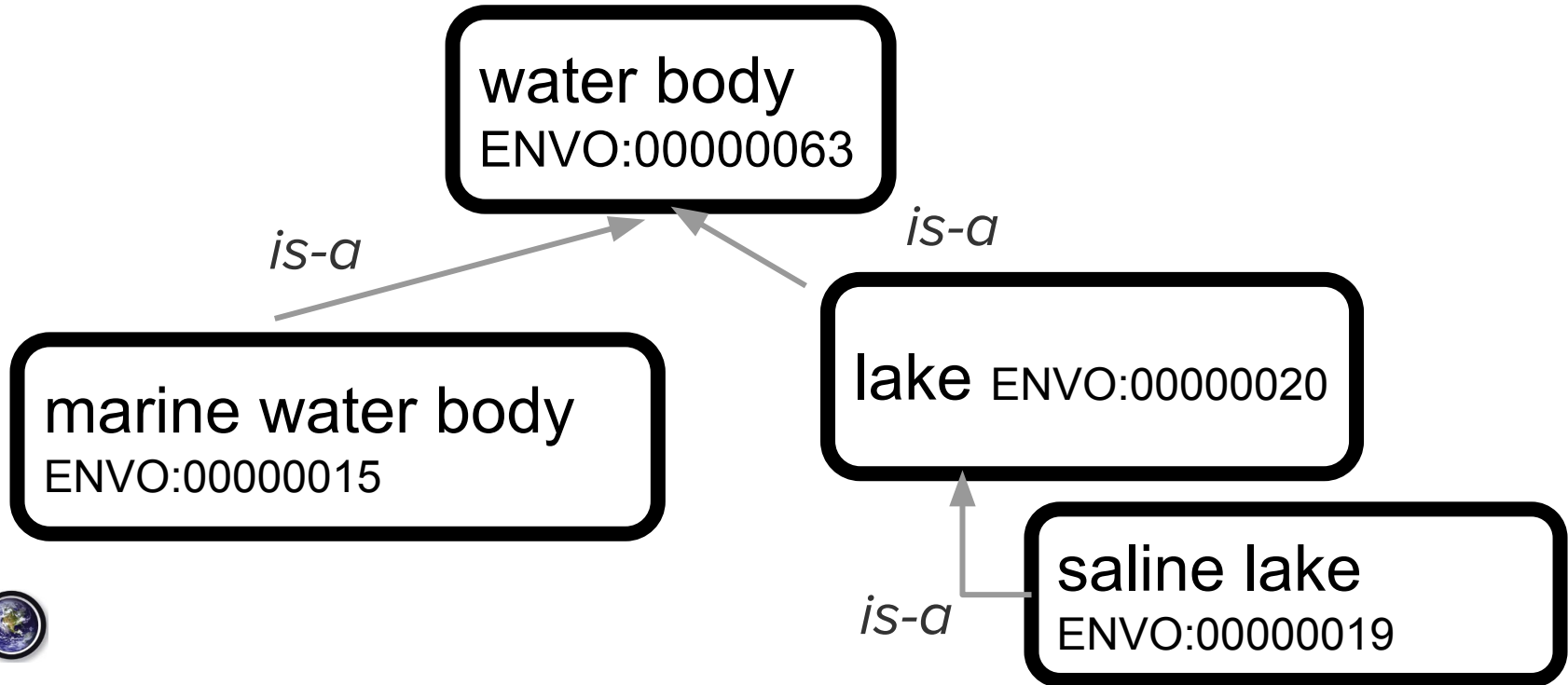
Jump to:

- chemical food component
- entity
 - continuant
 - generically dependent continuant
 - independent continuant
 - anatomical entity
 - biological entity
 - immaterial entity
 - material entity
 - anthropogenic litter
 - chemical entity
 - construction
 - fiat object part
 - astronomical body part

| Details | Visualization | Notes (0) | Class Mappings (0) |
|----------------|---------------|-------------|---|
| Preferred Name | | | lake |
| Synonyms | | | <ul style="list-style-type: none">llyncatch basinpasteur lakebroadtarnopen watermortlakelochmerelochanlough |



Common vocabularies are key



Example: biosample datasets

Pacific Ocean Sample Dataset

| depth | bacteria | type |
|-------|----------|---------------|
| 22 cm | p | ENVO:00001999 |
| 22 cm | p | ENVO:00001999 |

Crater Lake Sample Dataset

| depth | bacteria | type |
|-----------|----------|---------------|
| 22 inches | x | ENVO:00000020 |
| 15 feet | x, p | ENVO:00000020 |



wait...I need epipelagic zones... cm to meters to inches to feet?

| depth |
|---|
| N40.1164_W88.2543 |
| 25 <u>santimeters</u> |
| 0 – 20 cm |
| 3.149 |
| 30-60cm <u>replicate6</u> |
| Surface soil from deep water |
| Metamorph4 (19dpf) biological replicate 3 |

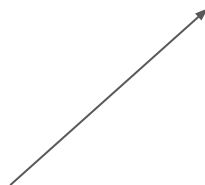
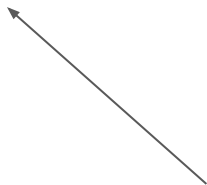
Example: biosample datasets

Pacific Ocean Sample Dataset

| depth | bacteria | type |
|-------|----------|---------------|
| 22 cm | p | ENVO:00001999 |
| 22 cm | p | ENVO:00001999 |

Crater Lake Sample Dataset

| depth | bacteria | type |
|-----------|----------|---------------|
| 22 inches | x | ENVO:00000020 |
| 15 feet | x, p | ENVO:00000020 |



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