LinkML for GSC MIxS

GSC-CIG Meeting April 2021

Chris Mungall
cjmungall@lbl.gov

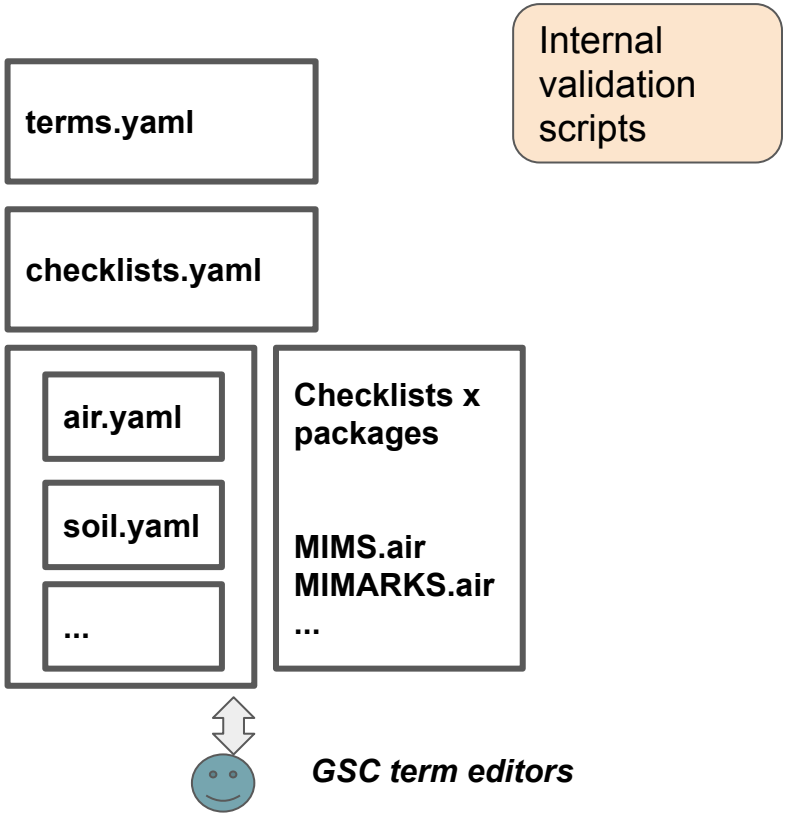
Motivation: FAIRifying MIxS

- Currently the MIxS standard is maintained as a spreadsheet
- Drawbacks for **maintenance**:
  - Repetition (denormalized)
  - Lack of internal automated QC
  - Terms not mapped to vocabularies
- Drawbacks for **users**
  - Lack of validators
  - Lack of web pages for each term and package / checklist
  - Submission tool builders don’t have computable input
  - Doesn’t play well with modern frameworks (JSON, RDF, …)

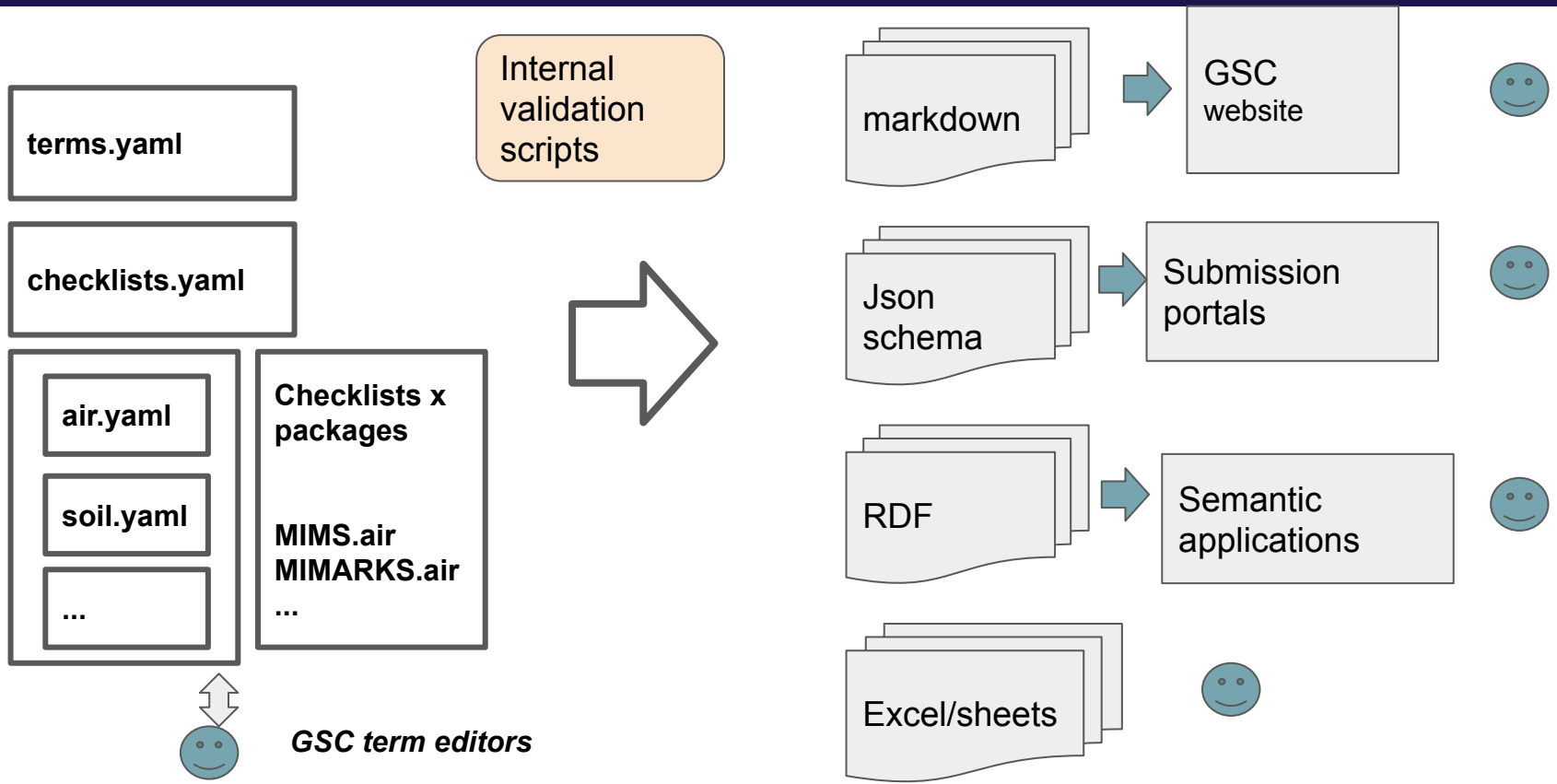
LinkML: A framework for data dictionaries

- LinkML is a way of specifying Schemas:
  - Datamodels
  - data dictionaries
  - check lists
- Can generate:
  - Markdown/HTML - to make a web site
  - JSON Schema - for developers
  - RDF, SQL DDL, ShEx, GraphQL, Excel, …
- LinkML is in production use for a range of projects
  - NCATS Data Translator
  - Gene Ontology
  - Alliance of Genome Resources
  - Center for Cancer Data Harmonization
  - **National Microbiome Data Collaborative**
    - We have been using our own LinkML rendering of MIxS

Manage files as YAML, export to other formats



Manage files as YAML, export to other formats



Demo Repo

main mixs-source / src / schema /

cmungall regenerated from latest sheets

..	
air.yaml	initial-commit
built_environment.yaml	initial-commit
checklists.yaml	regen
core.yaml	initial-commit
host_associated.yaml	initial-commit
human_associated.yaml	initial-commit
human_gut.yaml	initial-commit
human_oral.yaml	initial-commit
human_skin.yaml	initial-commit
human_vaginal.yaml	initial-commit
hydrocarbon_resources_cores.yaml	initial-commit
hydrocarbon_resources_fluids_swabs.yaml	initial-commit
microbial_mat_biofilm.yaml	initial-commit
miscellaneous_natural_or_artificial_environment.yaml	initial-commit
mixs.yaml	regen
plant_associated.yaml	initial-commit
ranges.yaml	initial-commit
sediment.yaml	initial-commit
soil.yaml	initial-commit
terms.yaml	regenerated from latest sheets

Seeded from mixs6 google sheets

(we would switch to YAML being source)

https://github.com/cmungall/mixs-source/
(this would move to GSC)

terms.yaml

```yaml
elev:
  is_a: environment field
  aliases:
  - elevation
  description: Elevation of the sampling site is its height above a fixed reference
    point, most commonly the mean sea level. Elevation is mainly used when referring
    to points on the earth's surface, while altitude is used for points above the
    surface, such as an aircraft in flight or a spacecraft in orbit
  range: quantity value
  examples:
  - value: 100 meter
  comments:
  - 'Expected value: measurement value'
  - 'Position: 7.0'
  - 'This field is used in: 9 packages: air, host-associated, hydrocarbon resources-cores,
    microbial mat_biofilm, miscellaneous natural or artificial environment, plant-associated,
    sediment, soil, water'
geo_loc_name:
  is_a: environment field
  aliases:
  - geographic location (country and/or sea,region)
  description: The geographical origin of the sample as defined by the country or
    sea name followed by specific region name. Country or sea names should be chosen
    from the INSDC country list (http://insdc.org/country.html), or the GAZ ontology
    (v 1.512) (http://purl.bioontology.org/ontology/GAZ)
  range: string
  examples:
  - value: Germany;North Rhine-Westphalia;Eifel National Park
  comments:
  - 'Expected value: country or sea name (INSDC or GAZ);region(GAZ);specific location
    name'
  - 'Position: 8.0'
  pattern: '{term};{term};{text}'
  see_also: https://github.com/GenomicsStandardsConsortium/mixs/issues/17
```

The format follows the LinkML standard

All terms (fields) in MIxS go in this file

These can be re-used in different packages and checklists

checklists.yaml

```yaml
1504          - conditional mandatory
1505    MIMAG:
1506      mixin: true
1507      description: Minimum Information About a Metagenome-Assembled Genome
1508      aliases:
1509      - MIMAG
1510      slots:
1511      - submitted_to_insdc
1512      - investigation_type
1513      - sample_name
1514      - project_name
1515      - experimental_factor
1516      - lat_lon
1517      - geo_loc_name
1518      - collection_date
1519      - env_broad_scale
1520      - env_local_scale
1521      - env_medium
1522      - env_package
1523      - ref_biomaterial
1524      - source_mat_id
1525      - rel_to_oxygen
1526      - sample_collect_device
1527      - sample_collect_method
1528      - samp_mat_process
1529      - size_frac
1530      - samp_size
1531      - nucl_acid_ext
1532      - nucl_acid_amp
1533      - lib_size
1534      - lib_reads_seqd
1535      - lib_layout
1536      - lib_vector
1537      - lib_screen
```

```yaml
1567      slot_usage:
1568        submitted_to_insdc:
1569          required: true
1570        investigation_type:
1571          required: true
1572        sample_name:
1573          required: true
1574        project_name:
1575          required: true
1576        experimental_factor:
1577          comments:
1578          - conditional mandatory
1579        lat_lon:
1580          required: true
1581        geo_loc_name:
1582          required: true
1583        collection_date:
1584          required: true
1585        env_broad_scale:
1586          required: true
1587        env_local_scale:
1588          required: true
1589        env_medium:
1590          required: true
1591        env_package:
1592          comments:
1593          - conditional mandatory
1594        ref_biomaterial:
1595          required: false
```

Fields can be declared required etc on a per-checklist level

Packages files

80 lines (80 sloc) | 1.5 KB

```yaml
id: http://w3id.org/mixs/soil
name: soil
imports:
- linkml:types
- terms
prefixes:
  linkml: https://w3id.org/linkml/
  mixs.vocab: https://w3id.org/mixs/vocab/
  MIXS: https://w3id.org/mixs/terms/
default_prefix: mixs.vocab
slots: {}
classes:
  soil:
    description: soil
    mappings: []
    slots:
    - lat_lon
    - depth
    - alt
    - elev
    - geo_loc_name
    - collection_date
    - env_broad_scale
    - env_local_scale
    - env_medium
    - cur_land_use
    - cur_vegetation
    - cur_vegetation_meth
    - previous_land_use
    - previous_land_use_meth
    - crop_rotation
    - agrochem_addition
    - tillage
    - fire
    - flooding
    - extreme_event
    - horizon
    - horizon_meth
    - sieving
    - water_content
    - water_content_soil_meth
    - samp_vol_we_dna_ext
    - pool_dna_extracts
    - store_cond
```

One yaml file per package (easier to manage?)

List all slots + overrides used in that package

Making a new package is easy - copy an existing file and change the slot list!

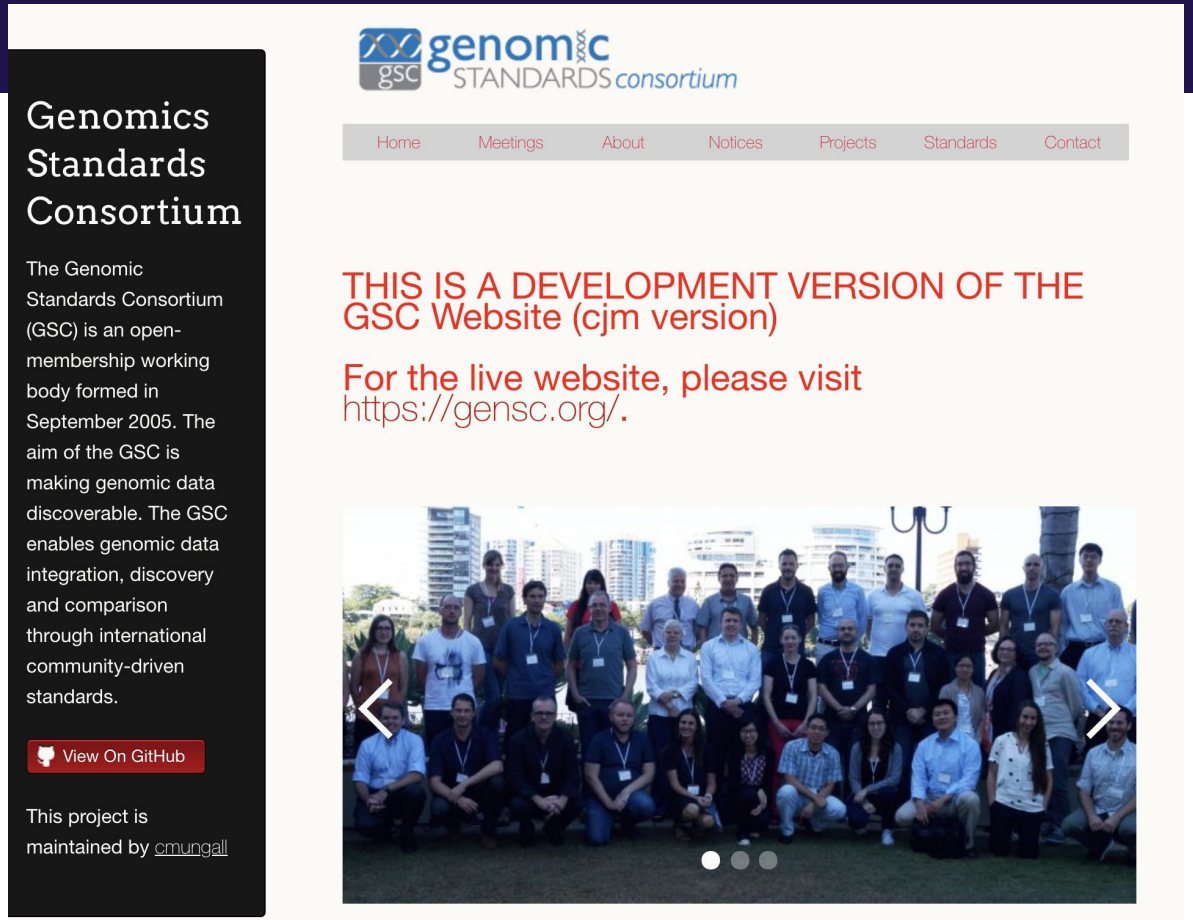
Combinatorics

```
2795    plant-associated MIMARKS survey:
2796      is_a: plant-associated
2797      mixins:
2798      - MIMARKS survey
2799      description: 'Combinatorial checklist Minimal Information about a Marker Specimen:
2800        survey with environmental package plant-associated'
2801    plant-associated MISAG:
2802      is_a: plant-associated
2803      mixins:
2804      - MISAG
2805      description: Combinatorial checklist Minimum Information About a Single Amplified
2806        Genome with environmental package plant-associated
2807    plant-associated MIMAG:
2808      is_a: plant-associated
2809      mixins:
2810      - MIMAG
2811      description: Combinatorial checklist Minimum Information About a Metagenome-Assembled
2812        Genome with environmental package plant-associated
2813    plant-associated MIUVIG:
2814      is_a: plant-associated
2815      mixins:
2816      - MIUVIG
2817      description: Combinatorial checklist Minimum Information About an Uncultivated
2818        Virus Genome with environmental package plant-associated
```

Enumerate all checklist x env package combos

Can be automated

Demo website



Current prototype: created by Chris H

Static site hosted with GitHub pages

Auto-generated pages easily added

Genomics Standards Consortium

The Genomic Standards Consortium (GSC) is an open-membership working body formed in September 2005. The aim of the GSC is making genomic data discoverable. The GSC enables genomic data integration, discovery and comparison through international community-driven standards.

View On GitHub

This project is maintained by cmungall

- Soil - soil
  - SoilME - Combinatorial checklist Metagenome or Environmental with environmental package soil
  - SoilMIGSBacteria - Combinatorial checklist Minimal Information about a Genome Sequence: cultured bacteria/archaea with environmental package soil
  - SoilMIGSEukaryote - Combinatorial checklist Minimal Information about a Genome Sequence: eukaryote with environmental package soil
  - SoilMIGSOrg - Combinatorial checklist Minimal Information about a Genome Sequence: org with environmental package soil
  - SoilMIGSPlant - Combinatorial checklist Minimal Information about a Genome Sequence: plant with environmental package soil
  - SoilMIGSVirus - Combinatorial checklist Minimal Information about a Genome Sequence: cultured bacteria/archaea with environmental package soil
  - SoilMIMAG - Combinatorial checklist Minimum Information About a Metagenome-Assembled Genome with environmental package soil
  - SoilMIMARKSSpecimen - Combinatorial checklist Minimal Information about a Marker Specimen: specimen with environmental package soil
  - SoilMIMARKSSurvey - Combinatorial checklist Minimal Information about a Marker Specimen: survey with environmental package soil
  - SoilMISAG - Combinatorial checklist Minimum Information About a Single Amplified Genome with environmental package soil
  - SoilMIUVIG - Combinatorial checklist Minimum Information About an Uncultivated Virus Genome with environmental package soil

- WastewaterSludge - wastewater/sludge
  - WastewaterSludgeME - Combinatorial checklist Metagenome or Environmental with environmental package wastewater_sludge
  - WastewaterSludgeMIGSBacteria - Combinatorial checklist Minimal Information about a Genome Sequence: cultured bacteria/archaea with environmental package wastewater_sludge

Experiment: I forked the GSC site and added generated markdown files

Note: this is the generic LinkML layout -- can be customized!

https://cmungall.github.io/gensc.github.io/mixs6

Page per term

cmungall.github.io/gensc.github.io/mixs6/elev.html

Genomics Standards Consortium

The Genomic Standards Consortium (GSC) is an open-membership working body formed in September 2005. The aim of the GSC is making genomic data discoverable. The GSC enables genomic data integration, discovery and comparison through international community-driven standards.

View On GitHub

This project is maintained by cmungall

Hosted on GitHub Pages

Slot: elev

Elevation of the sampling site is its height above a fixed reference point, most commonly the mean sea level. Elevation is mainly used when referring to points on the earth's surface, while altitude is used for points above the surface, such as an aircraft in flight or a spacecraft in orbit

URI: mixs.vocab:elev

Domain and Range

None -> OPT QuantityValue

Parents

- is_a: environment field

Children

Used by

- Air
- AirME
- AirMIGSBacteria
- AirMIGSEukaryote
- AirMIGSOrg
- AirMIGSPlant
- AirMIGSVirus
- AirMIMAG
- AirMIMARKSSpecimen

Note: this is the generic LinkML layout, we can customize

***For the semweb/FAIR people*:**

We can also use content-negotiation for RDF/JSON-LD access

Other generated files


```
{
  "$id": "http://w3id.org/mixs",
  "$schema": "http://json-schema.org/draft-07/schema#",
  "definitions": {
    "Air": {
      "additionalProperties": false,
      "description": "air",
      "properties": {
        "alt": {
          "$ref": "#/definitions/QuantityValue",
          "description": "Altitude is a term used to identify heights of objects such as airplanes, space sh
object which is above the earth's surface. In this context, the altitude measurement is the vertical distanc
        },
        "barometric_press": {
          "$ref": "#/definitions/QuantityValue",
          "description": "Force per unit area exerted against a surface by the weight of air above that surf
        },
        "carb_dioxide": {
          "$ref": "#/definitions/QuantityValue",
          "description": "Carbon dioxide (gas) amount or concentration at the time of sampling"
        },
        "carb_monoxide": {
          "$ref": "#/definitions/QuantityValue",
          "description": "Carbon monoxide (gas) amount or concentration at the time of sampling"
        },
        "chem_administration": {
          "description": "List of chemical compounds administered to the host or site where sampling occurre
ology (chebi) (v 163), http://purl.bioontology.org/ontology/chebi",
          "type": "string"
        },
        "collection_date": {
          "description": "The time of sampling, either as an instance (single point in time) or interval. In
01-23T19:23:10; 2008-01-23; 2008-01; 2008; Except: 2008-01; 2008 all are ISO8601 compliant",
          "type": "string"
        },
        "depth": {
```


JSON Schema

Generated RDF/OWL

mixs (http://w3id.org/mixs)

hydrocarbon resources-cores 〉 hydrocarbon resources-cores MIMAG

Active ontology × | Entities × | Individuals by class × | Individual Hierarchy Tab × | DL Query ×

Annotation properties | Datatypes | Individuals

Classes | Object properties | Data properties

Class hierarchy: hydrocarbon resources-cores MIMAG

Asserted

- host_sex_enum
- human-associated
- human-gut
- human-oral
- human-skin
  - human-skin ME
  - human-skin MIGS bacteria
  - human-skin MIGS eukaryote
  - human-skin MIGS org
  - human-skin MIGS plant
  - human-skin MIGS virus
  - human-skin MIMAG
  - human-skin MIMARKS specimen
  - human-skin MIMARKS survey
  - human-skin MISAG
  - human-skin MIUVIG
- human-vaginal
- hydrocarbon resources-cores
  - hydrocarbon resources-cores ME
  - hydrocarbon resources-cores MIGS bacteria
  - hydrocarbon resources-cores MIGS eukaryot
  - hydrocarbon resources-cores MIGS org
  - hydrocarbon resources-cores MIGS plant
  - hydrocarbon resources-cores MIGS virus
  - hydrocarbon resources-cores MIMAG
  - hydrocarbon resources-cores MIMARKS spec

hydrocarbon resources-cores MIMAG — https://w3id.org/mixs/

Annotations | Usage

Annotations: hydrocarbon resources-cores MIMAG

Annotations

rdfs:label

hydrocarbon resources-cores MIMAG

skos:definition

Combinatorial checklist Minimum Information About a Metagenome-Ass package hydrocarbon resources-cores

Class description | Taxon constraints

Description: hydrocarbon resources-cores MIMAG

Equivalent To

SubClass Of

- 'hydrocarbon resources-cores'
- adapters max 1 string
- annot max 1 string
- assembly_name max 1 string
- assembly_qual exactly 1 assembly_qual_enum
- assembly_software exactly 1 string
- bin_param exactly 1 bin_param_enum
- bin_software exactly 1 bin_software_enum
- compl_appr max 1 compl_appr_enum

Enums

```yaml
rel_to_oxygen:
  is_a: nucleic acid sequence source field
  aliases:
  - relationship to oxygen
  description: Is this organism an aerobe, anaerobe? Please note th
    anaerobic are valid descriptors for microbial environments
  range: rel_to_oxygen_enum
  examples:
  - value: aerobe
  comments:
  - 'Expected value: enumeration'
  - 'Position: 29.0'
  pattern: '[aerobe|anaerobe|facultative|microaerophilic|microanaer
    aerobe|obligate anaerobe]'
```

```yaml
rel_to_oxygen_enum:
  permissible_values:
    aerobe:
      description: dependent on oxygen
      meaning: PATO:0001455
    anaerobe:
      description: independent on oxygen
      meaning: PATO:0001456
    facultative: {}
    microaerophilic: {}
    microanaerobe: {}
    obligate aerobe:
      description: requires oxygen to grow
      meaning: ECOCORE:00000179
    obligate anaerobe:
      description: cannot grow in the presence of oxygen
      meaning: ECOCORE:00000178
source_uvig_enum:
```

SOP/Workflow

MIxS Maintainers <u>edit yaml files</u> managed in GitHub

- GitHub Actions takes care of
  - Validating
  - Generation of website, other files

Anyone in the community is free to make a Pull Request

- (But in general they would interact some other way)

See README for proposed SOP for

- New checklists
- New packages
- New terms
- New mappings of enums

Demo/Discussion

- **Can demo now if useful**

- **When to make the move?**
  - **Which GitHub pages template to use?**
    - **Current one does not support search**

- **Any developers able to help for customization?**

Credits

GSC CIG, RDF groups

- Ramona Walls
- Pier Luigi Buttigieg
- Bill Duncan
- Chris Hunter
- Ilene Mizraki
- Josie Burgin
- Lynn Schriml

NMDC Metadata team

- Bill Duncan
- Jagadish Sundramurthi
- David Hayes
- Sam Purvine
- Stan Martin
- Lee Anne McCue
- Montana Smith
- Pajau Vangay
- Elisha Wood-Charlson
- Emiley Eloe Fadrosh

LinkML framework

- Harold Solbrig (Johns Hopkins University)
- Dazhi Jiao (Johns Hopkins University)
- Deepak Unni (Berkeley Lab)
- Richard Bruskiewich (Star Informatics)
- Jim Balhoff (RENCI)
- William Duncan (Berkeley Lab)
- Harshad Hegde (Berkeley Lab)
- Mark Miller (Berkeley Lab)
- Sierra Moxon (Berkeley Lab)
- Donnie Winston (Berkeley Lab)
- Matthew Brush (OHSU)
- Nico Matentzoglu (Semanticly)
- Anne Thessen (Oregon State University)
- Melissa Haendel (University of Colorado)

nmdc
National Microbiome Data Collaborative

gsc genomic STANDARDS consortium