Leveraging Ontologies within the National Microbiome Data Collaborative

William D. Duncan[1], Faiza Ahmed[2], Fnu Anubhav[3], Jeffrey Baumes[2], Jonathan Beezley[2], Mark Borkum[3], Lisa Bramer[3], Shane Canon[1], Patrick Chain[4], Danielle Christianson[1], Yuri Corilo[3], Karen Davenport[4], Brandon Davis[2], Meghan Drake[5], Kjiersten Fagnan[1], Mark Flynn[4], David Hays[1], Bin Hu[4], Marcel Huntemann[1], Julia Kelliher[4], Sofya Lebedeva[1], Po-E Li[4], Mary Lipton[3], Chien-Chi Lo[4], Douglas Mans[3], Stanton Martin[5], Lee Ann McCue[3], David Millard[3], Kayd Miller[1], Nigel Mouncey[1], Paul Piehowski[3], Elais Player Jackson[4], Anastasiya Prymolenna[3], Samuel Purvine[3], TBK Reddy[1], Rachel Richardson[3], Migun Shakya[4], Montana Smith[3], Jagadish Chandrabose Sundaramurthi[1], Mark A. Miller[1], Deepak Unni[1], Pajau Vangay[1], Bruce Wilson[5], Donald Winston[6], Elisha Wood-Charlson[1], Yan Xu[4], Emiley Eloe-Fadrosh[1] and Christopher J. Mungall[1]

[1] *Lawrence Berkeley National Laboratory, Berkeley, CA, USA*
[2] *Kitware, Clifton Park, NY, USA*
[3] *Pacific Northwest National Laboratory, Richland, WA, USA*
[4] *Los Alamos National Laboratory, Los Alamos, NM, USA*
[5] *Oak Ridge National Laboratory, Oak Ridge, TN, USA*
[6] *Polyneme LLC, New York, NY, USA*

Abstract

The National Microbiome Data Collaborative (NMDC) is a multi-organizational effort to integrate microbiome data across diverse areas in environmental science. Data provided by the NMDC can then undergo advanced analysis and provide new insights into metagenomics, metatranscriptomics, metaproteomics, and metabolomics. To address these challenges, we have developed our schema using the Linked data Modeling Language (LinkML). This allows us to easily map data to existing standards and ontologies.

Keywords

Ontology, environmental science, environmental metagenomics

1. Introduction

The National Microbiome Data Collaborative (NMDC) is a multi-organizational effort to integrate microbiome data across diverse areas in environmental science. Data provided by the NMDC can then undergo advanced analysis and provide new insights into metagenomics, metatranscriptomics, metaproteomics, and metabolomics.

A major challenge for the NMDC is that data are heterogeneous and complex, and existing standards and ontologies are lacking or incomplete. To address these challenges, we have developed our schema using the Linked data Modeling Language (LinkML). This allows us to easily map data to existing standards where appropriate. This includes mapping of both schema elements and data values. For instance, in the NMDC schema, the LinkML syntax specification maps the NMDC term *biosample processing* to the Ontology for Biomedical Investigations' (OBI) term *material processing*.

International Conference on Biomedical Ontologies 2021, September 16–18, 2021, Bozen-Bolzano, Italy
EMAIL: wdduncan@gmail.com
ORCID: 0000-0001-9625-1899

```
biosample processing:
  aliases:
    - material processing
  is_a: named thing
  description: >-
    A process that takes one or more biosamples as inputs and generates one or
    as outputs.
  slots:
    - has input
  broad_mappings:
    - OBI:0000094
```

In this example, outputs of the *biosample processing* are not specified, since not every biosample process will necessarily have an output, and the mapping relation is a "broad mapping", meaning that the OBI term is more general that the NMDC term. LinkML's *broad_mapping* relation implements the *broadMatch* predicate from the Simple Knowledge Organization System Namespace (SKOS). Thus, by using SKOS predicates, the NMDC schema (via LinkML) leverages a well established standard for mapping terms.

For each biosample in the NMDC database, we record a number of important properties about the biosample's environment. We standardize this information in two ways. First, we utilize the *env_broad_scale*, *env_local_scale*, and *env_medium* terms defined by the Genomics Standards Consortium (GSC) MIxS (Minimal Information about any Sequence) standard (note that MIxS is in the process of migrating to LinkML). Roughly speaking, the *env_medium* defines the material containing the microorganism, the *env_local_scale* defines geographic features of the material, and the *env_broad_scale* defines the biosample's biome. Second, we use terms from the Environment Ontology (EnvO) to provide values for the aforementioned MIxS terms. For example, the following JSON formatted record from the NMDC database clearly shows the material, geographic feature, and biome of the biosample identified by the compact URI (i.e., CURIE) *gold:Gb0115217*:

```
{
   "id": "gold:Gb0115850",
    "env_medium": {"has_raw_value": "ENVO:00005802" },      # bulk soil
    "env_local_scale": { "has_raw_value": "ENVO:00000291"},  # drainage basin
    "env_broad_scale": { "has_raw_value": "ENVO:00000446"}, # terrestrial biome
    ...
}
```

In other words, the microorganisms within this biosample were found in a portion of bulk soil taken from a drainage basin in a terrestrial biome. Moreover, using EnvO's terms to define the environmental context of biosamples also permits us to leverage the ontology's semantics. For instance, since *bulk soil* is a kind of *soil*, and a *drainage basin* is a kind of (geographic) *depression*, we can make use of EnvO's hierarchy to find other kinds of soil (e.g., *dry soil*) that are found in other kinds of depressions (e.g., *dry lake*).

Finally, in the NMDC, we track many aspects of data provenance. This is especially important for computational workflows that produce files used for genomic analysis, such as metabolomics files. For this, we make use of the Provenance Ontology (PROV). In the following record, we use PROV's *wasGeneratedBy* predicate to specify that the file resulted of the activity identified as *nmdc:6fdeaf901c4c4c8fa19ec94696a2d03a*:

```
{
   "id": "nmdc:eadcd3b883c0da0f9d42a9fb1162ffcf",
   "name": "Froze_Core_2015_S2_0_10_7_Metab.csv",
   "description": "MetaMS GC-MS metabolomics output detail CSV file",
   "file_size_bytes": 565558,
```

```
    "md5_checksum": "eadcd3b883c0da0f9d42a9fb1162ffcf",
    "url": "https://nmdcdemo.emsl.pnnl.gov/metabolomics/results/
            Froze_Core_2015_S2_0_10_7_Metab.csv",
    "was_generated_by": "nmdc:6fdeaf901c4c4c8fa19ec94696a2d03a"
}
```


2. Acknowledgements

This work is supported by the Genomic Science Program in the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research (BER) under contract numbers DE-AC02-05CH11231 (LBNL), 89233218CNA000001 (LANL), DE-AC05-00OR22725 (ORNL), and DE-AC05-76RL01830 (PNNL).