Enhanced metadata standards supported by the National Microbiome Data Collaborative

Chris Mungall[1]* (cjmungall@lbl.gov), Faiza Ahmed[2], 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], William Duncan[1], Kjiersten Fagnan[1], Mark Flynn[4], David Hays[1], Bin Hu[4], Marcel Huntemann[1], Julia Kelliher[4], Sonya 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], Deepak Unni[1], Pajau Vangay[1], Bruce Wilson[5], Donny Winston[6], Elisha Wood-Charlson[1], Yan Xu[4], **Emiley Eloe-Fadrosh**[1]

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

https://microbiomedata.github.io/nmdc-metadata/

Project Goals: Short statement of goals. (Limit to 1000 characters)
The National Microbiome Data Collaborative (NMDC) is a pilot initiative launched to support microbiome data exploration and discovery through a collaborative, integrative science gateway. With a community-centered design approach, the NMDC team is building an open-source, integrated data science ecosystem that leverages existing data standards, data resources, and infrastructure within the DOE complex.

Abstract

To understand microbiomes we need to integrate, analyze, and query large amounts of data, including multi-omics data (e.g., metagenome, metatranscriptome, metaproteome, and metabolome) and environmental data. This is challenging because these data are heterogeneous and complex, and existing standards and ontologies are lacking or incomplete.

For the NMDC project, we created a FAIR (findable, accessible, interoperable, and reusable) schema for handling data and metadata of multiple aspects of microbiome data, including environmental metadata about a sample and a study, metadata and provenance for all processing and workflows, and searchable information arising from annotation workflows (for example, functional annotations and results of binning).

The schema leverages and maps to existing standards where appropriate. For describing sample metadata and environmental characteristics, we leveraged the Genomics Standards Consortium (GSC) MIxS (Minimal Information about any Sequence) and use a combination of ENVO

(Environment Ontology) and GOLD used for classifying environments. This includes a mechanism for uniquely identifying source samples using identifier systems such as IGSN (International Geo Sample Number), allowing us to link together data from different omics processing pipeline connected to the same source sample. We extended the W3C PROV standard (https://www.w3.org/TR/prov-overview/) for metadata about computational workflows. For outputs of genomics/transcriptomics workflows, we built on standards such as GFF3, using standardized systems such as KEGG for functional annotation; and for metabolomics/metaproteomics we map to existing ontologies such as PSI-MS where possible.

Our schema weaves together these different standards into a coherent whole. It is rendered as JSON-Schema which allows for precise validation of data input streams using standard validators, as well . We also aim for FAIR compliance by also providing an RDF (Resource Description Framework) version of the schema, including mappings to existing standards.

We used this schema to integrate multiple diverse types of data into JSON-LD files, and to drive search in a web portal.

Funding statement.

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).