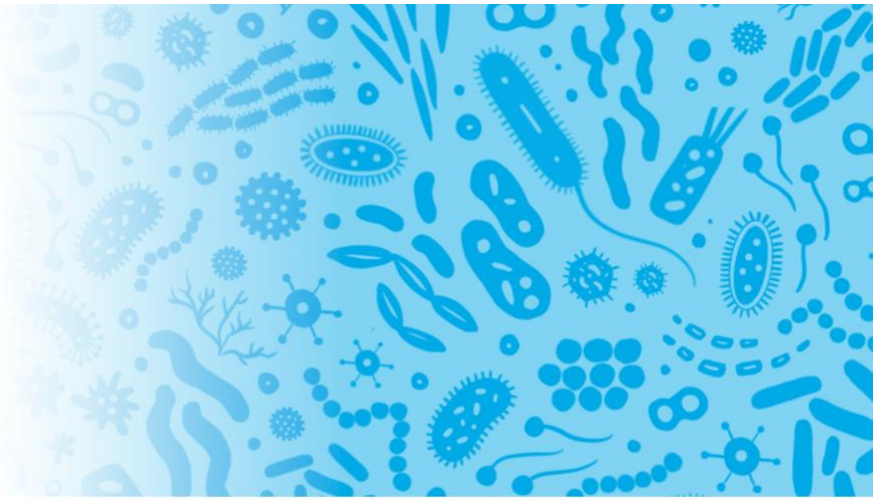

DOE BSSD Performance Management Metrics Report Q3

June 28, 2024

Investigators: Emiley A. Eloe-Fadrosh[1] (Lead, eaeloefadrosh@lbl.gov), Patrick S. G. Chain[2], Shreyas Cholia[1], Kjiersten Fagnan[1], Douglas Mans[3], Lee Ann McCue[3], Christopher J. Mungall[1], Nigel J. Mouncey[1]

Participating Institutions: [1]Lawrence Berkeley National Laboratory, Berkeley, CA 94720; [2]Los Alamos National Laboratory, Los Alamos, NM 87545; [3]Pacific Northwest National Laboratory, Richland, WA 99354.

Deliverable Q3: Report on the strategy to engage with other entities for access to data resources and/or modeling efforts for microbiome research.

Executive Summary

Microbiome data is complex, spanning information from microbial genomes within diverse communities, protein and metabolite readouts, and contextual information (metadata) captured from the environments from which these samples were collected. While the variety and scale of microbiome data generation has dramatically expanded over the past twenty years, infrastructure to support data management, sharing, and access has lagged. New ways to improve interoperability across existing resources and advancing community standards are necessary to support how researchers create, use, and reuse data. The National Microbiome Data Collaborative (NMDC) aims to advance a microbiome data sharing network through infrastructure, data standards, and community building.

The NMDC leverages a federated data model with multi-omics microbiome data and metadata hosted across various locations, and is centered around the NMDC schema to ensure microbiome data are findable, accessible, interoperable, and reusable (FAIR). Our schema serves as a unified data model that weaves together existing community standards and ontologies along with the use of persistent identifiers (PIDs) to provide globally unique, persistent, and machine resolvable identifiers to connect data objects created within the NMDC infrastructure (e.g., studies, samples, and workflow runs).

All NMDC data and metadata can be accessed through a user-friendly Data Portal *and* programmatically through a public Application Programming Interface (API). The NMDC API can be used broadly by the research community to query and access biosample and workflow outputs, and we have provided tutorials for researchers to learn how to use the NMDC API. NMDC's overall software architecture thus supports the programmatic exchange and linking of data across DOE's Biological and Environmental Research (BER) program User Facilities, the Joint Genome Institute (JGI) and Environmental Molecular Sciences Laboratory (EMSL). Our architecture also supports linking of data with BER resources, the Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) and DOE's Systems Biology Knowledgebase (KBase), towards a larger FAIR data ecosystem.

The NMDC serves as a data integration "hub" for standardized microbiome data for DOE's BER program and beyond. Microbiome data generated at EMSL and JGI are available through the NMDC in collaboration with the primary research teams. The Submission Portal supports both legacy and prospective studies to adhere to community standards and comply with community best practices. The NMDC also links to ESS-DIVE for archived environmental data and for future metabolic modeling efforts in KBase. Herein, we describe NMDC's strategy to engage additional data and modeling resources to maximize microbiome data accessibility and interoperability.

The NMDC infrastructure

The NMDC production platform leverages a federated data model with multi-omics microbiome data and metadata hosted across various locations including the Environmental Molecular Sciences Laboratory (EMSL) and Berkeley Lab's National Energy Research Scientific Computing Center (NERSC) (**Figure 1**). The NMDC user-facing software products: the Submission Portal (a user friendly spreadsheet-like web interface to enter and validate multi-omics metadata), NMDC EDGE (a web interface for external users to run NMDC standardized bioinformatics workflows), the Data Portal (a portal for data exploration and access through an integrated, distributed data framework aligned with the FAIR data principles), and the newly developed Field Notes app (a mobile application for metadata collection in the field), are derived from the NMDC schema and provide information to the central metadata database. This MongoDB database maintains references to the locations of the data and access protocols, allowing the data to be retrieved when requested by users. Programmatic access to the data is provided by the NMDC runtime services relying on schema-validated data stored with persistent identifiers.

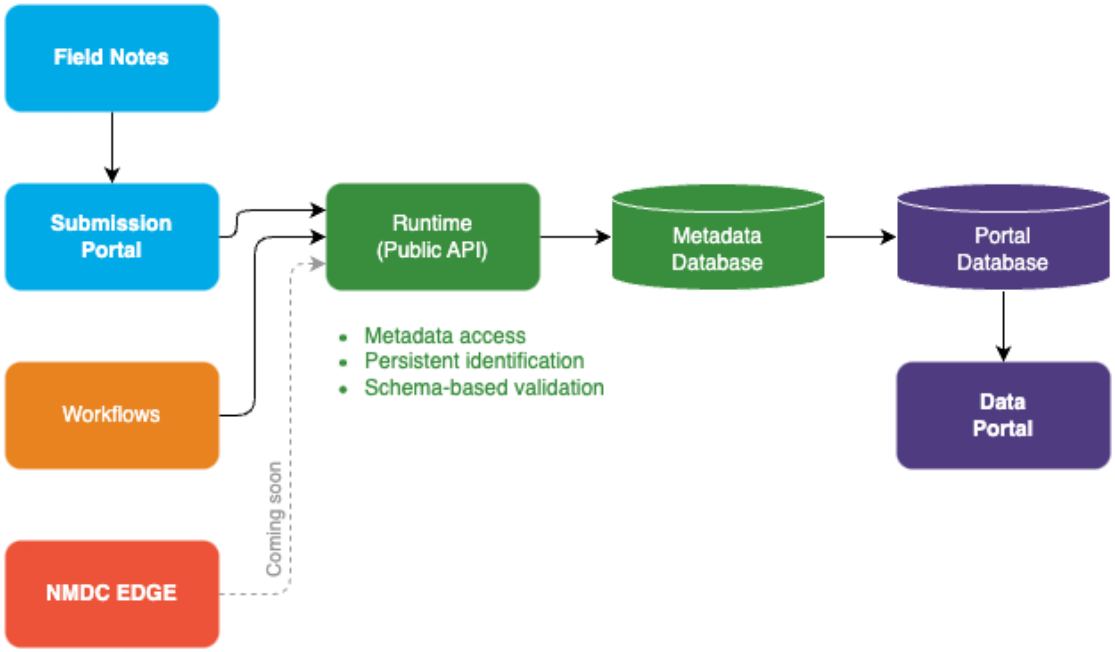

Figure 1. Overview of the NMDC infrastructure. The NMDC infrastructure consists of several interlinked components, including the user-facing software products (the Submission Portal, NMDC EDGE, the Data Portal, and the newly developed Field Notes app) and the backend data management services (NMDC runtime services, the MongoDB Metadata Database, and a Postgres Portal Database). The Submission Portal serves as the primary entry point for study and sample information from researchers. The Field Notes mobile application provides a specialized interface for entering this data from the field, and sends metadata to the Submission Portal, where researchers can refine it in preparation for submission. The Submission Portal submits data into the Metadata Database via the Public API of the NMDC Runtime. Processed multi-omics data are handled by workflows running at NERSC and EMSL, and registered through the Runtime API. Finally, an ingest pipeline pushes this metadata into the Portal Database, to be served up along with associated data through the Data Portal. Connections with workflows and metadata from NMDC EDGE are under development.

Schema-Guided Design

The NMDC schema serves as the foundation of our data infrastructure and ensures that microbiome data are findable, accessible, interoperable, and reusable (FAIR) [1]. This unified data model weaves together vocabulary selected from external resources, including the Environment Ontology (EnvO) [2] and the Genomic Standards Consortium's (GSC) Minimum Information about any (x) Sequence standard (MIxS) [3]. The NMDC schema provides a consistent, modular representation of the whole lifecycle of multi-omics analysis of environmental samples.

The schema, along with other NMDC components, follows best practices like monthly releases. It serves as a specification for many functionalities of NMDC's Application Programming Interface (API). The NMDC schema is written in the Linked Data Modelling Language (LinkML) and takes advantage of conveniences like automated documentation generation. Through additional human-curated documentation and contributor guidelines, NMDC is poised to support long-term data collaboration efforts.

Community Standards

As mentioned above, the NMDC schema benefits from employing community metadata standards and ontologies, notably the MIxS standard [3] and EnvO [2] for environmental terms, along with alignment with the Ontology for Biomedical Investigations (OBI) [4]. Our engagement with the Proteomics Standards Initiative and the Metabolomics Standards Initiative have guided our development of new classes for sample processing and data generation from mass spectrometry experiments. Engagement with these community-driven efforts has ensured interoperability of the data in the NMDC and has resulted in tangible benefits for the NMDC as well as for our partners.

Through our longstanding and productive partnership with the GSC, we have successfully transitioned the MIxS reporting standard [3] to a machine actionable format using LinkML. This entailed collecting the multiple different spreadsheets that had been used historically by the GSC, reconciling them and resolving differences and inconsistencies, and then translating this into a computable form. This work resolved a number of long-standing points of confusion, and put the GSC on track for improved management of the standard. It also allows for automated validation of data according to the standard. This past year, we have also improved access to the MIxS standard and its documentation. A description of the components of the MIxS standard and a practical example using the standard was published recently [5].

Persistent identifiers

Identifiers are crucial for the NMDC, both for the data objects we *create* and for the external objects we *reference.* We use internally generated persistent identifiers (PIDs) to provide globally unique, persistent, and machine resolvable identifiers to connect data objects created within the NMDC infrastructure (e.g., studies, samples, and workflow runs). These PIDs are defined by our schema and generated via an API endpoint, providing the unique digital reference to NMDC data objects that are the cornerstone for collaborating across data and modeling resources. In addition, by using ORCiD

identifiers, the NMDC is able to credit researchers with the data they have contributed to the NMDC, associate them with their professional contributions (e.g., data, publications, software, grants), and in the future, we will be able to connect to researcher's contributions stored by the many other resources adopting ORCiDs. NMDC's broad adoption of PIDs is increasing the discoverability of microbiome research, alleviating data validation issues and improving attribution and citation.

Data sharing across BER User Facilities and Resources

Our team closely collaborates with colleagues across DOE's Biological and Environmental Research (BER) program to support an ever-growing portfolio of microbiome research. While BER's User Facilities and resources have developed their own independent data management processes, there is increasing recognition that BER-funded microbiome researchers need solutions to seamlessly access multi-omics data across these systems to support their scientific goals and enable cross-study comparisons. Our team has focused on adopting persistent identifiers and advancing support for community standards as a means for data integration and interoperability with BER User Facilities and resources (**Figure 2**), as well as with research projects like the Science Focus Areas (SFAs). Ongoing and future activities are described below.

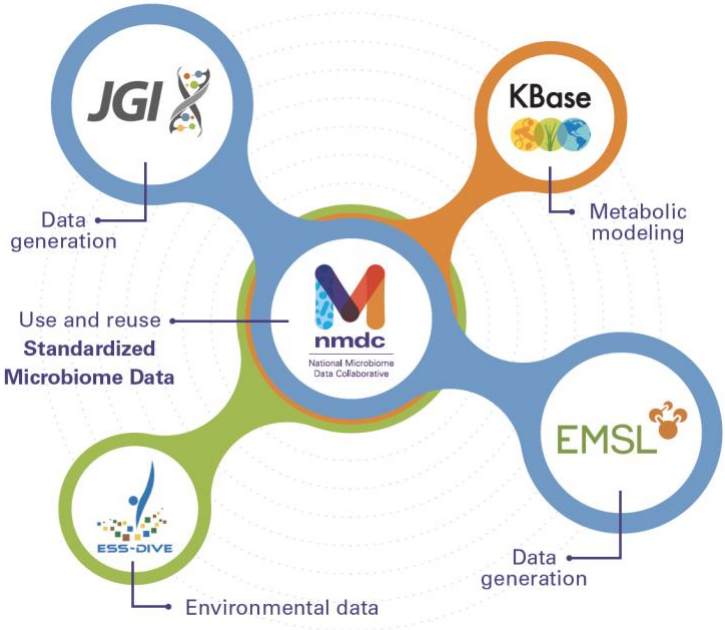

Figure 2. The NMDC serves as a data integration "hub" for standardized microbiome data for DOE's BER program and beyond. Microbiome data generated at the two user facilities, EMSL and the JGI, are made available through the NMDC Data Portal in collaboration with the primary research team. The Submission Portal supports both legacy and prospective studies to adhere to community standards and comply with user facility requirements. The NMDC also links to ESS-DIVE for archived environmental data and for future metabolic modeling efforts in KBase.

User Facilities: EMSL & JGI

As DOE BER User Facilities, the JGI and EMSL provide cutting-edge capabilities to a worldwide community of researchers studying microbiomes from a broad array of environmental systems. This user community can access (1) JGI for metagenomic and metatranscriptomic data generation, (2) EMSL for metaproteomic, metabolomic and natural organic matter data generation, or (3) the combined data generation capabilities

of both facilities through the Facilities Integrating Collaborations for User Science (FICUS) program. In addition to data generation, the JGI and EMSL embrace data management practices to ensure that the data are FAIR, including collection of metadata about samples, how they were processed, and the data generation protocols applied to the submitted biological samples.

The NMDC has collaborated closely with the JGI and EMSL on sample metadata standards, providing functionality in the Submission Portal to validate compliant metadata consistent with the sample submission requirements of both User Facilities. This includes supporting EMSL's sample management requirements and JGI's DNA and RNA quality metrics. Furthermore, the Submission Portal supports metadata harmonization across the JGI and EMSL, and compliance with the community metadata standards implemented by the NMDC. These Submission Portal improvements support both DOE User Facility researchers and the broader microbiome research community.

We also have a longstanding and highly productive collaboration across JGI's data management systems, the Genomes OnLine Database (GOLD) and Integrated Microbial Genomes & Microbiomes (IMG/M) platform. Last year, we developed an automated process to fetch study and sample metadata from the GOLD API and regularly share metadata updates across systems. We have also worked to share processed metagenome data more seamlessly with the IMG/M team for interoperability across the NMDC Data Portal and IMG/M. This year, the IMG/M team made use of the NMDC API to pull data into the IMG/M platform and built out the NMDC Metagenome Study List. Together, these efforts aim to harmonize and complement existing JGI data and comparative analysis services to support microbiome research.

Through our collaboration with EMSL, we are advancing the community data standards for natural organic matter data generated by high resolution mass spectrometry. Specifically, the NMDC has incorporated metadata standards for sample processing, data generation, and data processing developed by EMSL for the Molecular Observation Network (MONet) into the NMDC schema. As MONet data become available, EMSL and the NMDC will share these metadata via APIs and the NMDC Data Portal will point to the data locations in EMSL's archive.

Additionally, in collaboration with EMSL's Computing and Data Operations group, we have established a backup instance of the NMDC Data Portal and Submission Portal on EMSL's Kubernetes resources and initiated an allocation on Tahoma to provide compute hours for processing multi-omics data. We are able to ensure reliable access to microbiome data and resilience for data processing demands by leveraging these BER infrastructure resources.

ESS-DIVE

The NMDC team has worked with the ESS-DIVE team to develop methods for linking samples and metadata across data infrastructures. Together, we detailed these collaborations in a blog post this past year and have further engaged in these activities

with the broader environmental research community through the ESS-DIVE Open Data Workshop. Specifically, our coordinated efforts have used persistent identifiers to add links and references on relevant ESS-DIVE and NMDC landing pages to connect data across these systems. For example, original source sample International Generic Sample Numbers (IGSNs) from the System for Earth Sample Registration (SESAR) used in ESS-DIVE connect to NMDC's sample identifiers. Additionally, ESS-DIVE Dataset landing pages and NMDC Study landing pages reference JGI award DOIs, associated journal publication DOIs, and data DOIs to provide cross linkages across the systems (e.g., Genome Resolved Open Watershed study, East River Watershed study).

Further, team members at ESS-DIVE are collaborating with the NMDC team to test the sample metadata validation tool used by the Submission Portal (DataHarmonizer) with the goal of incorporating the ESS-DIVE sample identifier and metadata reporting format into the NMDC sample Submission Portal. This new effort will make it easier for researchers to submit Environmental System Science samples that have been assigned IGSNs to the NMDC, JGI, and EMSL to further support data sharing and interoperability.

KBase

Available multi-omics microbiome data in the NMDC can support advanced analyses and modeling of microbiomes. Towards this effort, we initiated collaboration with the KBase team to support data sharing of NMDC's multi-omics data within their narrative infrastructure. These initial efforts were prototyped in 2021 for metagenome data from six studies (https://narrative.kbase.us/#org/nmdc), yet presented challenges with maintaining data harmonization across systems. To overcome these challenges, the KBase team has initiated the development of a "Data Transfer Service" (DTS) with the JGI to prototype data sharing and preserve high-level credit and metadata information for attribution and provenance. We have recently started working with the KBase team to prototype the use of this solution with NMDC data. This will allow data in the NMDC to be directly accessed through the KBase platform and narrative infrastructure.

Further, the NMDC team is also engaged in proactive discussions with the KBase team for data modeling efforts and are exploring mechanisms to integrate the NMDC schema into a revised KBase architecture. We are also developing interactive notebooks that will serve as examples for how NMDC data and metadata can be used in the KBase platform for data analysis and modeling of microbiomes.

Supporting interagency partners in the environmental sciences

While it is essential to partner across BER user facilities and resources to advance microbiome data sharing, there is a much broader sphere of microbiome research spanning federal agencies, industry, and internationally. Below we describe our current and future efforts to support interagency partnerships.

The National Ecological Observatory Network (NEON)

In 2021, DOE and NSF established a memorandum of understanding that included the NMDC and the National Ecological Observatory Network (NEON). NEON pairs metagenome data for soil, benthic, and surface water samples with other information such as carbon flux measurements that are valuable for researchers interested in continental-scale environmental science. As part of our collaboration, we have developed tools for mapping NEON sample metadata to the community metadata standard terms in NMDC's schema. This work included leveraging interoperability between National Land Cover Database (NLCD) terms used by NEON and the Environment Ontology (EnvO) terms recommended by the GSC, as well as adding new terms to EnvO to better categorize aquatic samples. Once the metadata terms between NEON and the NMDC were mapped, the metagenome data for thousands of samples dating back to 2015 were processed with the NMDC workflows, and the metadata and processed data ingested into NMDC's Data Portal for broad accessibility. Highlighting this work, the NMDC and NEON teams co-hosted a workshop at the 2023 Ecological Society of America (ESA) Annual Meeting. This effort was also showcased in a recent NEON Data Skills Webinar.

In addition, NEON and the JGI have partnered to generate higher depths of sequencing for metagenome samples, for which NEON has been awarded a JGI CSP. As part of this partnership, NMDC's Submission Portal is being used to collect sample metadata. A more in-depth description of this partnership was highlighted in a recent blog.

The National Center for Biotechnology Information (NCBI)

When International Nucleotide Sequence Database Collaboration (INSDC) identifiers for BioProjects and BioSamples exist at NCBI, the NMDC schema stores these persistent identifiers in our metadata database. The NMDC team is also currently establishing a data submission protocol that will register new BioProjects and BioSamples with NCBI and submit sequencing data to NCBI's Short Read Archive (SRA). This effort will leverage the current metadata GSC submission packages made available at NCBI, will provide a service to researchers who aren't working with a DOE User Facility that manages submission to NCBI, and will crosslink the NMDC persistent identifiers with the NCBI INSDC identifiers.

Proteomic and Metabolomic Data Repositories

Data repositories for mass spectrometry-based proteomic and metabolomic data are adopting data DOIs to serve as persistent identifiers. For example, the NMDC schema stores DOIs minted by Mass Spectrometry Interactive Virtual Environment (MassIVE) data repository and Global Natural Products Social Molecular Networking (GNPS) system to crosslink metaproteomic and metabolomic data that are available on NMDC's Data Portal.

Towards future data integration to enable microbiome research

Our strategy to engage with DOE's BER User Facilities and data resources, along with broader interagency partnerships, has positioned the NMDC as a data integration "hub" for standardized microbiome data. Combined with our team's engagement with the BER

Advisory Committee (BERAC) activities over the past year to define infrastructure needs and prioritize recommendations (described in the workshop report "A Unified Data Infrastructure for Biological and Environmental Research: Report from the BER Advisory Committee"), the NMDC is poised to guide the next generation of data management and analysis framework for BER. Towards that end, our team has focused on advancing support for community standards across EMSL, the JGI, ESS-DIVE, and KBase as a means for data integration and interoperability.

References

1. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3: 160018.

2. Buttigieg PL, Morrison N, Smith B, Mungall CJ, Lewis SE, ENVO Consortium. The environment ontology: contextualising biological and biomedical entities. J Biomed Semantics. 2013;4: 43.

3. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nat Biotechnol. 2011;29: 415–420.

4. Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, et al. The Ontology for Biomedical Investigations. PLoS One. 2016;11: e0154556.

5. Eloe-Fadrosh EA, Mungall CJ, Miller MA, Smith M, Patil SS, Kelliher JM, et al. A Practical Approach to Using the Genomic Standards Consortium MIxS Reporting Standard for Comparative Genomics and Metagenomics. In: Setubal JC, Stadler PF, Stoye J, editors. Comparative Genomics: Methods and Protocols. New York, NY: Springer US; 2024. pp. 587–609.