

PNNL-35573

DOE BSSD Performance Management Metrics Report Q1

January 2024

Emiley Eloie-Fadrosh
Chris J. Mungall
Patrick Chain
Lee Ann McCue
Shreyas Cholia
Kjiersten Fagnan
Douglas M. Mans
Nigel J. Mouncey

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.** Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from
the Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062

www.osti.gov
ph: (865) 576-8401
fox: (865) 576-5728
email: reports@osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
or (703) 605-6000
email: info@ntis.gov
Online ordering: <http://www.ntis.gov>

DOE BSSD Performance Management Metrics Report Q1

January 2024

Emiley Eloie-Fadrosh
Chris J. Mungall
Patrick Chain
Lee Ann McCue
Shreyas Cholia
Kjiersten Fagnan
Douglas M. Mans
Nigel J. Mouncey

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354



DOE BSSD Performance Management Metrics Report Q1

January 16, 2024

Investigators: Emiley A. Elloe-Fadrosh¹ (Lead, eaelloefadrosh@lbl.gov), Patrick S. G. Chain², Shreyas Cholia¹, Kjersten Fagnan¹, Douglas Mans³, Lee Ann McCue³, Christopher J. Mungall¹, Nigel J. Mouncey¹

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

Deliverable Q1: Describe the overall background, strategy, and challenges to developing an online, open access microbiome data collection for the research community.

Executive Summary

Microbes play key roles in our biosphere, from driving global nutrient cycling to impacting plant, animal and human health and disease. Complex data from microbial genomes, proteins, and metabolites provide a window into these tiny engines that drive life on our planet. Yet these data are dispersed among researchers' laboratories and various repositories, making it difficult to access. This calls for new ways of managing data, improving data interoperability, advancing community standards, and creating an infrastructure where data are shared efficiently. We have built the National Microbiome Data Collaborative ([NMDC](#)) to advance how scientists create, use, and reuse data to redefine the way we understand and harness the power of microbes.

The vision of the National Microbiome Data Collaborative (NMDC) is to drive a microbiome data sharing network **connecting data, people, and ideas** to advance microbiome innovation and discovery. The NMDC was launched in 2019 and brought together DOE National Laboratories to collaborate across resources, capabilities, and expertise. The NMDC team was strategically assembled to include software developers, microbial researchers, metadata experts, and multi-omics specialists. The diversity of the NMDC team reflects the inherently interdisciplinary nature of microbiome science, and we leverage the strengths of the DOE National Laboratory system.

Towards BER's goal of advancing an iterative systems biology approach to the understanding of microbial genomes, the NMDC serves as a foundation for infrastructure, data standards, and community building. Together with the flagship DOE User Facilities, the Joint Genome Institute (JGI) and the Environmental Molecular Sciences Laboratory (EMSL), we are developing core capabilities in metadata standards for environmental descriptors and sample handling and processing; standardized bioinformatic workflows; an interface for data search and access; and robust community engagement activities.

The NMDC production platform supports long-term data infrastructure and community building for BER's bioenergy and environmental research goals. Our approach leverages lessons learned and an ambitious framework for collaborative, interdisciplinary data infrastructure to support microbiome research. The NMDC supports data, information, and knowledge access through three defined software tools – the [Submission Portal](#), [NMDC EDGE](#), and the [Data Portal](#) – driven by community needs. Herein, we describe the value proposition for the microbiome research community, our overarching strategy, and challenges and opportunities for developing the NMDC as both an infrastructure and community engagement program.

Background

Given the key role that microbes play in the health of every biosphere on Earth, and the intense research efforts funded by multiple United States government agencies, the 2018 Interagency Strategic Plan for Microbiome Research outlined three areas of focus for strategic investments [1]. Importantly, the Interagency Strategic Plan highlighted the need to develop platform technologies in support of open and transparent data sharing with a user-friendly, robust, and integrated system. This area of focus acknowledged that the volume of microbiome data that is now being generated by efficient high-throughput technologies is overwhelming the available infrastructure resources for collection, processing, and distribution of metagenome, metatranscriptome, metabolome, metaproteome, and lipidome data in an effective, uniform, and reproducible manner. Further, technical advances are needed for data and metadata standards, interoperable database systems for open data sharing and advanced analytical technologies that scale with the volumes of data that are being generated.

When the NMDC launched in 2019, our aim was to lay the groundwork for tackling these needs by focusing on both data sharing infrastructure and engagement with the scientific community (**Figure 1**). The three core infrastructure elements of the NMDC framework are: (1) the [Submission Portal](#) to support collection of standardized study and biosample information; (2) [NMDC EDGE](#), an intuitive user interface to access standardized bioinformatics workflows; and (3) the [Data Portal](#), a resource for consistently processed and integrated multi-omics data enabling search, access, and download [2]. Our engagement strategy includes partnerships with complementary data resources like DOE's Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) and DOE's Systems Biology Knowledgebase (KBase); partnerships with DOE User Facilities, the JGI and EMSL; coordinating with interagency programs outside of the DOE ecosystem such as NSF's National Ecological Observatory Network (NEON); and development of our flagship engagement programs, the NMDC [Ambassadors](#) and [Champions](#). This community-centric framework leverages unique capabilities, expertise, and resources available at the DOE National Laboratories to create an enabling environment for findable, accessible, interoperable, and reusable (FAIR) multi-omics microbiome data.

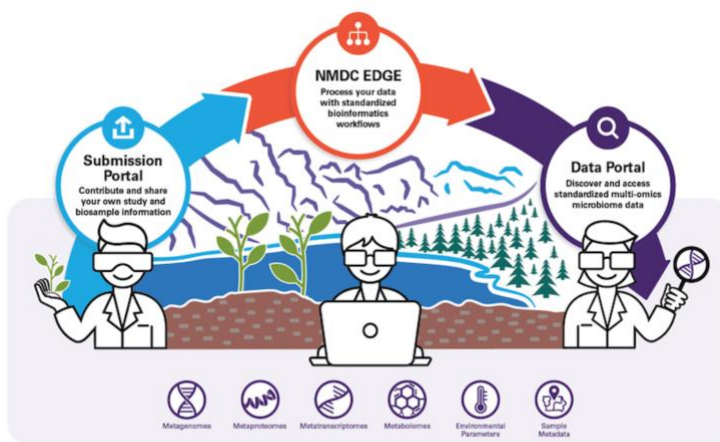


Figure 1. Overview of the National Microbiome Data Collaborative software tools to support the research community.

The NMDC's commitment to developing metadata and data standards for collection of samples and data generation will expand the use and reuse of generated data and promote collaboration and integration among researchers studying different ecosystems, thereby accelerating the discovery of underlying principles of microbiome function. The

NMDC Data Portal supports new ways for researchers to ask questions about how individual microbes and their genes, proteins, metabolites form a microbial community and drive environmental processes.

Strategy

Our strategy that underpins the vision and mission of the NMDC focuses on fostering strong community partnerships to advance our powerful products into tools that drive scientific impact. We have worked across stakeholders to identify the most pressing needs of the research community [3,4]. We continue to leverage user-centered design methodology to collect feedback from the scientific community and help prioritize the most important issues. Here, we outline our efforts to build a robust, yet flexible, software ecosystem for handling multi-omics microbiome data that includes our three products: the Submission Portal, NMDC EDGE and the Data Portal.

Adopting a framework for tracking tasks & implementation

In 2022, we developed a strategic roadmap that outlined 81 infrastructure and engagement milestones spanning a 3-year timeline. To organize specific tasks and ensure accountability for the implementation of the milestones, we adopted the Agile framework. Clear deliverables for each two-week sprint are tracked in the NMDC's public [GitHub](#), and we review all new and existing GitHub tickets before, during, and after each sprint. Additionally, to promote focused work on single deliverables, we adopted a loose version of *Product Squads* with the idea that these smaller, focused teams will be able to tackle infrastructure and user research milestones in a more productive manner. Each squad lasts multiple sprints and has clearly defined goals and outcomes which directly reference the NMDC milestones. Since the creation of the squads, 25 squads have completed with a total 557 tickets and 48 milestones completed. This organizational structure allows for transparency of NMDC activities with the community and NMDC stakeholders and aligns daily tasks with program priorities.

Building a robust software and data ecosystem

To deliver on a larger data ecosystem, we developed the [NMDC schema](#) to support handling and modeling of data. This foundation has enabled the weaving together of several different community standards, such as the Minimal Information about any (x) Sequence (MIxS) standard [5] from the Genomic Standards Consortium ([GSC](#)) and the Ontology of Biomedical Investigations (OBI) framework [6] for modeling sample processing and data generation. The schema was developed using the Linked Data Modeling Language ([LinkML](#)) which provides a robust, yet flexible solution to data modeling. A major accomplishment for us was teaming up with the GSC to translate their widely used standard from a spreadsheet into the LinkML framework ([version 6.2](#)), making the metadata machine operable and conversion between various formats for different tools straightforward, thus opening the doors to future AI applications.

Every [NMDC schema release](#) provides detailed information on schema improvements, updates, and contributor information. Further, we have recently initiated a major [update](#) of the schema to (i) better capture laboratory processing through generation of omics data

and (ii) better aggregate laboratory and analysis workflow processes. To our knowledge, this is the **first effort to fully model multi-omics microbiome data from sample to processed data**. It offers the research community a path towards data integration across studies and across computational platforms.

We are also leading the way towards FAIR through the NMDC persistent identifier [service](#) that was launched in January 2023. Persistent identifiers support links across studies, samples, and workflow runs in stable, unique, and long-term ways to reference digital objects, which fosters interoperability, attribution, and linking across resources.

Supporting a FAIR microbiome data sharing network

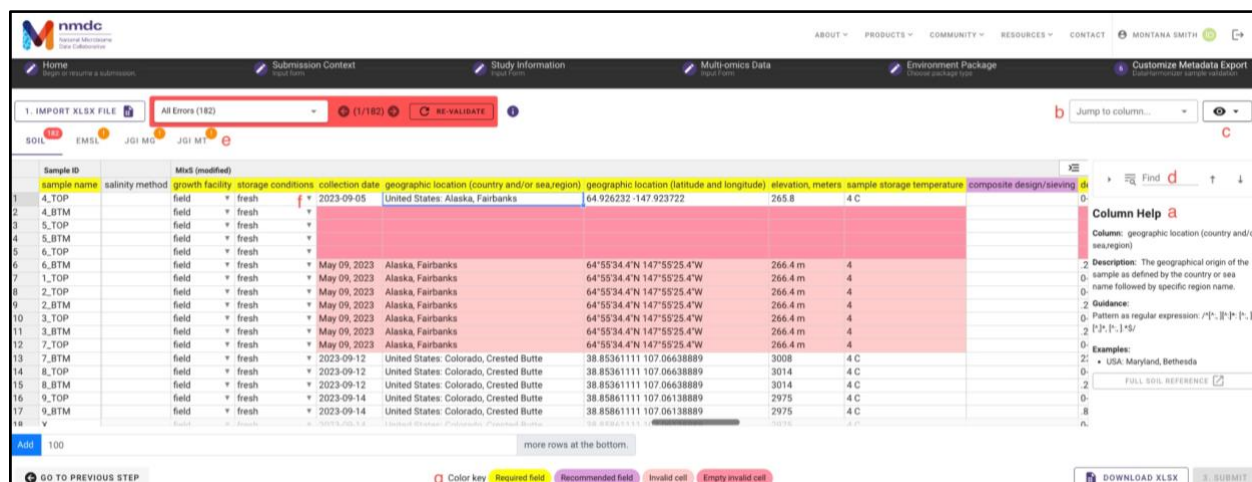
With the foundation of community standards and a robust data model, NMDC's three powerful software tools advance the creation, use, and reuse of data. Since the inception of the NMDC, [user-centered design](#) has been at the core of our products and mission. User-centered design places user feedback at the forefront of consideration. We have engaged extensively with the community to ensure we capture diverse perspectives to continuously improve our infrastructure and engagement strategies and meet our users' needs. Our user research efforts consist of asking researchers exploratory questions to collect information on researcher priorities, methodologies, and perceptions to ensure that we are aware of the current state of microbiome research. Our usability testing provides researchers with prototypes or live versions of the NMDC products, and we capture valuable information on how users interact to make improvements. This past year, we have adopted a standardized process for performing user-centered design activities that has resulted in 185 insights and 78 actions.



The NMDC Submission Portal: making it easy to follow standards

The Submission Portal is a flexible, template-driven tool designed to lower the barrier to collecting and reporting cohesive, standardized metadata about studies, samples, and assays (**Figure 2**). This tool was designed to make the capture and adherence to community standards easy for sample contextual information by leveraging the MlxS environmental extensions and the validation functions of the [DataHarmonizer](#) tool to check entered metadata values against the standards in the NMDC schema. The flexible framework **addresses the critical obstacle to FAIR microbiome data by easing the process of collecting and reporting the necessary metadata describing a study and biosamples**.

We created a comprehensive [user guide](#) and [tutorial](#) to guide researchers on how to use the MlxS environmental extensions and the real-time validation functions. To support data submission across the DOE user facilities, the JGI and the EMSL, the Submission Portal validates compliant metadata consistent with sample submission requirements. This includes supporting EMSL's sample management requirements and the JGI's DNA and RNA quality metrics. These user facility metadata requirements have been implemented in the Submission Portal through a new framework design that is intuitive and supports metadata harmonization across the JGI and EMSL.



Sample ID	Sample Name	Salinity Method	Growth Facility	Storage Conditions	Collection Date	Geographic Location (country and/or sea/region)	Geographic Location (latitude and longitude)	Elevation, meters	Sample Storage Temperature	Composite Design/Sieving
1	4_TOP	field	fresh	*	2023-09-05	United States: Alaska, Fairbanks	64.926232 -147.923722	265.8	4 C	
2	4_BTM	field	fresh	*						
3	5_TOP	field	fresh	*						
4	5_BTM	field	fresh	*						
5	6_TOP	field	fresh	*						
6	6_BTM	field	fresh	*	May 09, 2023	Alaska, Fairbanks	64°55'34.4"N 147°55'25.4"W	266.4 m	4	
7	1_TOP	field	fresh	*	May 09, 2023	Alaska, Fairbanks	64°55'34.4"N 147°55'25.4"W	266.4 m	4	
8	2_TOP	field	fresh	*	May 09, 2023	Alaska, Fairbanks	64°55'34.4"N 147°55'25.4"W	266.4 m	4	
9	2_BTM	field	fresh	*	May 09, 2023	Alaska, Fairbanks	64°55'34.4"N 147°55'25.4"W	266.4 m	4	
10	3_TOP	field	fresh	*	May 09, 2023	Alaska, Fairbanks	64°55'34.4"N 147°55'25.4"W	266.4 m	4	
11	3_BTM	field	fresh	*	May 09, 2023	Alaska, Fairbanks	64°55'34.4"N 147°55'25.4"W	266.4 m	4	
12	7_TOP	field	fresh	*	May 09, 2023	Alaska, Fairbanks	64°55'34.4"N 147°55'25.4"W	266.4 m	4	
13	7_BTM	field	fresh	*	2023-09-12	United States: Colorado, Crested Butte	38.85361111 107.06638889	3008	4 C	
14	8_TOP	field	fresh	*	2023-09-12	United States: Colorado, Crested Butte	38.85361111 107.06638889	3014	4 C	
15	8_BTM	field	fresh	*	2023-09-12	United States: Colorado, Crested Butte	38.85361111 107.06638889	3014	4 C	
16	9_TOP	field	fresh	*	2023-09-14	United States: Colorado, Crested Butte	38.85861111 107.06138889	2975	4 C	
17	9_BTM	field	fresh	*	2023-09-14	United States: Colorado, Crested Butte	38.85861111 107.06138889	2975	4 C	
18	V	field	fresh	*	2023-09-14	United States: Colorado, Crested Butte	38.85861111 107.06138889	2975	4 C	

Figure 2. The NMDC Submission Portal. Valid (white), invalid (pink), and missing (red) metadata support real-time validation of standardized and schema compliant metadata for samples. Key features are labeled as follows: (a) column help with descriptions and examples of required information and formatting, (b) jump to, (c) filtering, (d) 'find and replace' for easy access and organization, (e) a single place for metadata submission to DOE User Facilities, (f) enumerated drop downs for easy selection and completion, (g) and color coding for clear requirements and status.



NMDC EDGE: providing user-friendly bioinformatics tools

NMDC EDGE makes it easy to use standardized bioinformatics workflows, supporting big computing needs and reproducibility (**Figure 3**). The web application is designed to provide broad access to the NMDC workflows for all microbiome scientists. The lightweight architecture enables installation and configuration in different computing environments, thereby **addressing challenges in bioinformatics software portability, along with providing support for scalable compute resources**.

Standard workflows are available in NMDC EDGE for a variety of omics data types and are the same production workflows used at EMSL and the JGI. The metagenomics workflow provides end-to-end processing of sequence data from reads quality control to assembly, gene annotation, and taxonomy classification. In addition, workflows are available to categorize contigs into genomes from metagenome data (i.e., metagenome-assembled genomes, MAGs) and to identify viruses and plasmids by leveraging the recently released geNomad tool [7]. The NMDC metaproteomics workflow is an end-to-end data processing workflow for LC-MS/MS proteomics data. The workflow uses raw MS data and a matched metagenome to produce protein identifications with functional annotations and rank abundances. The natural organic matter workflow processes direct infusion Fourier Transform mass spectrometry (DI FT-MS) data to assign molecular formulas to observed peaks. Formula assignments are determined from a defined molecular search space and based on the mass accuracy and fine isotopic structure of the data.

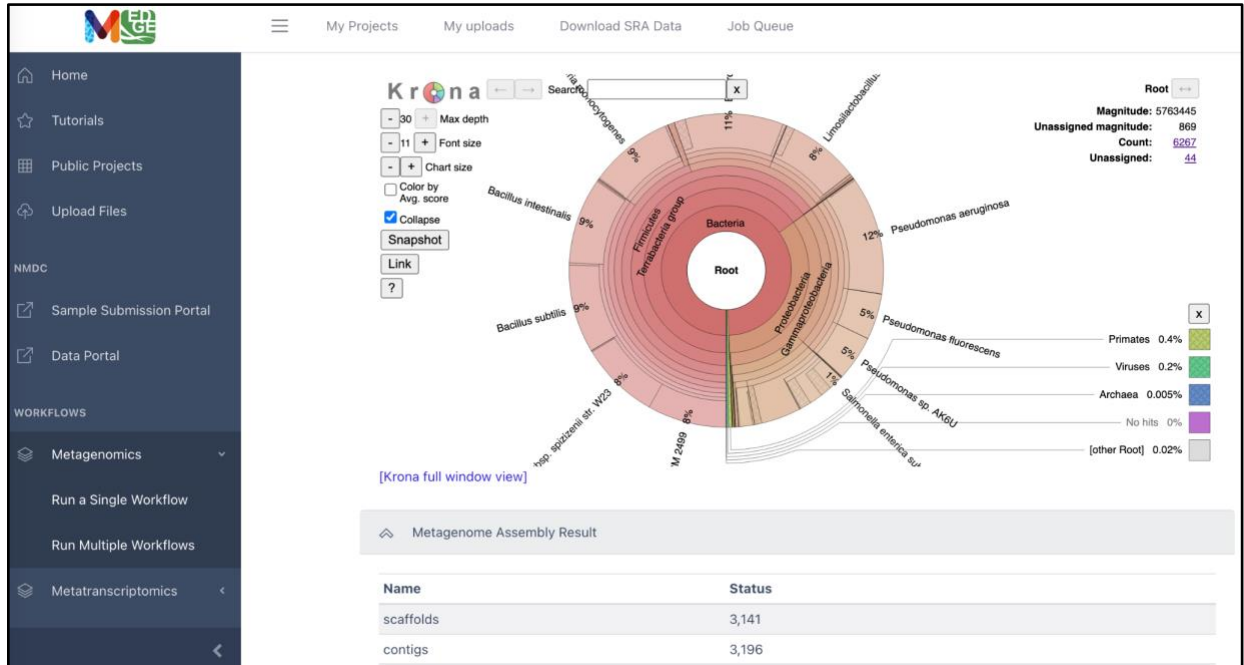


Figure 3. NMDC EDGE with available multi-omics workflows. The results for a metagenome workflow reads-based taxonomy classification analysis with visualization of the taxonomic composition using a Krona plot.

The [workflow documentation](#) and [tutorials](#) are continually updated to assist both novice and experienced users in their understanding of the various components of the data processing steps. Aligned with our commitment to increasing the accessibility of the NMDC products, our team has updated and translated each workflow [user guide](#) into Spanish and we anticipate additional language translations this year.



The NMDC Data Portal & Public API: simple access to multi-omics data

The Data Portal democratizes access to studies integrating different data types, supporting persistent identifiers and links to external repositories (**Figure 4**). Launched in March 2021, the infrastructure relies on a distributed framework and sets the foundation for future development to **address an immediate and pressing need for access to integrated multi-omics studies**.

The Data Portal design enables the research community to discover multi-omics data through a variety of search functionalities, such as faceted search and interactive visualizations using both the environmental sample information and functional annotations through KEGG Orthology, module, and pathway terms. Users may also search through two systems for environmental ecosystem classifications, including the [GOLD](#) ecosystem classification paths [8] and the Environment Ontology ([EnvO](#)) classification terms [9].

The Data Portal currently contains nearly 117,000 metagenomic, metatranscriptomic, metaproteomic, metabolomic and natural organic matter data files, generated by the

NMDC standard workflows, available for download. These data are associated with 7,786 biosamples from a broad range of environmental ecosystems, spanning river water and sediments, plant-microbe associations, and a range of diverse soils, among others. New studies, biosamples, and standardized data products are continually added to the Data Portal, in partnership with the JGI, EMSL, NEON and the microbiome research community.

For programmatic access, we have developed a web application programming interface (API). The public [NMDC API](#) enables query and access for all study and biosample information. This allows NMDC metadata and data to be directly pulled into programmatic scripts or analysis tools. The NMDC API can be used broadly by the research community and has been adopted this past year by JGI's IMG/M platform [10] to access and share data. These major accomplishments form the basis of the NMDC's overall software architecture for exchanging and linking data across resources.

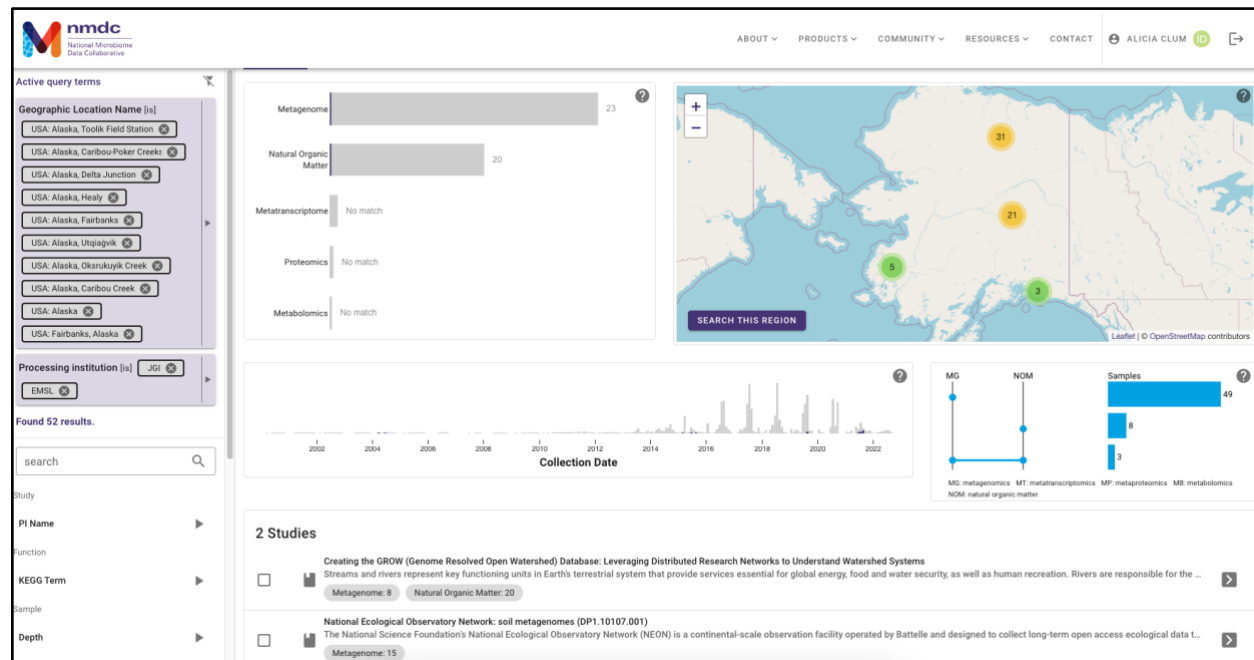


Figure 4. Data Portal search results for samples from Alaska where data was generated at either the JGI or EMSL DOE User Facilities.

Challenges & Opportunities

Many studies now collect multiple omics data types (e.g., metagenomics, metatranscriptomics, metaproteomics, and metabolomics) to better understand microbial composition, function, and impacts on ecosystem processes. However, integration of multiple microbiome data types are *ad hoc* and meta-analyses across studies are rare. The NMDC offers tools to make it easy to adhere to data standards and ways to access integrated multi-omics data as a foundation for researchers to discover community dynamics, pathways, metabolic currencies, and other relationships. Over the past four years, we have met several challenges that have been instructive.

First, we have found it critical to assess the benefit and cost of support for legacy data. Petabytes of microbiome data already exist across diverse data repositories. One of the primary challenges we faced during the initial development stages was determining the *value of existing data* for integration into the NMDC infrastructure. To adhere to community standards, many of the studies and available data required manual curation to ensure minimum metadata requirements were met. This required contacting the primary research team to gather missing metadata, along with coordinating with User Facilities and public data repositories. The human resources required to fill in these gaps was an unanticipated effort, and led to a focus on multi-omics studies derived from large consortia or research groups with the resources and staffing to assist our efforts to compile accurate metadata about samples and experiments that were completed years ago. This model has informed our strategy to focus on data management best practices at the beginning of the data lifecycle, and importantly resourcing for the Submission Portal and a planned metadata mobile app, NMDC Field Notes, to ensure all necessary data are collected and managed to avoid back-filling legacy data.

We have also found that creating distributed infrastructure for data sharing across computational resources presents unique challenges. A long history of investments, both by DOE and other federal agencies, has created a mosaic of data infrastructures and computational resources that are not connected and cannot easily share data. While this existing ecosystem has supported the diverse needs of the research community, it has also created significant siloes across research areas and data types. An early challenge we faced was developing distributed infrastructure to share the primary data generated by and archived in existing User Facility systems. This required in-depth knowledge of data and compute requirements and dependencies across JGI, EMSL, ESS-DIVE, and KBase, along with NEON for external DOE connections. Data integration across heterogeneous DOE resources will continue to be an ongoing challenge. We look forward to participating in the broader DOE-wide discussions on enabling pathways towards an integrated data ecosystem and believe that our approach to structured metadata submission and API-driven access can help in this space.

Despite these challenges, the strategy we have adopted continues to be refined and developed in close collaboration with the research community. What we have accomplished over the past four years is seeding the future of open, collaborative microbiome research. Looking to the year ahead, we are well underway in our development of the mobile app, NMDC Field Notes, for field researchers to record sample information onsite. By connecting mobile devices to sensors, researchers would be able to efficiently collect real-time measurements compared to writing in notebooks and manually organizing data. We are also developing machine learning tools to translate free text into standardized terms. These innovations will reduce the cost of data management and curation, allowing researchers to focus instead on using microbiome data to tackle important issues like climate change and ecosystem health. We look forward to supporting the community in harnessing data to understand the tiny microbial engines that drive life on our planet.

References

1. Interagency strategic plan for microbiome research, FY 2018-2022. US DOE Office of Science (SC) (United States); 2018 Apr. doi:10.2172/1471707.
2. Elie-Fadrosh EA, Ahmed F, Anubhav A, Babinski M, Baumes J, Borkum M, et al. The National Microbiome Data Collaborative Data Portal: an integrated multi-omics microbiome data resource. *Nucleic Acids Res.* 2021;50: D828–D836.
3. Vangay P, Burgin J, Johnston A, Beck KL, Berrios DC, Blumberg K, et al. Microbiome Metadata Standards: Report of the National Microbiome Data Collaborative’s Workshop and Follow-On Activities. *mSystems.* 2021;6. doi:10.1128/mSystems.01194-20.
4. Hu B, Canon S, Elie-Fadrosh EA, Anubhav, Babinski M, Corilo Y, et al. Challenges in Bioinformatics Workflows for Processing Microbiome Omics Data at Scale. *Front Bioinform.* 2022;1. doi:10.3389/fbinf.2021.826370.
5. 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 (MlxS) specifications. *Nat Biotechnol.* 2011;29: 415–420.
6. Bandrowski A, Brinkman R, Brochhausen M, Brush MH, Bug B, Chibucos MC, et al. The Ontology for Biomedical Investigations. *PLoS One.* 2016;11: e0154556.
7. Camargo AP, Roux S, Schulz F, Babinski M, Xu Y, Hu B, et al. Identification of mobile genetic elements with geNomad. *Nat Biotechnol.* 2023. doi:10.1038/s41587-023-01953-y.
8. Mukherjee S, Stamatidis D, Li CT, Ovchinnikova G, Bertsch J, Sundaramurthi JC, et al. Twenty-five years of Genomes OnLine Database (GOLD): data updates and new features in v.9. *Nucleic Acids Res.* 2023;51: D957–D963.
9. 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.
10. Chen I-MA, Chu K, Palaniappan K, Ratner A, Huang J, Huntemann M, et al. The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Res.* 2023;51: D723–D732.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov