nmdc
National Microbiome
Data Collaborative

National Microbiome Data Collaborative

Progress Report 2023
U.S. Department of Energy

TABLE OF CONTENTS

Cover credit: Photo courtesy of NEON Science
Photos taken by Edward Pablo and Andrea Starr,
Pacific Northwest National Laboratory.

Vision

To connect data, people, and ideas to advance microbiome innovation and discovery.

Mission

To support a findable, accessible, interoperable, and reusable (FAIR) microbiome data-sharing network, through infrastructure, data standards, and community building, that addresses pressing challenges in environmental science.

Advancing microbiome science for the benefit of all

Emiley Eloe-Fadrosh
National Microbiome Data Collaborative Lead

In January of this year, the White House Office of Science and Technology Policy launched the Year of Open Science with new actions to advance open and equitable science policies across the federal government. The National Microbiome Data Collaborative (NMDC) is committed to open and equitable research in microbiome science. This commitment is the foundation for our infrastructure development activities and has been a prominent driver this past year across our three products: the **Submission Portal**, **NMDC EDGE**, and the **Data Portal**. It has been an exciting and busy time for our team as we work to serve the scientific community in a way that enables microbiome innovation and discovery. I am exceptionally proud of the progress the NMDC team has made this past year in fostering strong community partnerships and advancing our powerful products into tools that drive scientific impact.

We launched the NMDC persistent identifier service in January 2023 and deployed programmatic access to all NMDC data through a public application programming interface (**API**). Together, these efforts support a larger findable, accessible, interoperable, and reusable (FAIR) data ecosystem to programmatically exchange and link data across resources. The Data Portal now hosts over 7,700 biosamples with a collective nearly 90 TB of multi-omics microbiome data and links across complementary data platforms, including the Integrated Microbial Genomes and Microbiomes (IMG/M) and Genomes OnLine Database (GOLD) of the Joint Genome Institute (JGI), the Department of Energy (DOE) Systems Biology Knowledgebase (KBase), the National Center for Biotechnology Information (NCBI), the Mass Spectrometry Interactive Virtual Environment (MassIVE), the Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE), and the National Science Foundation's (NSF) National Ecological Observatory Network (NEON) Data Portal. This trove of standardized multi-omics data represents a large and unique microbiome collection that spans geographically diverse samples available for cross-study comparisons. We have also provided access to the NMDC standardized bioinformatics workflows through NMDC EDGE, along with the newly available virus and plasmid identification tool, geNomad, that was collaboratively published this summer in ***Nature Biotechnology***.

The **2023 NMDC Ambassadors** made tremendous strides this past year to embrace and broadly share data stewardship best practices. This outstanding cohort hosted over 20 events reaching 550 researchers in three languages. We have also sustained our commitment to our previous cohort through a published synthesis of the Ambassador Program in ***Nature Microbiology***. This year, we have fully embedded our scientific engagement and user research throughout the NMDC development lifecycle and called on our active **Champions**. In June, we launched a web page with information about our **user research**, which includes how we recognize this service through ORCiD and related publications, and any open calls for contributions.

As 2023 ends and we look to the year ahead, I am energized to boldly push forward on the collaborative path that the NMDC team has forged for the research community. Through broadening and strengthening our partnerships across research teams, organizations, federal agencies, and the international sphere, our team is poised to advance open and equitable microbiome research for all. We invite you to read on to learn about our team's accomplishments this past year and to help us shape the future of microbiome innovation.

Emiley Eloe-Fadrosh
National Microbiome
Data Collaborative Lead

Delineating Soil Sampling Plots in high winds.
Photo courtesy of NEON Science

A community-driven data infrastructure

The NMDC is tackling existing gaps in microbiome research by using proven approaches and new innovations in distributed data infrastructure and linked data technologies. Our three products — the **Submission Portal**, **NMDC EDGE**, and the **Data Portal** — are driven by community needs. They support data, information, knowledge sharing, and access. This past year, we worked closely with the research community to strengthen our existing infrastructure in ways that will catalyze new research. This included launching a new persistent identifier service, supporting programmatic access to NMDC data, expanding the amount of available high-quality data and workflows, and contributing to major updates of community data standards.

Sevilleta Long Term Ecological Research (LTER) Site in New Mexico. Credit: Buck Hanson

Building a robust software and data ecosystem

The NMDC schema underlies all aspects of how data are handled and made FAIR. Using the Linked Data Modeling Language (LinkML), our team can weave together several different community standards, such as the Minimal Information about any (x) Sequence (MIxS) standard from the Genomic Standards Consortium (GSC) and the Ontology for Biomedical Investigations (OBI) framework for modeling sample processing and data generation. This past year, our team has significantly updated the NMDC schema along with coordinated data migrations to streamline data management and processing. The new NMDC schema releases provide detailed information on schema improvements, updates, and contributor information. The evolution of the NMDC schema follows a robust, yet flexible, approach to support large studies and long-term data collection efforts driven in close partnership with NEON. The schema now supports new classes for sample processing, including pooling, extraction, and library preparation. Further, our team has worked toward improved support for database conversion to Resource Description Framework triple store with corresponding SPARQL queries and has overhauled the definition and usage of CURIE prefixes.

In January, we launched the NMDC persistent identifier service consistent with the NMDC schema to support links across studies, samples, and workflow runs. We also deployed programmatic access to all NMDC study and biosample metadata through the public NMDC API. The NMDC API can be used broadly by the research community and has been adopted this past year by the IMG/M platform 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.

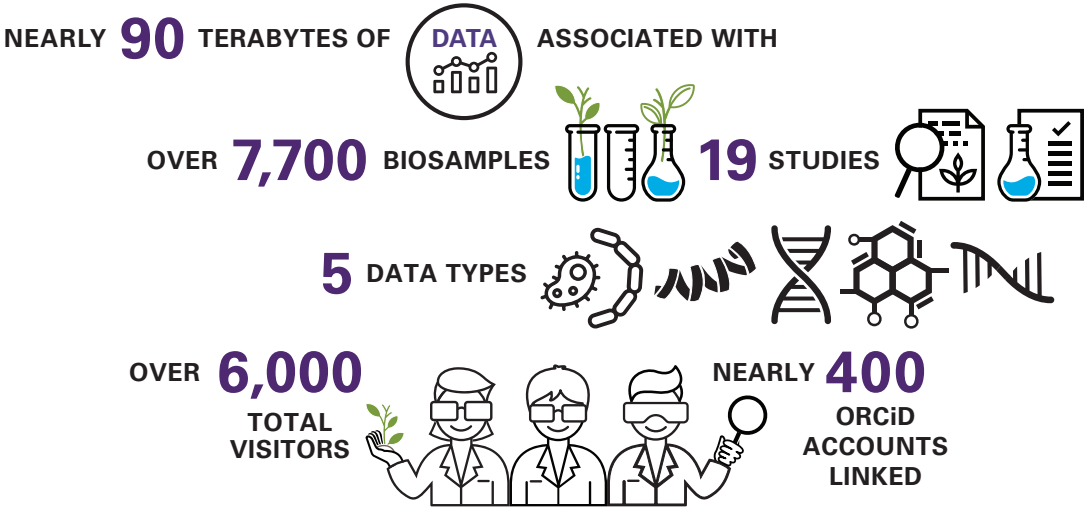
ACCOMPLISHMENTS

- **Made significant updates to the NMDC schema, which are captured in the new NMDC schema releases with detailed information on improvements and contributor information**
- **Deployed programmatic access to NMDC metadata through the NMDC API**
- **Transitioned the NMDC workflow automation to use *Perlmutter* at the National Energy Research Scientific Computing Center (NERSC)**
- **Created a backup instance of the Data Portal and Submission Portal running on the Kubernetes resources at the Environmental Molecular Sciences Laboratory (EMSL)**
- **Enabled Globus access to NMDC data for automated, bulk, and high-speed transfers**

With the retirement of the NERSC supercomputer Cori in May, the NMDC workflow automation has been updated to use *Perlmutter*. Our team has enhanced the automation to manage version updates of the workflows, along with improvements toward importing processed data from the JGI to avoid duplication. We have further streamlined our database infrastructure and ingest processes to significantly reduce the time to import new data from over a day to under an hour. These improvements have enabled us to add more than 70 TB of data to the Data Portal this year alone.

To improve our overall resilience, we have instituted automated backups of our databases and hosted an archival copy of the NMDC products. We have also implemented several improvements to increase the robustness of our system deployments. Our use of containerized microservices (with Docker and Kubernetes) has allowed us to easily deploy multiple environments for production and development at NERSC, along with a backup instance of the Data Portal and Submission Portal running on EMSL's Kubernetes resources. Further, to allow for high bandwidth access to NMDC data, we are introducing Globus access through NERSC. This provides users with a mechanism for automated and bulk downloads of data that can take advantage of high-speed networking capabilities.

BY THE NUMBERS



Lowering barriers to FAIR data collection with the Submission Portal

Our team is constantly seeking new ways to lower barriers to data collection and to improve the adoption of standards. In April 2022, we released a beta version of the NMDC **Submission Portal** that supports the collection of study and biosample information. Through improved technical specifications of the GSC's MIxS standard, the Submission Portal allows users to adhere to standards and enables the data to be machine-readable and interoperable. This past year, our team improved the usability of the Submission Portal and created a comprehensive **user guide** and **tutorial**. These materials guide researchers on how to use the MIxS environmental extensions and the real-time validation functions.

To support data submission across the DOE user facilities, the JGI and 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 also supports metadata harmonization across the JGI and EMSL.

This past summer, we conducted five user interviews to gather feedback on the Submission Portal's design, and compiled action items to make improvements to support both DOE user facility researchers and the broader microbiome research community.

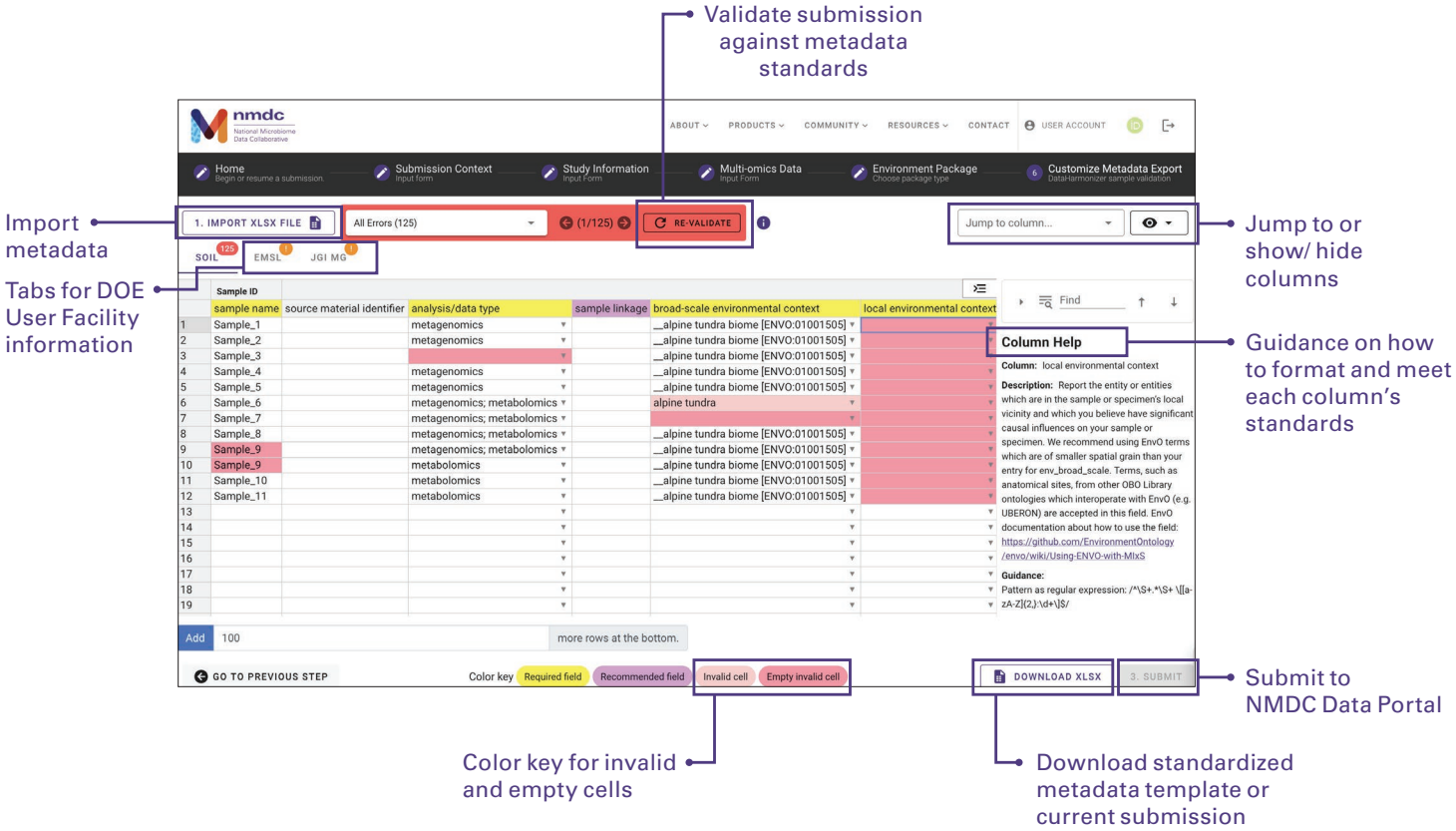
The new "**sandbox**" site for demonstration purposes is now available, hosting temporary data for a week to test and explore functionality. Interested users can test our tools and get familiar with the functions and required metadata before they are ready to submit a study or sample information for integration into the Data Portal.

ACCOMPLISHMENTS

- **Launched a new "sandbox" site for testing purposes and workshop demonstrations**
- **Implemented a new design to separate sample and assay metadata along with DOE user facility information for ease of use**
- **Developed a comprehensive user guide and tutorial to showcase features and support usage by the research community**
- **Implemented submission status with quality control checks to support user submissions**
- **Performed usability testing that yielded 72 insights, resulting in 20 action items to improve the user interface and the accessibility of portal features**

The NMDC Submission Portal

Making it easy to follow standards



Standardized bioinformatics workflows accessible through NMDC EDGE

Our team continues to support production-quality open-source bioinformatics workflows to process multi-omics data and produce interoperable and reusable data products. This year, we released two new workflows, the NMDC standardized referenced-based metaproteomics workflow and the viruses and plasmids workflow (geNomad). 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 viruses and plasmids workflow leverages the newly released **geNomad tool** along with **CheckV** to detect plasmids and viral contigs from genomic, metagenomic, transcriptomic, or metatranscriptomic sequencing data. In September 2023, several team members collaboratively published a description of geNomad and its availability through NMDC EDGE in ***Nature Biotechnology***.

The **workflow documentation** and **tutorials** have been 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 multiple languages.

For 2023, we expanded the **Ambassador Program** to encompass training activities on standardized

ACCOMPLISHMENTS

- **Collaboratively published a description of the geNomad workflow for detecting plasmids and viruses and its availability in NMDC EDGE in *Nature Biotechnology***
- **Integrated the standardized metaproteomics workflow into NMDC EDGE**
- **Developed updated user guides, tutorial videos, and technical documentation**
- **Translated user guides into multiple languages to increase accessibility**
- **Performed beta testing of the NMDC EDGE workflows, which resulted in 63 insights and 40 actions**

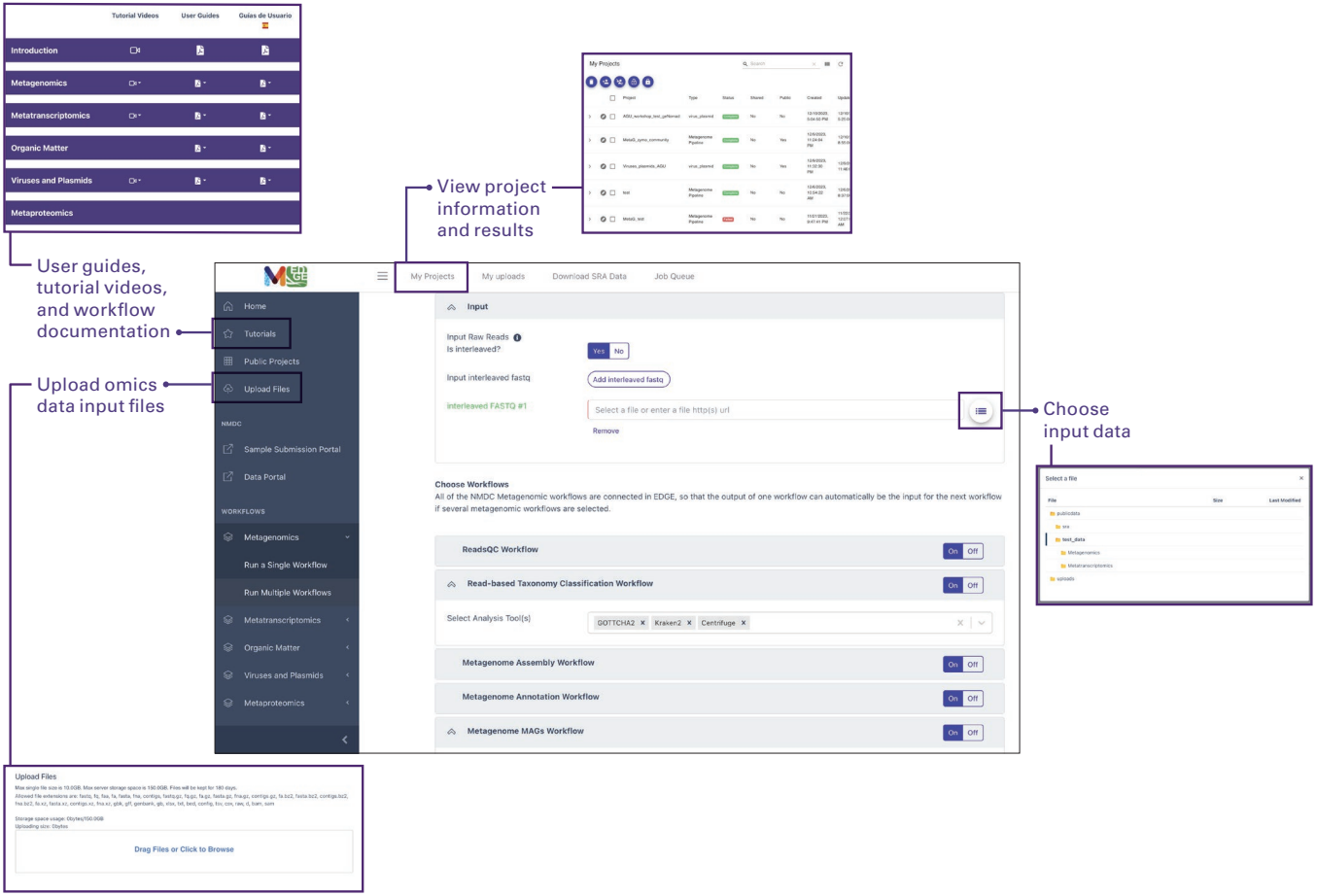
bioinformatics workflows with NMDC EDGE. All Ambassadors featured this product in their workshops, and NMDC EDGE was also prominently featured in several NMDC-led events.

The Ambassadors and **Champions** continue to provide critical feedback regarding the workflows, user interface, output visualizations, and training materials. They have helped us conduct extensive beta testing and user research on NMDC EDGE and the newly released workflows, leading to 63 insights and 40 action items. Beta testing feedback included requests for additional outputs, such as publication-ready figures, updates to the technical documentation, and further guidance in the workflow submission pages.

Ambassador Reid Longley and NMDC team member Julia Kelliher present during a workshop at the Sequencing to Function: Analysis and Application for the Future conference (SFA²F). Credit: Leah Johnson

NMDC EDGE

Providing user-friendly bioinformatics tools



Allowing new access to multi-omics data with the Data Portal

Since the launch of the **NMDC Data Portal** in 2021, we have continued to add studies, biosamples, and standardized data products to it from our multi-omics bioinformatics workflows. A key feature of the Data Portal is enabling 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 and the Environment Ontology (**EnvO**) classification terms.

The Data Portal currently contains nearly 117,000 data files available for download, which are associated with 7,786 biosamples from a broad range of environmental microbiomes, spanning river water and sediments, plant-microbe associations, and a range of diverse soils, among others.

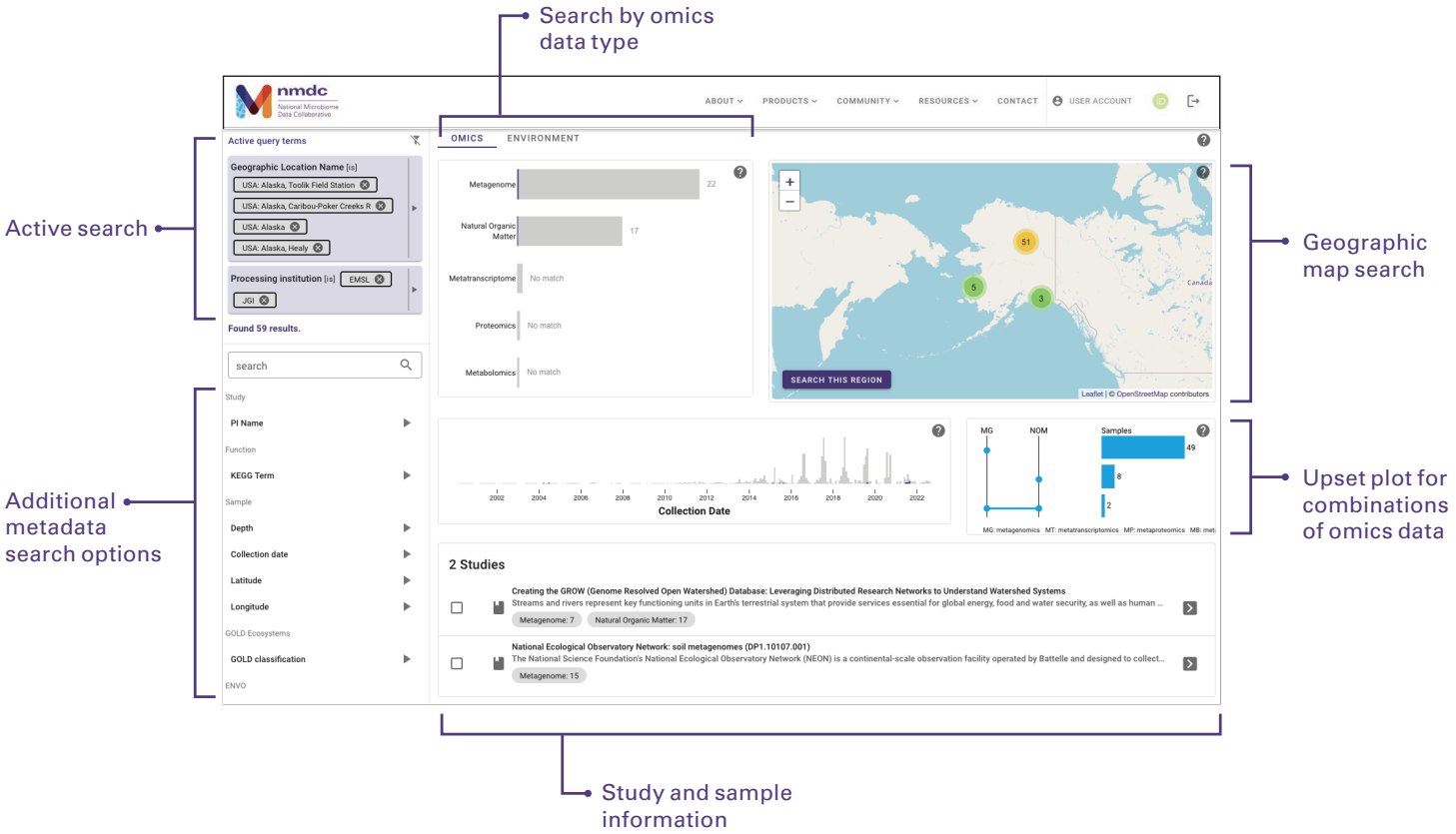
In partnership with several research teams and campaigns, we have added thousands of unique soil, sediment, and aquatic samples this past year, representing a large and diverse geographic distribution. These data derive from the Genome Resolved Open Watershed (GROW) project, the 1,000 Soils Research Campaign, and microbiome samples collected from NEON's 81 field sites, including 47 terrestrial and 34 aquatic sites throughout the ecoclimatic domains of the United States.

ACCOMPLISHMENTS

- **Updated the NMDC Data Portal to host a six-fold increase in data, with over 7,700 biosamples from 19 studies**
- **Made available thousands of new workflow results for download, primarily for metagenomics and natural organic matter analysis**
- **Used our API minting service to create persistent identifiers for records, including studies and outputs generated by our workflows**
- **Ensured NMDC study pages support an increased number of study identifiers with links to external resources including MassIVE, ESS-DIVE, KBase, and project-relevant websites and publications**

NMDC Data Portal

Simple access to multi-omics data



Promoting an inclusive and connected scientific community

The NMDC's role within the exponentially growing field of microbiome research is to serve the community in a way that enables innovation and discovery. To achieve our vision of connecting data, people, and ideas, it is imperative that the scientific community be deeply involved in the creation and execution of all aspects of the NMDC. We engage partners across research teams, organizations, federal agencies, scientific societies, and the international sphere. Over this past year, we have expanded our Ambassador and Champions programs, strengthened our user research activities, launched new ways to engage through social media, and advanced large-scale data initiatives within and beyond the DOE.

Oksrukuyik Creek (OKSR), AK. Credit: NEON Science

Toward an inclusive, diverse, and equitable future for microbiome science

In 2023, our team attended attended over 40 events, gave over 30 presentations, reaching over 2,100 researchers. At the American Society for Microbiology (ASM) Microbe conference, the team hosted a town hall session showcasing the work of three NMDC Ambassadors and the 2023 ASM Microbiome Data Prize awardee. The team hosted a panel discussion that was included in the ASM's new featured Climate Change and Microbes scientific track. The NMDC team presented and cohosted a workshop with an Ambassador at the Sequencing to Function: Analysis and Application for the Future (SFA^2F) conference, and presented an overview of the NMDC to over 250 conference attendees. We presented at the National Summer Undergraduate Research Project (NSURP) and hosted a panel discussion at the National Diversity in STEM conference. Several team members presented at the GSC annual meeting, and team members also hosted a hands-on workshop centered on the Submission Portal and how the NMDC incorporates the GSC standards into this product. The NMDC team also cohosted a workshop and session with our NEON collaborators at the Ecological Society of America (ESA) conference. Through our participation at

ACCOMPLISHMENTS

- **Collaboratively published an article in *Nature Microbiology* with past Ambassadors highlighting the program and future activities**
- **Gathered user feedback from Champions, Ambassadors, and community members and created a user research web page to inform the community of opportunities for involvement**
- **Ambassadors hosted 21 events throughout 2023 in the United States and globally, reaching 550 researchers**
- **Created a 2024 inclusion, diversity, equity, and accountability (IDEA) Action Plan and deployed it onto the website to increase accountability**
- **Launched the NMDC LinkedIn and Instagram accounts to broaden reach**

conferences, we meet researchers eager to learn about the best ways to collect and share their findings with others. Our activities become a setting for conversations about new trends and best practices in microbiome research.

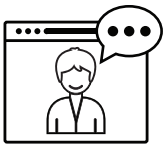
Our team supports an inclusive culture that is aware of the diversity of experiences, expertise, backgrounds, needs, and perspectives of the microbiome research community. Our initial **IDEA strategic plan**, which included 27 actions, helped us achieve our goals and move toward creating an even more equitable community. Our team completed 22 actions, with 5 currently in progress. Some highlights include the launch of the NMDC IDEA working group and the translation of product user guides into multiple languages. We also outlined and implemented relevant metrics for tracking NMDC progress toward our IDEA goals. The team also created a **2024 IDEA Action Plan**.

This year, we launched the NMDC **LinkedIn** and **Instagram** accounts. We have continued to grow our social media presence and expand to new audiences. We provided updates to our community through our quarterly newsletter, ***The Microbiome Standard***, and published several **blog posts** highlighting specific accomplishments and collaborations. We also launched **NMDC Snapshots** introducing our 2023 Ambassador cohort to our wider community.

BY THE NUMBERS

13 AMBASSADORS

53 CHAMPIONS



21 AMBASSADOR EVENTS

30+ NMDC TEAM Presentations

2,650+ PEOPLE REACHED at events

27 USER RESEARCH Participants

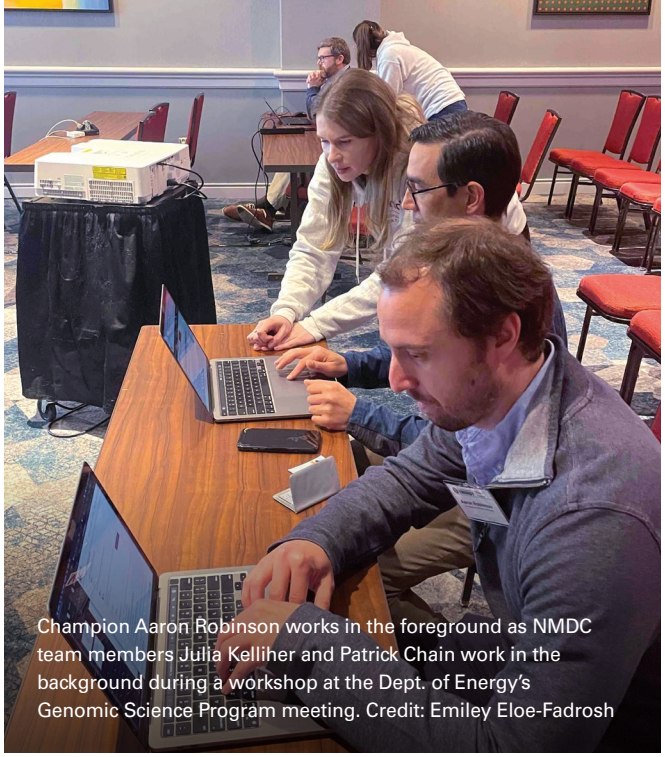
185 INSIGHTS

78 ACTION ITEMS in total

User research underpins infrastructure innovation

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. This year, we created the **NMDC user research resource page** to communicate current and upcoming areas of user research to the community and provide an avenue for the research community to sign up to be involved. User research participants can opt out of acknowledgments but are credited for their contributions through an ORCiD service designation and acknowledgments in product-related publications, as well as on our web pages.

This year, we conducted five user interviews and performed beta testing to gain deeper insights into improving the Submission Portal and NMDC EDGE, respectively. For the Submission Portal usability testing, our team gathered 72 insights that resulted in 20 action items. These insights included improved tutorials and help guidance, more direct methods to contact the NMDC team from the Submission Portal, and updates to the user interface to increase feature accessibility, such as more intuitive methods to download a blank metadata template.



Champion Aaron Robinson works in the foreground as NMDC team members Julia Kelliher and Patrick Chain work in the background during a workshop at the Dept. of Energy's Genomic Science Program meeting. Credit: Emiley Eloe-Fadrosh

Beta testing of the NMDC EDGE workflows focused on the reference-based metaproteomics workflow and the viruses and plasmids workflow (geNomad). The 63 insights generated from the beta testing form responses led to 40 action items to update the usability of the web platform and the utility of the workflows. Our users requested additional post-processing outputs, such as publication-ready figures and data summaries, updates to the technical documentation to clarify workflow parameters and inputs, and further help guidance within the workflow submission pages.

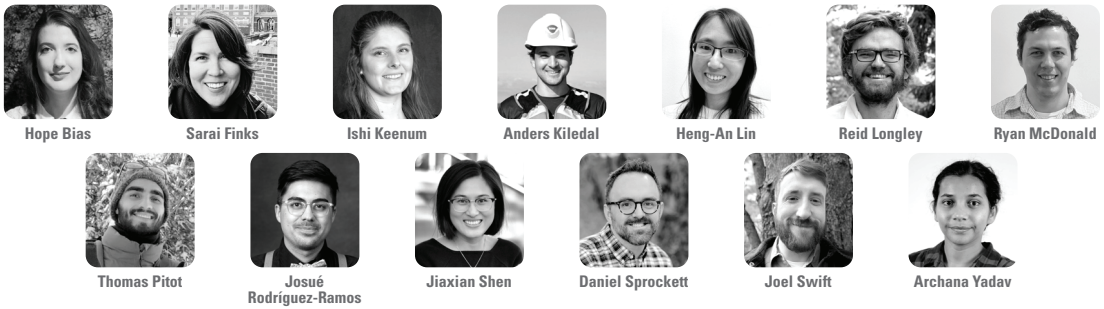
NMDC Ambassadors

The **Ambassador Program** uses a community learning approach to train and support early career researchers motivated to engage with their respective microbiome research communities. During their year-long term, Ambassadors undergo extensive training in microbiome data stewardship best practices. Afterward, they host events within their community that spread awareness and knowledge regarding data stewardship, metadata standards, and bioinformatics workflows. These events also highlight the NMDC products and how researchers can use these products.

Last year, we asked our first cohort of Ambassadors for suggestions on ways to enhance the program. This year, based on their feedback, we expanded our program's curriculum from a focus on metadata standards to data management best practices more broadly. Our 2023 cohort of 13 NMDC Ambassadors started their term by selecting one of three pathways to focus on: (1) microbiome metadata standards, MIxS templates, and the **NMDC Submission Portal**; (2) multi-omic data processing, standardized bioinformatics workflows, and **NMDC EDGE**; or (3) microbiome data stewardship, data management, and the **NMDC Data Portal**.

Throughout 2023, the Ambassadors organized and hosted 21 events, in nine states and three countries, which included hands-on workshops and presentations. They reached a combined audience of more than 550 researchers through their engagement activities. Two NMDC Ambassadors hosted events for Spanish and French-speaking audiences. The Ambassadors provided feedback from their own experiences and those of their event attendees. Event attendees were asked to fill out a post-event survey, and the results and feedback will be used to improve the Ambassador Program and NMDC products. This information will also be captured in a manuscript written by the NMDC team and the 2023 Ambassadors.

MEET OUR AMBASSADORS



NMDC Champions

Our NMDC **Champions** give us important insights into how our products can help the larger microbiome research community. Throughout the year, we work with them to create training materials, refine our products through user research, and brainstorm ways to connect with new users. The Champions Program now includes over 50 microbiome researchers from around the globe, who are working in many different institutional settings. Champions attend quarterly meetings, where they learn about recent infrastructure developments and opportunities for engagement and networking. This year, several Champions cohosted events with the NMDC team, including the panel focused on data management at the National Diversity in STEM conference. They are also working toward collaborative manuscripts with our team members and are involved in the NMDC-led IDEA working group.

MEET OUR CHAMPIONS

Science, partnerships, and impact

Cultivating strategic partnerships and building community has always been at the core of the NMDC. Through coordination of cross-cutting efforts and working with diverse stakeholders, our team is helping to shape the next generation of microbiome data stewardship. This year, our team has been deeply embedded in conversations to advance a unified data infrastructure for the DOE Office of Science Biological and Environmental Research (BER) program. Our work with the GSC, ASM, and the Microbiome Centers Consortium continues to thrive and further builds toward a collaborative ecosystem. We have also worked closely to demonstrate value and scientific impact through unique data collections with an eye toward scaling microbiome data to address pressing environmental challenges.

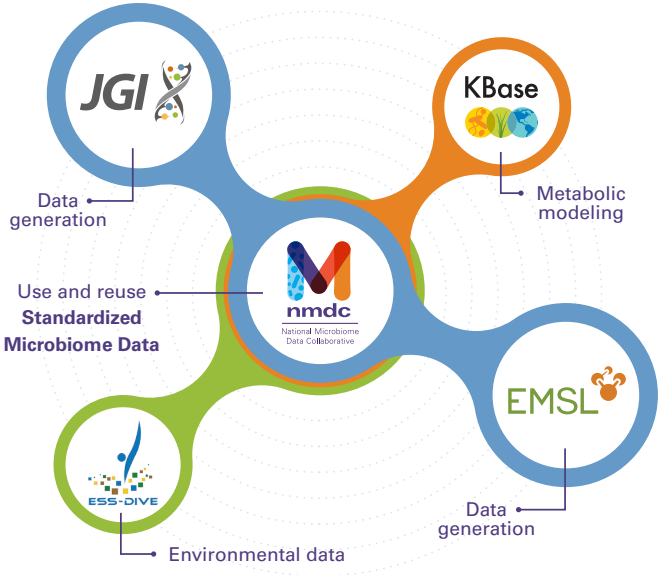
Rocky Mountain National Park, CASTNET (RMNP), CO. Credit: NEON Science

Bolstering the DOE data ecosystem

Our team closely collaborates with colleagues across the DOE's Office of Science BER program to support an ever-growing portfolio of microbiome research. Our team was heavily involved in the BER Advisory Committee unified data infrastructure activities this past year, including the *Towards a unified data framework for DOE BER* workshop series in July 2023. We look forward to a forthcoming workshop report focused on developing a new strategy for the next generation of data management and analysis within a unified framework.

We have a longstanding and highly productive collaboration across the JGI's GOLD and IMG/M platform teams. 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 more seamlessly share processed metagenome data 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.

We have also established a productive collaboration with EMSL's Computing and Data Operations group. During the past year, we 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 data. We can ensure reliable access to microbiome data and resilience for data processing demands by leveraging these BER infrastructure resources.

In addition to our collaborations with EMSL and the JGI, we coordinate with KBase and ESS-DIVE to share data and support joint outreach activities. This year, we have advanced discussions with the ESS-DIVE team to develop methods for linking samples and metadata across data infrastructures. Together with the ESS-DIVE team, we detailed these collaborations in a **blog post** and have further engaged in these activities with the broader environmental research community through the **ESS-DIVE Open Data Workshop**.

Driving technical solutions for community standards

Our team has a longstanding and productive partnership with the GSC to support several facets of the community standards. Since the launch of the NMDC, the GSC's President, Lynn Schriml, has served on the NMDC **Scientific Advisory Board**. Several members of the NMDC team are heavily involved in both the GSC's Compliance and Interoperability Working Group and Technical Working Group. This past year, we worked closely with the GSC to advance the use of LinkML to manage the MIxS standard for new releases and introduced new ways for user access (**https://github.com/GenomicsStandardsConsortium/mixs**). Our team was also instrumental in releasing the new MIxS **documentation website**, which allows for browsing of the metadata terms via several different entry points, for example, a checklist of interest or an extension of interest. Our close partnership with the GSC continues to advance computable and automated validation for this critical community standard.

The NMDC has also partnered with the GSC on several outreach activities to communicate the value of microbiome data standards. Our team provides extensive MIxS and data standards training to the annual cohort of NMDC Ambassadors, who are then tasked with hosting their own workshops and events to distribute this information and provide hands-on experiences with MIxS for microbiome data. This year, we have made these **training materials** widely available. The GSC awarded NMDC Submission Portal product owner Montana Smith the Dawn Field Award this year. NMDC team members also presented at the GSC's annual meeting and ran a workshop.

NMDC team members from left to right: Julia Kelliher, Mark Miller, Montana Smith, and Yuri Corilo traveled to the Genomic Standards Consortium meeting in Bangkok, Thailand. Credit: GSC23 Conference

Tackling continental-scale biology through partnerships

This year, the National Academies held a series of listening sessions for the public information gathering event **Paving the Way for Continental Scale Biology: Connecting Research Across Scales**. It was an honor for our team to present on the many ways the NMDC is helping microbiome researchers translate knowledge across scales. This work is exemplified in our ongoing partnership with the NSF's NEON. Together with the NEON team, we have worked to coordinate metadata exchange for over 4,000 soil, sediment, and aquatic samples to host in the NMDC Data Portal and process metagenome and natural organic matter data. In August, NMDC and NEON team members cohosted a workshop at the ESA conference to showcase access to this treasure trove of data. We are also excited to partner with the NEON team through a newly supported **Community Science Program** project with the JGI to sequence over 1,000 metagenomes. These new data will be integrated with a decade of existing microbiome data to further bolster continental-scale analyses. We have already incorporated the data sequenced at the JGI to pilot this partnership.

In a similar thread, we have partnered with the EMSL team on the newly launched Molecular Observation Network (**MONet**). This is a new open science network designed to produce regional-scale molecular and microstructural information about soils and their microbial communities. Launched in March 2023 and accepting user proposals on a quarterly basis, MONet will generate new data to advance understanding of diverse soils with a focus on training and networking across research communities. Our team will support accessibility of MONet metagenome data generated at the JGI and natural organic matter data from EMSL through the NMDC Data Portal. We have already demonstrated integration of sample metadata through support for MONet's pilot the **1,000 Soils Research Campaign**. Our team will also partner with the MONet team for future outreach and training activities.

1,000 Soils Research Campaign. Credit: Edward Pablo.

A spotlight on scientific impacts

The NMDC aims to serve as a catalyst for innovative research leveraging multi-omics data across diverse environmental microbiomes. This past year, we advanced our collaborations with research teams doing cutting-edge research and bioinformaticians building up computational capabilities for cross-study multi-omics comparative analyses. The **Bio-Scales project**, led by Mitchel J. Doktycz at Oak Ridge National Laboratory, is a model for integrated and FAIR microbiome science. This project is driven by the hypothesis that combinations of host and microbial traits influence nitrogen transformation patterns and fluxes across the coupled plant-soil-microbial system. In collaboration with the JGI and NMDC teams, the Bio-Scales project has completed data generation for soil, rhizosphere, root endosphere, and leaf samples derived from 27 different *Populus trichocarpa* genotypes grown in two different environments. This tremendous integrated dataset spans 318 metagenomes, 98 plant transcriptomes, and 314 metabolomic profiles that are correlated with diverse soil measurements. Data have been made available through the **Bio-Scales** study on the NMDC Data Portal, and a forthcoming data report will be published in early 2024.

As with the Bio-Scales project, the NMDC team has similarly collaborated this past year with the GROW consortium. Leveraging previous successful models of community crowd-sourced microbiome science, the GROW consortium aims to create a catalog of diverse river microbiomes. To fill data gaps for river microbiomes, 250 rivers, including 35 of the largest rivers in the United States, were sampled over time and across interconnected hydrological units. The GROW consortium, in collaboration with the JGI, KBase, and NMDC teams, released a **preprint** earlier this year. The work describes initial insights into the nearly 1,500 metagenome-assembled genomes recovered from this unique sampling effort. The NMDC Data Portal hosts hundreds of metagenome and natural organic matter datasets from **GROW**, with links to KBase and ESS-DIVE for complementary data.

East River Samples. Credit: Patrick Sorensen

For more information, contact:
Emiley Eloe-Fadrosh,
National Microbiome Data Collaborative Lead
eaeloefadrosh@lbl.gov

Design by Creative Services, Berkeley Lab

In October 2023, NMDC staff attended a two-day retreat in Monterey, California. The team members from PNNL, LANL, and LBNL brainstormed implementation strategies and planned milestones. Credit: NMDC; A series of photos on top from the National Microbiome Data Collaborative. Starting from left: Montana Smith in Prosser, Washington, Credit: Graham Bourque; NMDC team member Leah Johnson at the Sevilleta LTER site, Credit: Buck Hanson; NMDC team member Francie Rodriguez in the Waimea Valley of Hawaii, Credit: Thomas Chapin.

A series of photos from the National Microbiome Data Collaborative. Starting top down: Location: Schynige Platte in Switzerland, Credit: Julia Kelliher; Sevilleta LTER site, Credit: Buck Hanson; Prosser Soil Sampling in Washington, Credit: Montana Smith; East River, Credit: Patrick Sorensen; Oksrukuyik Creek (OKSR) Alaska Credit NEON Science.

nmdc
National Microbiome Data Collaborative

U.S. DEPARTMENT OF ENERGY | Office of Science

microbiomedata.org @microbiomedata @microbiomedata https://www.linkedin.com/company/microbiomedata/

23-EGSB-14053