DOE BSSD Performance Management Metrics FY24 End-of-Year Summary

September 20, 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.

FY24 End-of-Year Summary.

Executive Summary

Our team has made excellent progress this past year towards advancing NMDC’s infrastructure and processes to better serve the microbiome research community. This progress is driven by the need for scalable, robust solutions that align with the evolving scientific ecosystem and ensure that data remains findable, accessible, interoperable, and reusable (FAIR). This year, our key achievements include the introduction of a monthly software release cycle; launching the persistent identifier service; and initiating an overhaul of the schema to better capture complex laboratory processing methodologies and workflow processes.

The software release process allows for regular updates to the NMDC's public-facing products, ensuring that improvements in runtime, server management, and schema handling are consistently implemented. These updates are carefully tracked and documented through versioning on GitHub, enhancing transparency and facilitating easier access to improvements and contributor details. The persistent identifier service, launched in January 2023, now generates stable, unique identifiers for all hosted data, providing long-term reference points for studies, samples, and workflows. This year, we successfully integrated these identifiers into all legacy NMDC data to be fully consistent with FAIR best practices.

Across our four products, the Submission Portal, NMDC EDGE, the Data Portal, and the newly developed NMDC Field Notes, we have made improvements based on our robust user research process. This year, we streamlined connections from the Submission Portal to the Data Portal, as well as syncing with the new NMDC Field Notes mobile app. We have enhanced programmatic access to NMDC data, providing researchers with scalable, efficient ways to interact with and query large datasets via our public API. We are also aiming to complete the data lifecycle process by connecting to primary data archives to preserve data. Towards this, we have initiated direct submission of sequence data to NCBI following protocols for batch submission as a ‘trusted broker.’

We have maintained a strong focus on community engagement and inclusivity. The Ambassador and Champions programs have grown dramatically this past year, with the Ambassador program nearly doubling in cohort size. These flagship programs foster collaboration, feedback, and build a strong network across the diverse areas of microbiome research. By engaging with a diverse group of researchers, we ensure that NMDC’s software, data, and services reflect the needs of the broader microbiome research community. User-centered design practices, including extensive usability testing, have been instrumental in guiding product development and ensuring that NMDC products remain user-friendly and accessible. The past year’s developments have positioned NMDC as a leader in advancing infrastructure, standards, and community building for microbiome research. We continue to advance scientific progress by providing critical tools and resources that enable collaboration and ensure the long-term usability of microbiome data.

Advancing NMDC's infrastructure & processes

Our team has made excellent progress this past year towards advancing NMDC's mission with robust process improvements and major infrastructure updates. These improvements will enable future scaling, align the NMDC towards a larger data ecosystem, and ensure stable production services to our user community. Below we describe these advances in software and data management, and the new processes our team has developed this year.

Monthly software releases

In December 2023, we established a monthly release cycle to push updates from our development environment to production. This monthly schedule supports regular updates across public-facing software products for the NMDC, including the runtime, server, and schema. Each production deployment receives a version tag and is tracked across the NMDC github repos with detailed information on improvements, updates, and contributor information (for example, NMDC schema release).

Integrated ORCiD authentication

We have implemented a standard, harmonized authentication process across all NMDC products and services, using ORCiD. ORCiD provides a secure authentication service for users to access resources using their institutional logins. This integration ensures that every NMDC offering now utilizes ORCiD authentication. We leverage ORCiD member capabilities which grants our team access to valuable ORCiD user metrics and additional user information, such as email and affiliation, when authorized. We have implemented a "three-legged OAuth" based authentication protocol, utilizing a common backend that allows us to exchange tokens across different services. Our approach enhances the user experience and data integration across the NMDC ecosystem.

Persistent Identifiers

Persistent identifiers are essential to make data findable, accessible, interoperable, and reusable (FAIR) and support links across studies, samples, and workflow runs in stable, unique, and long-term ways to reference digital objects. We are leading the way towards interoperability, attribution, and linking across resources through the NMDC persistent identifier service that was launched in January 2023. This service generates persistent identifiers for all data hosted within the NMDC and is documented here. This past Spring, we successfully completed a large update across all legacy data to include NMDC persistent identifiers as the primary identifier to be fully compliant with FAIR best practices. This effort included updates to all studies, biosamples, and multi-omics workflow outputs, along with enabling future minting of persistent identifiers across all NMDC workflows.

The NMDC schema

The NMDC schema was developed to support handling and modeling of data in a robust, yet flexible manner. Using the Linked Data Modeling Language (LinkML), we are able to model sample processing and data generation components and support community standards, such as the Minimum Information about any (x) Sequence (MIxS) standard [1] from the Genomic Standards Consortium (GSC) and the Ontology of Biomedical Investigations (OBI) framework [2]. This past year, we worked closely 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. This effort is described in the book chapter, "A Practical Approach to Using the Genomic Standards Consortium MIxS Reporting Standard for Comparative Genomics and Metagenomics" [3].

A major effort initiated this past year was to update the NMDC schema to better capture laboratory processing methodology and analysis workflow processes. The impetus for this effort stems from the complex nature samples are handled and managed to generate multi-omics data. Capturing this information is necessary both for being able to trigger the appropriate workflow, and for users to accurately interpret and analyze the processed omics data. To our knowledge, this is a new approach **to fully model multi-omics microbiome data from sample to processed data** and holds promise for data integration across studies and across computational platforms. The effort introduced a new, cleaner upper-level organization with the introduction of "Material Processing" and "Data Generation" classes, together with more granular subtypes (e.g., Mass Spectrometry, Library Preparation, Chromatographic Separation). This new structure makes the schema easier to maintain moving forward, as new kinds of sample processing and instrumentation can be easily slotted in, and also provides a consistent and coherent structure for programmatic queries of the data. We anticipate all NMDC systems to be fully transitioned to the new schema by the end of this year.

Infrastructure Advancements

We have continued to make advancements in our backend infrastructure capabilities to improve usability, scaling, and resilience. Improvements to our NMDC's Application Programming Interface (API) and backend runtime services included support for advanced custom queries, aggregations, and new endpoints based on user needs. We made significant updates to our database validation pipelines and migration tools, to improve how we maintain data integrity across schema versions. We have expanded our capabilities for high-performance bulk data downloads by making all NMDC data available through Globus. We have also instituted processes to ensure automated data/metadata backups and infrastructure redundancy.

Unique resources to enable microbiome science

Our commitment to open and equitable research in microbiome science is the foundation for NMDC's infrastructure development and has been a driver this past year across our four products: the Submission Portal, NMDC EDGE, the Data Portal, and the newly developed NMDC Field Notes. Bolstered by our infrastructure and process improvements, we have made improvements to streamline connections from the Submission Portal to the Data Portal, along with connecting to the new NMDC Field Notes. Further, we have initiated direct submission of sequence data to NCBI following protocols for batch submission as a ‘trusted broker.’ Below we describe these improvements to the NMDC products developed this year.

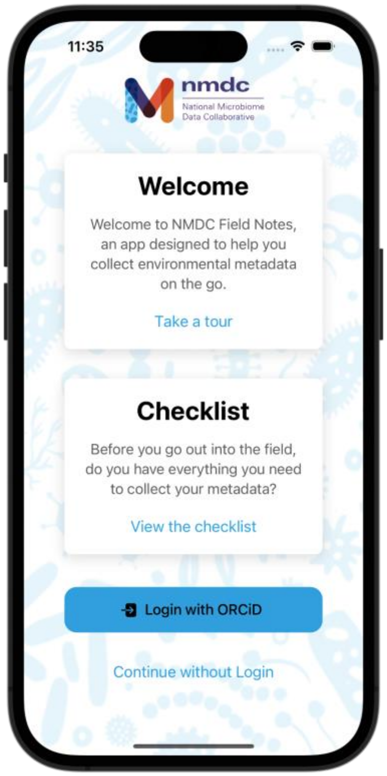
Field Notes and Submission Portal: making metadata capture easy

The NMDC Field Notes App and Submission Portal are flexible, template-driven tools designed to make metadata capture and adherence to community standards easy for information about studies, samples, and assays. A major new effort initiated at the NMDC team retreat in October 2023 was the development of a metadata mobile app to collect sample information in the field. The mobile app streamlines collection of sample information with automated syncing with the NMDC Submission Portal through a common login using ORCiD credentials. Sample information is validated in real time against the NMDC metadata standards which are based on MIxS standard and uses the environment-specific templates (e.g., soil or water). Automated data collection features like geolocation and date and time can be easily selected using the basic functions on either an Android device or an iOS device. Further, the app was designed for both online and offline mode to accommodate access in the field regardless of internet connectivity.

The mobile app was first previewed during the NMDC Town Hall at the 2024 Biological Systems Science Division’s Annual PI meeting. Over the course of the summer, we actively recruited nearly 50 beta testers and created a detailed document for how to access the app through Apple’s TestFlight app or Android's APK files. NMDC Field Notes has been showcased at ASM Microbe in Atlanta, the GSC’s Annual Meeting in Tucson, and the ISME19 meeting in Cape Town, South Africa.

Since the NMDC schema underpins both the NMDC Field Notes App and the Submission Portal, users are able to seamlessly transition across the two platforms. This design flexibility accommodates several ways users can interact and collect information throughout the sampling and data management process. We continue to evaluate how to make improvements across the mobile app and portal through our user research processes (described below).

The NMDC Submission Portal provides a place for data generators to complete standardized, machine readable, and structured metadata that is compliant with the NMDC and the Genomic Standards Consortium (GSC). Researchers can submit metadata via the Submission Portal for non-DOE and DOE projects, and they can use this product for submitted samples to a DOE User facility, ensuring their metadata is compliant and complete prior to data generation. This tool captures the required information



Welcome to NMDC Field Notes, an app designed to help you collect environmental metadata on the go.

Before you go out into the field, do you have everything you need to collect your metadata?

Figure 1. The NMDC Field Notes App is designed to capture sample metadata in the field.

for sharing a study on the Data Portal and the metadata required for a variety of sample types, based on the GSC's environmental Extensions [3]. Users are provided an interface with multiple tools to simplify the metadata completion process, such as, show/hide terms based on criteria, a term search function, real-time validation, and expanded term details like description, examples, and guidance. This product and the tools it provides lower the barriers to standardized metadata capture and FAIR data.

Beyond the Field Notes App to Submission Portal automated syncing, the Submission Portal has also had several new features and updates this past year to optimize submissions and improve the user experience, and specifically with a focus on DOE Facility users from EMSL and the JGI. This past year, we made many updates to the Submission Portal user interface (UI) and functionality based on user research and feedback. Among these, some key UI updates include providing easy access links to documentation and how-to guides, and providing confirmation of a saved or submitted submission. Additionally, updates were made to the NMDC Schema to capture more information about a study, and the Submission Portal UI has been updated to capture this information from submitters. Functionality was added to allow multiple editors access to a submission via ORCiD incorporation. These updates are outlined below.

New Features

- The Submission Portal home page was updated to include easy links to the quick start user guide, how-to's for submitting, video tutorials, and reference material to help users better navigate submitting study and biosample information. Further, submissions are now sortable once logged in to allow submitters to easily find submissions if there are multiple submissions.
- We have ongoing work to improve the user experience for JGI and EMSL users to submit project and sample information that 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, as well as the requirements for long-read sequencing.
- We added access permissions to a submission, allowing multiple contributors or viewers on a single submission enabling collaboration and connection across research teams for direct contribution and access. Further, we have established a process for maintaining private submissions that adhere to embargo policies that will trigger sharing of data publicly.

NMDC EDGE: providing user-friendly bioinformatics tools

NMDC EDGE is a web application that provides broad access to the NMDC workflows for all microbiome scientists and was modeled after the generalized EDGE bioinformatics platform [4]. To date, the standardized production-quality workflows developed by the JGI and EMSL have largely been limited to the facilities for which they were developed. NMDC EDGE provides access to these standardized workflows to process raw multi-omics data and produce interoperable annotated data from metagenomes, metatranscriptomes, metaproteomes, and natural organic matter characterizations. This past year, we have made improvements to NMDC EDGE to improve the user experience as outlined below.

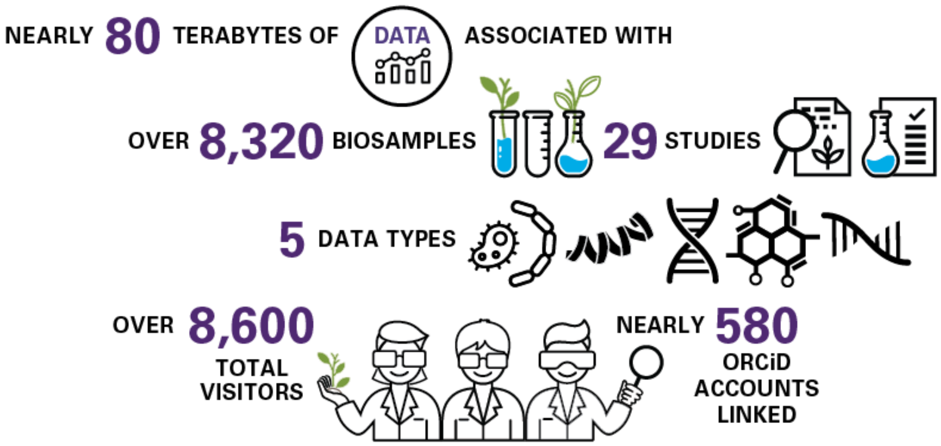
New Features

- Similar to the Submission Portal, the NMDC EDGE home page was updated to include the quick-start guide and tutorials that are continually updated. Aligned with our commitment to increasing the accessibility of the NMDC products, our team has updated and translated most of the workflow user guides into Spanish and French this past year.
- The 'Retrieve SRA Data' workflow enables users to access publicly available data from NCBI's Sequence Read Archive (SRA) and directly import into NMDC EDGE [5]. This feature streamlines the use of public data for data processing with NMDC's sequence-based workflows.
- Improvements to the metagenome workflow outputs have been made to include basic visualizations of taxonomic and functional abundance profiles. These new features were user requests and are aimed at helping researchers visualize and discover their processed data.

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

The Data Portal and public API democratize access to microbiome studies integrating different data types, supporting persistent identifiers and links to external repositories. This past year, we added three new studies and 1,349 samples to the Data Portal. Sample types added include aquatic, soil, and host-associated environments such as the phyllosphere. Most recently we added samples from the EcoFab ring 2 trial. We processed 624 natural organic matter datasets for EMSL's 1000 Soils research campaign and for an Alaska permafrost study and processed 767 datasets from various studies through our metagenome workflow. Additionally, we also updated to NMDC style identifiers as our primary identifier for eight studies and reprocessed those datasets through our proteomics and annotation workflows where applicable. Our latest Data Portal usability testing resulted in 136 insights and 42 action items.

BY THE NUMBERS



NEARLY 80 TERABYTES OF DATA ASSOCIATED WITH

OVER 8,320 BIOSAMPLES 29 STUDIES

5 DATA TYPES

OVER 8,600 TOTAL VISITORS

NEARLY 580 ORCiD ACCOUNTS LINKED

Improvements to our public API this year are primarily backend development including improved testing coverage, speed improvements with indexing, and an overhaul to more easily connect related records which is agnostic to future schema changes. User research on our API this year resulted in 84 insights leading to 18 action items to improve our documentation and accessibility of the API user interface and endpoints. Updates to support schema changes are expected to be released mid-October which include updates to tests, endpoint updates, and documentation updates.

New Features

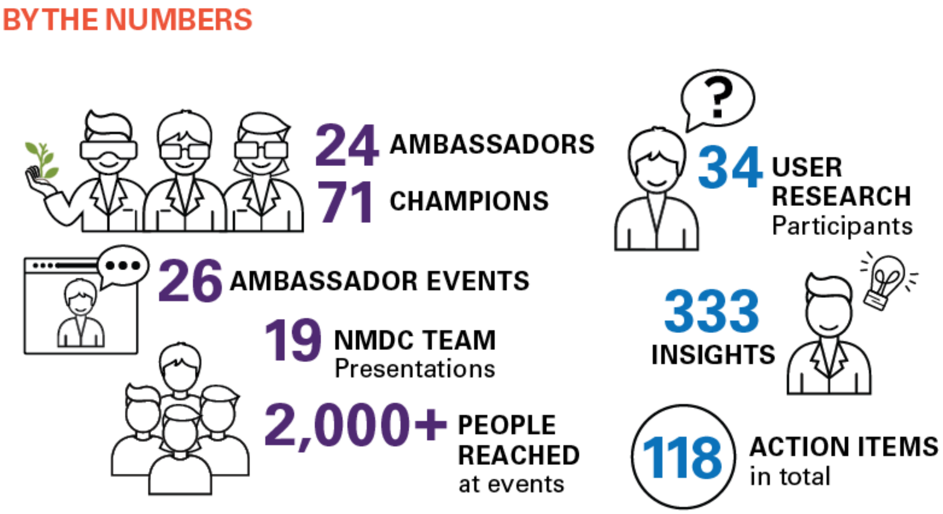
- This year we updated the Data Portal to better represent studies. This included updates to study landing pages and introducing groupings of related studies, for example by DOE Science Focus Area (SFA) and having a separate category for large research consortiums.
- We have also expanded the types of identifiers, names, and descriptions that can be searched on. Several action items identified from our user research have been addressed such as a freezing the faceted search pane, additional help guidance, and an improved upset plot interface. Backend and frontend work to support the extensive schema changes described above are expected to be released mid-October. Work is in progress to expand the Data Portal's functional search to COG, Pfam, and GO and we expect this to be completed before the end of the calendar year. The inclusion of COG and Pfam were driven by the user research we conducted this year.
- New API features include ORCiD authentication and the ability to run aggregation queries that support custom query pipelines. We have also updated and added new API endpoints based on evolving schema and user needs, such as endpoints for schema introspection and navigation of associations across data objects.

Supporting collaboration and building an inclusive community

This past year, we have expanded our Ambassador and Champions programs, strengthened our user research activities across all of the NMDC products, launched new ways to engage through social media, and advanced large-scale data initiatives within and beyond the Department of Energy. Our engagement efforts continue to build a strong foundation for how we work with the research community to shape the NMDC. In January, we expanded our engagement strategy to encompass communications and updated our social media playbook to include our new LinkedIn and Instagram accounts. We also conducted an overhaul of the NMDC website to add new pages for Data Standards, Data Integration, previous Newsletters, the new Field Notes mobile app, and refreshed our team page to specify team roles. We also published several blog posts highlighting specific accomplishments and collaborations, along with the NMDC Snapshot series introducing and promoting our Ambassador cohorts.

Ambassadors & Champions

From the beginning of the NMDC, we developed the Ambassador and Champions programs to foster learning, support networking, and translate feedback to actions through our user research program. To date, we have over 70 Champions actively participating in the program and recruited 24 early career researchers for the Ambassador program this year. In February, working groups were launched in the Champions program to focus on reusing microbiome data from the NMDC Data Portal. Both Champions and Ambassadors frequently contribute to user research, the development of training materials, and advocate for data stewardship within their networks. We also make all Ambassador training materials publicly available.


BY THE NUMBERS
24 AMBASSADORS
71 CHAMPIONS
26 AMBASSADOR EVENTS
19 NMDC TEAM Presentations
2,000+ PEOPLE REACHED at events
34 USER RESEARCH Participants
333 INSIGHTS
118 ACTION ITEMS in total

Fostering inclusion, diversity, and equity

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 Inclusion, Diversity, Equity, and Accountability (IDEA) Action Plan helps us achieve our goals and move towards creating an even more equitable community. In January, we released an updated IDEA Action Plan that outlines 41 action items for the year and also summarized our accountability metrics from the previous milestones. We also launched our IDEA working group to discuss making microbiome research more equitable. This discussion stemmed from our participation at the National Diversity in STEM conference in October.

User research: understanding the needs of microbiome researchers

As the NMDC team continues to drive new improvements and ideas, we follow a user-centered design process to ensure user feedback is incorporated into our product development. This approach allows us to create products that meet the needs of our users across the interdisciplinary field of microbiome science. This year, we focused on the Data Portal and API as well as gathered user feedback to support the development of the new NMDC Field Notes mobile app. We spoke with 24 researchers across diverse research fields, institutions, and career stages. Data Portal usability testing yielded 136 insights and 42 action items. These actions will improve our Data Portal accessibility through added features such as bulk download file size information, adjustments to our filtering and search capabilities, and additional search abilities such as taxonomy and functional annotation. The API user research round led to 84 insights and 18 action items.

These actions will make the API commands and endpoints more accessible through updates to the graphical user interface, streamlined documentation, and robust tutorials, examples, and help guidance.

To learn what researchers would need from a mobile app to facilitate metadata collection in the field, we spoke with 12 individuals across diverse fields. These discussions led to 50 insights and 18 action items to support the development and prioritization of features in the NMDC Field Notes mobile app. For instance, researchers requested the ability to create metadata templates to ensure they are capturing their necessary metadata. We have implemented this feature by providing the ability to toggle the visibility of the NMDC metadata fields in the app for each metadata package. Another example includes feature requests for capturing photographs in the field and using a barcode scanner for sample collection. We are currently hosting a beta testing round for the NMDC Field Notes app to capture ideas for improvement before the public release. With this year's user research rounds, we have successfully completed an iteration of user research on all of the NMDC products.

References

1. 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.

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

3. 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.

4. Li P-E, Lo C-C, Anderson JJ, Davenport KW, Bishop-Lilly KA, Xu Y, et al. Enabling the democratization of the genomics revolution with a fully integrated web-based bioinformatics platform. Nucleic Acids Res. 2016;45: 67–80.

5. Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR, O'Sullivan C. The Sequence Read Archive: a decade more of explosive growth. Nucleic Acids Res. 2022;50: D387–D390.