



nmdc

National Microbiome
Data Collaborative

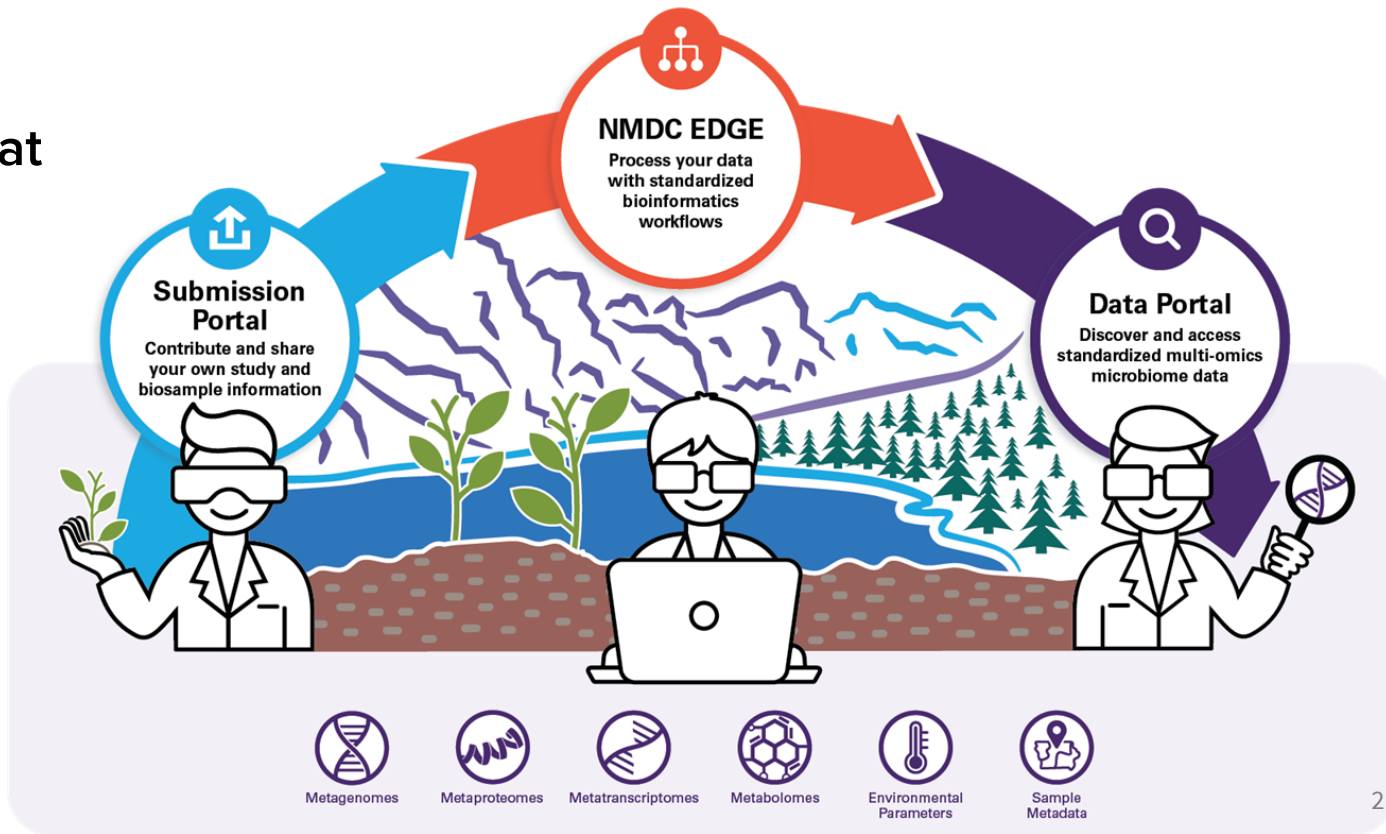
**Connecting data, people, and ideas with the
National Microbiome Data Collaborative**

Emiley Eloë-Fadrosch | Spring BERAC | April 21, 2023

What is the NMDC?

A microbiome data sharing network that supports:

- **Data standards**
- **Robust infrastructure**
- **Community building**



Supporting microbiome discovery

What role do microbes play in the persistence of soil carbon?

How do microbes mediate watershed scale nutrient transformations?



Support new ways for researchers to ask questions, compare across studies, and reuse data

How do microbes impact biogeochemical cycling locally and across continental scales?

Product Initiatives



Submission Portal



Lower barriers to collect study and biosample data

NMDC EDGE



Streamline multi-omics data processing

Data Portal & API



Access and discovery of microbiome information

Engagement

User Facilities



Individuals



Strategic Partners



Making standards FAIR



Minimal Information about any (X) Sequence (MlxS)

Machine-actionable, web- searchable standard

MlxS version 5



MlxS version 6



Reporting standard

- Managed as Excel spreadsheets
- Not machine-actionable
- Not following FAIR
- Not modular



Data Validation

Submission Tools

Support for persistent identifiers

Persistent identifiers and mappings to external resources enable interoperability

- Created a consistent system and format for all identifiers created within NMDC, ranging from

updated with expanded alternative identifier fields to support IGSN, GOLD, IMG, ESS-DIVE, INSDC, and massive

Three paths for metadata submission

Supporting metadata submission

User Facilities



User Facility templates and requirements for sample submission



Individuals



Templates for data generated anywhere to be compliant with community standards

Strategic Partners



Working with partners to support large-scale data sharing and metadata mapping

Submission Portal



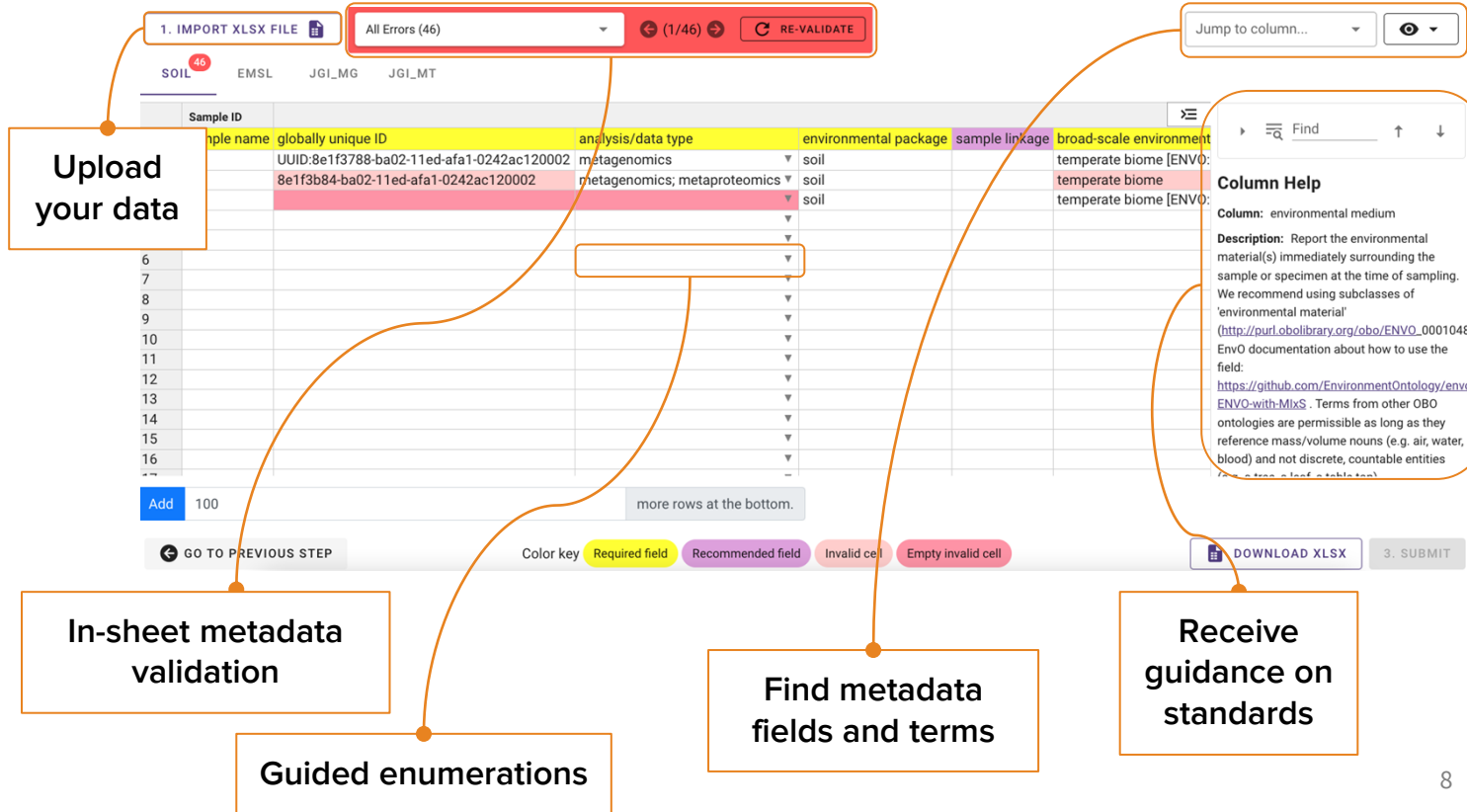
Lower barriers to collect study and biosample data

Streamlined metadata submission

Submission
Portal



Lower barriers
to collect
study and
biosample
data



The screenshot shows a web interface for submitting metadata. At the top, there is a '1. IMPORT XLSX FILE' button, a dropdown menu showing 'All Errors (46)', and a 'RE-VALIDATE' button. Below this is a table with columns: 'Sample ID', 'Sample name', 'globally unique ID', 'analysis/data type', 'environmental package', 'sample linkage', and 'broad-scale environment'. The table contains several rows of data, with some cells highlighted in yellow (required) or pink (invalid). A 'Column Help' sidebar is open on the right, providing details for the 'environmental medium' column. At the bottom, there are buttons for 'GO TO PREVIOUS STEP', 'DOWNLOAD XLSX', and '3. SUBMIT'. A legend for 'Color key' shows 'Required field' (yellow), 'Recommended field' (purple), 'Invalid cell' (pink), and 'Empty invalid cell' (light pink).

Upload your data

In-sheet metadata validation

Guided enumerations

Find metadata fields and terms

Receive guidance on standards

Column Help

Column: environmental medium
Description: Report the environmental material(s) immediately surrounding the sample or specimen at the time of sampling. We recommend using subclasses of 'environmental material' (http://purl.obolibrary.org/obo/ENVO_0001048: EnvO documentation about how to use the field: <https://github.com/EnvironmentOntology/envo> ENVO-with-MixS . Terms from other OBO ontologies are permissible as long as they reference mass/volume nouns (e.g. air, water, blood) and not discrete, countable entities (e.g. trees, leaf, antibodies).

Standardized bioinformatics workflows



nmdc

National Microbiome
Data Collaborative



Fully standardized and
containerized workflows to
support multi-omics
integration and data reuse



metagenomics



metatranscriptomics



metaproteomics



metabolomics



Standardized NMDC Workflows

Available



Under development



Discovery with the Data Portal & API



nmdc
National Microbiome
Data Collaborative

Data
Portal & API



Access and
discovery of
microbiome
information

Found 2449 results.

search

Study

PI Name

Function

KEGG Term

Sample

Depth

Collection date

Latitude

Longitude

Geographic Location Name

GOLD Ecosystems

GOLD classification

ENVO

Broad-scale Environmental Context

Local Environmental Context

Environmental medium

Omics Processing

Instrument name

Omics type

Processing institution

OMICS

- Metagenome: 2055
- Natural Organic Matter: 1570
- Metatranscriptome: 111
- Proteomics: 52
- Metabolomics: 34

ENVIRONMENT

Collection Date

14 Studies

- Defining the functional diversity of the Populus root microbiome
- Riverbed sediment microbial communities from the Columbia River, Washington, USA
- Bulk soil microbial communities from the East River watershed near Crested Butte, Colorado, United States
- Earth Microbiome Project Multi-omics (EMP500)
- Determining the genomic basis for interactions between gut fungi and methanogenic archaea

NMDC Dataset API

user

- GET /api/me Return The Current User Name
- GET /api/users Get Users
- POST /api/users/{id} Update User
- GET /logout Log Out Of The Current Session

aggregation

- GET /api/search Text Search
- GET /api/summary Get Database Summary
- GET /api/stats Get Aggregated Stats
- POST /api/environment/saskey Get Environmental Saskey
- POST /api/environment/geospatial Get Environmental Geospatial

biosample

- POST /api/biosample/search Search For Biosamples
- POST /api/biosample/facet Get All Values Of An Attribute
- POST /api/biosample/binned_facet Get All Values Of A Non-String Attribute With Binning

API enables data
access and sharing
across systems

Supporting data integration & access



nmdc

National Microbiome
Data Collaborative

MODELS

Improvement of
watershed models
to include chemical
and biological
processes

ecosys

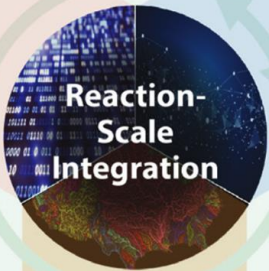


PFLOTRAN

DATA

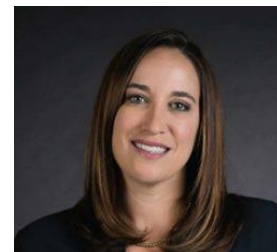


Data assembly,
integration,
and storage



DISTRIBUTED SCIENCE APPROACH

Geochemistry, hydrology, metabolites, metagenomes, and metatranscriptomes



Kelly Wrighton



Mikayla
Borton

U.S. DOE. 2019. Open Watershed Science by Design: Leveraging Distributed Research Networks to Understand Watershed Systems Workshop Report, DOE/SC-0200, U.S. Department of Energy Office of Science.

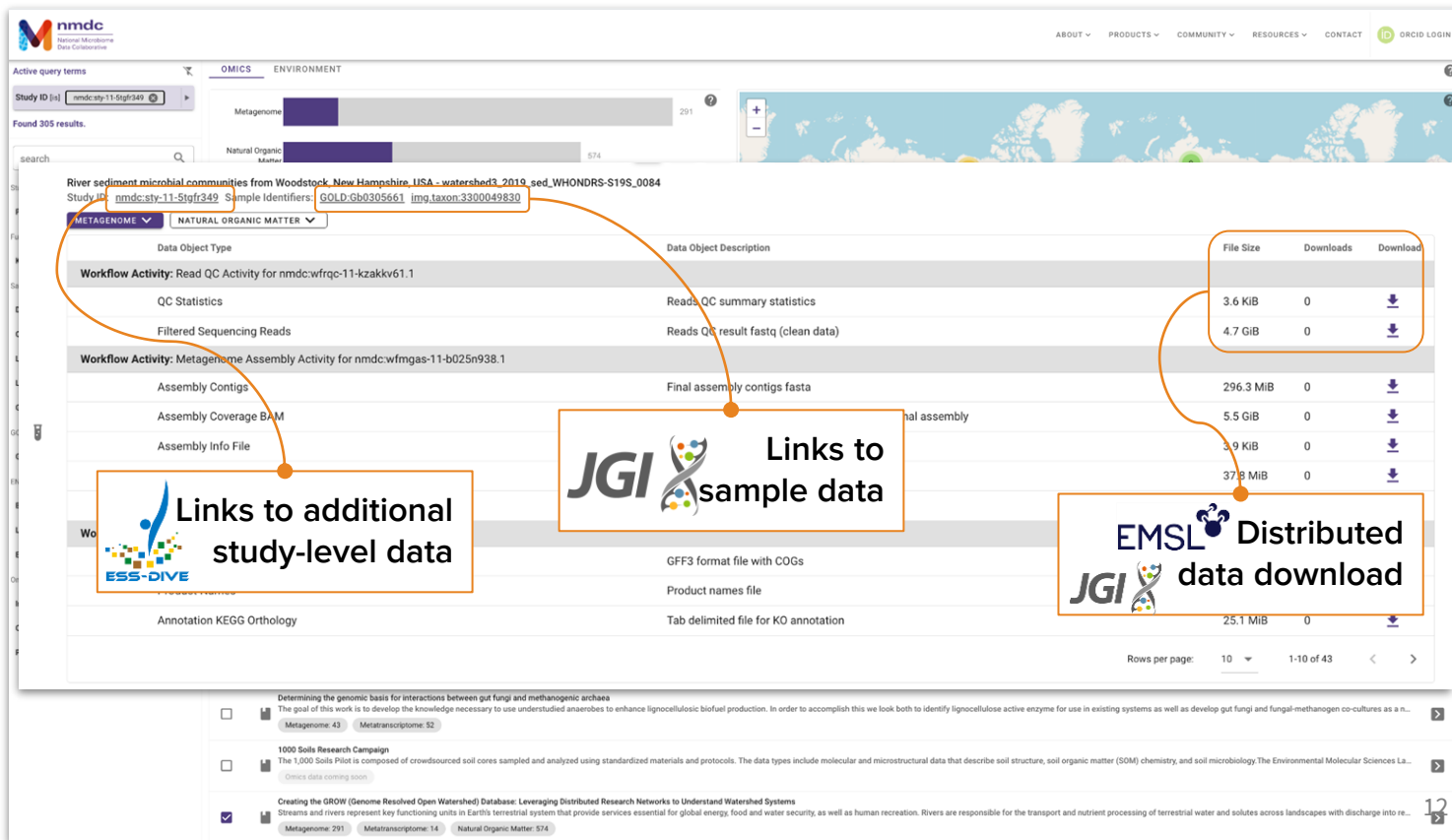
Supporting data integration & access

DATA



Integration of
metagenome,
metatranscriptome,
and NOM data
generated at JGI &
EMSL

Associated at the
biosample level



nmDC National Microbiome Data Collaborative

Active query terms: OMICS ENVIRONMENT

Study ID list: nmDC:st11-51grf349

Found 305 results.

search

Metagenome 291

Natural Organic Matter 574

River sediment microbial communities from Woodstock, New Hampshire, USA - watershed3_2019_sed_WHONDRS-S19S_0084

Study ID: nmDC:st11-51grf349 Sample Identifiers: GOLD:Gb0305661 img.taxon:3300049830

METAGENOME NATURAL ORGANIC MATTER

Data Object Type	Data Object Description	File Size	Downloads	Download
Workflow Activity: Read QC Activity for nmDC:wfrqc-11-kzakkv61.1	QC Statistics	3.6 KiB	0	Download
	Filtered Sequencing Reads	4.7 GiB	0	Download
Workflow Activity: Metagenome Assembly Activity for nmDC:wfmgas-11-b025n938.1	Assembly Contigs	296.3 MiB	0	Download
	Assembly Coverage BAM	5.5 GiB	0	Download
	Assembly Info File	3.9 KiB	0	Download
	Final assembly contigs fasta	37.8 MiB	0	Download
	GFF3 format file with COGs			
	Product names file			
	Tab delimited file for KO annotation	25.1 MiB	0	Download

Rows per page: 10 1-10 of 43

Determining the genomic basis for interactions between gut fungi and methanogenic archaea
Metagenome: 43 Metatranscriptome: 52

1000 Soils Research Campaign
The 1,000 Soils Pilot is composed of crowd-sourced soil cores sampled and analyzed using standardized materials and protocols. The data types include molecular and microstructural data that describe soil structure, soil organic matter (SOM) chemistry, and soil microbiology. The Environmental Molecular Sciences La...
Omics data coming soon

Creating the GROW (Genome Resolved Open Watershed) Database: Leveraging Distributed Research Networks to Understand Watershed Systems
Streams and rivers represent key functioning units in Earth's terrestrial system that provide services essential for global energy, food and water security, as well as human recreation. Rivers are responsible for the transport and nutrient processing of terrestrial water and solutes across landscapes with discharge into re...
Metagenome: 291 Metatranscriptome: 14 Natural Organic Matter: 574

Links to additional study-level data

Links to sample data

EMSL Distributed data download

Future opportunities

Support a microbiome data sharing network, through **infrastructure, data standards, and community building**, that addresses pressing challenges in environmental sciences

Current activities

- Machine-actionable standards in collaboration with the GSC
- Flexible schema supports persistent identifiers for biosamples & data types
- Distributed data resources for integration across User Facilities
- Community building to support data stewardship

Enabling new science

- Focus on data standards to support interoperability
- Evolve standards & workflows with the community
- Leverage partnerships & resources to support a sustainable infrastructure ecosystem