nmdc

National Microbiome Data Collaborative

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

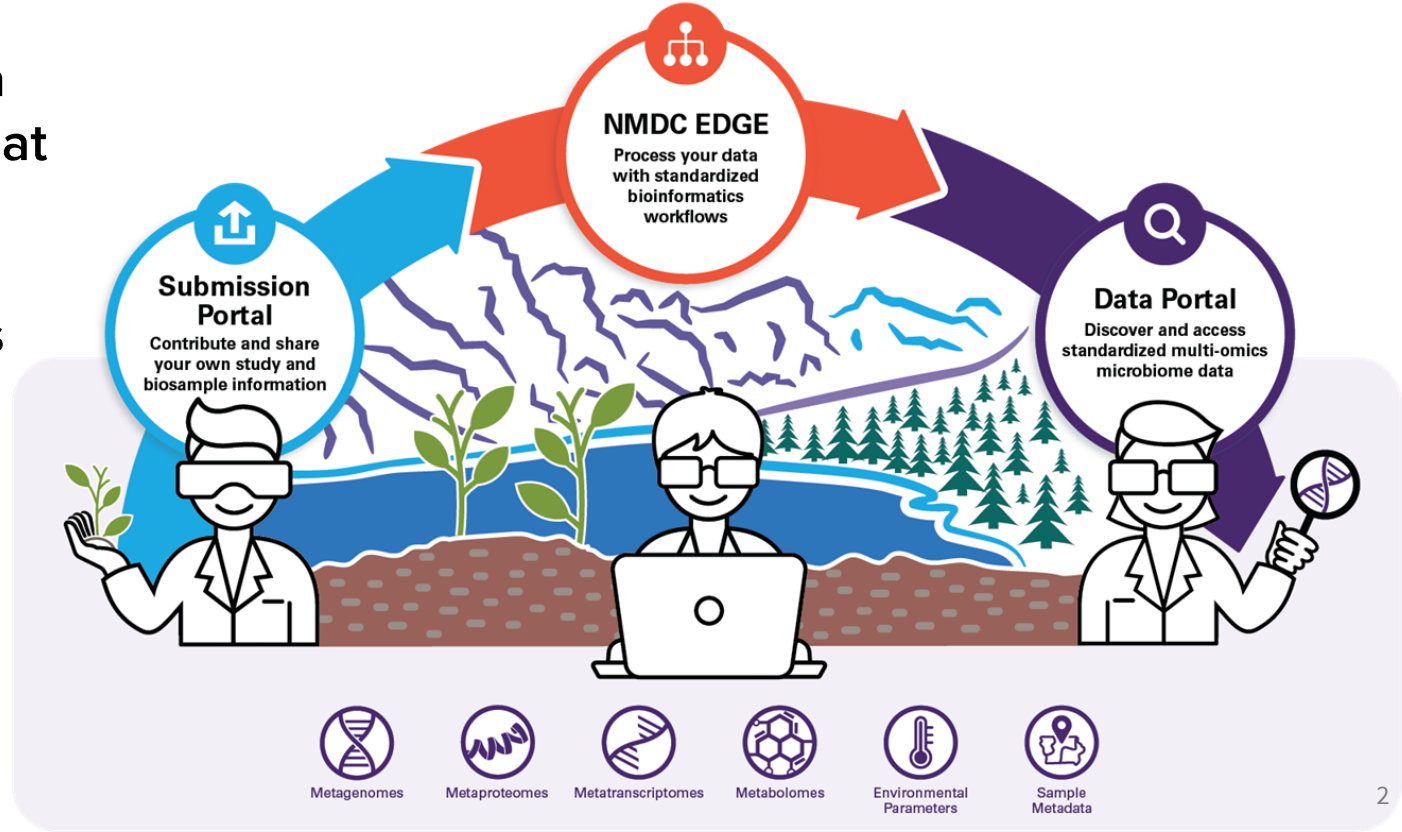
Emiley Eloe-Fadrosh | Spring BERAC | April 21, 2023

What is the NMDC?

nmdc
National Microbiome Data Collaborative

A microbiome data sharing network that supports:

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



Supporting microbiome discovery

nmdc
National Microbiome Data Collaborative

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

How do microbes mediate watershed scale nutrient transformations?

SFAs

1000 Soils RESEARCH CAMPAIGN

GROW Genome Resolved Open Watersheds

NSF

neon Operated by Battelle

earth microbiome project

FICUS Facilities Integrating Collaborations for User Science

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

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

Product Initiatives

nmdc
National Microbiome Data Collaborative

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

JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

EMSL

Individuals

Strategic Partners

NSF

neon Operated by Battelle

NASA GeneLab

Making standards FAIR

nmdc
National Microbiome Data Collaborative

genomic STANDARDS consortium

Minimal Information about any (X) Sequence (MIxS)

MIxS version 5

Reporting standard

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

Community Driven

Machine-actionable, web-searchable standard

MIxS version 6

JSON Schema

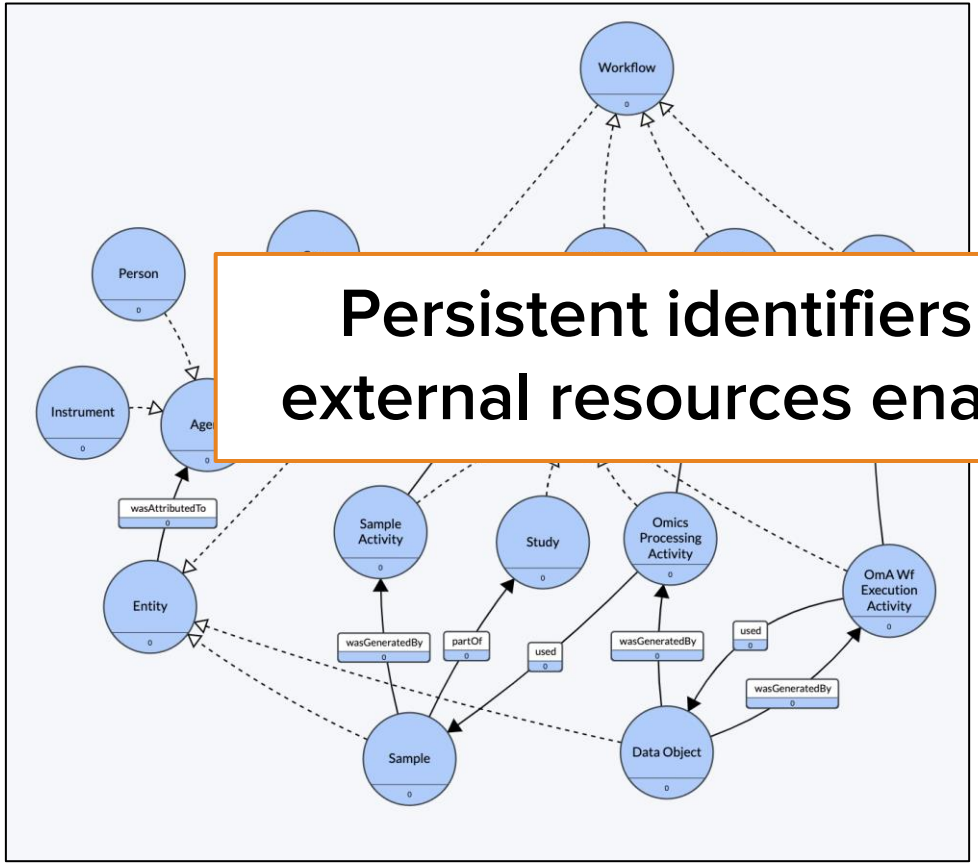
RDF

Submission Portal

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

Three paths for metadata submission

nmdc
National Microbiome Data Collaborative

Supporting metadata submission

User Facilities

JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

EMSL

Individuals

Strategic Partners

NSF

neon Operated by Battelle

NASA

GeneLab

User Facility templates and requirements for sample submission

FICUS Facilities Integrating Collaborations for User Science

Templates for data generated **anywhere** to be compliant with community standards

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

Submission Portal

Lower barriers to collect study and biosample data

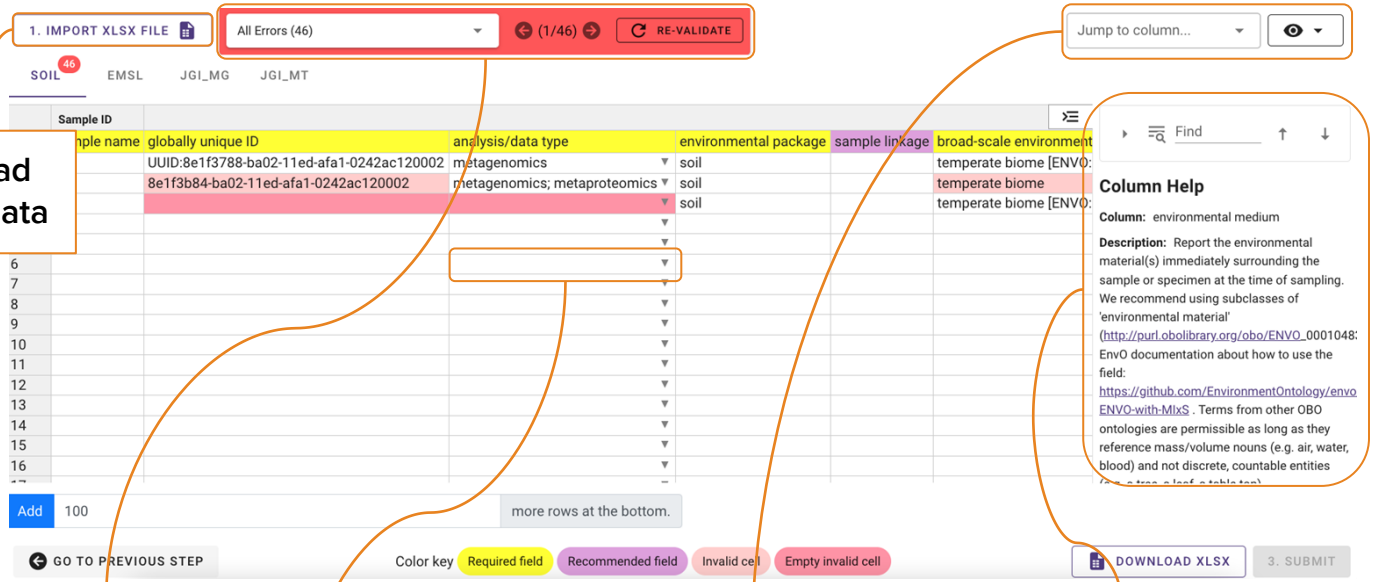
Streamlined metadata submission

nmdc
National Microbiome Data Collaborative

Submission Portal

Lower barriers to collect study and biosample data

Upload your data



In-sheet metadata validation

Guided enumerations

Find metadata fields and terms

Receive guidance on standards

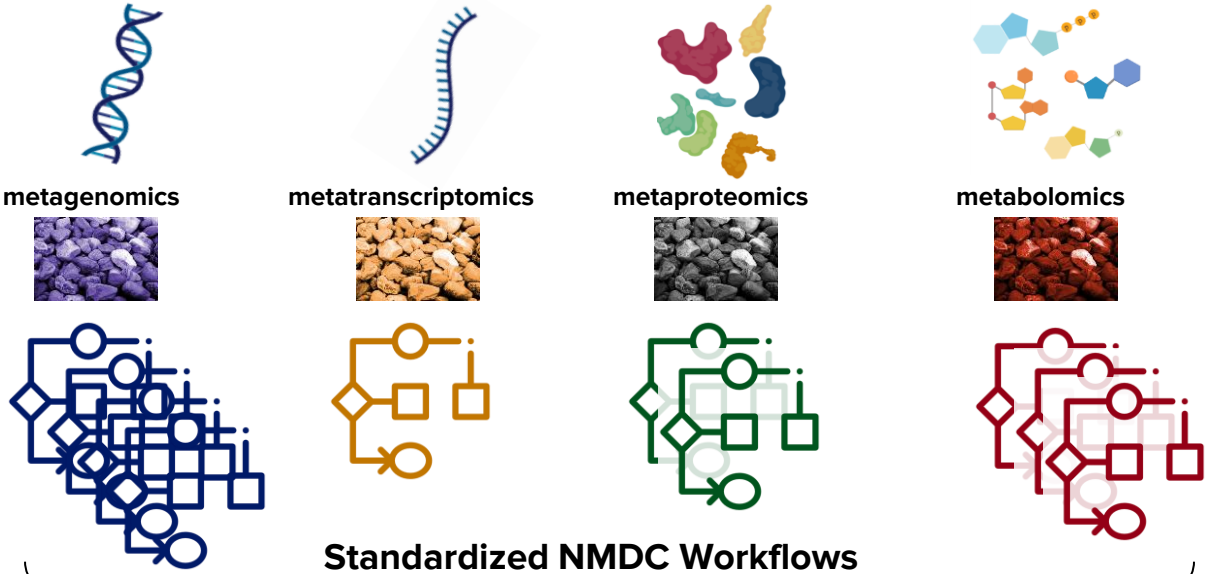
Standardized bioinformatics workflows

nmdc
National Microbiome Data Collaborative

JGI JOINT GENOME INSTITUTE DEPARTMENT OF ENERGY

EMSL

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



Standardized NMDC Workflows

Available

MEDGE

{wdl}

{wdl}

Under development

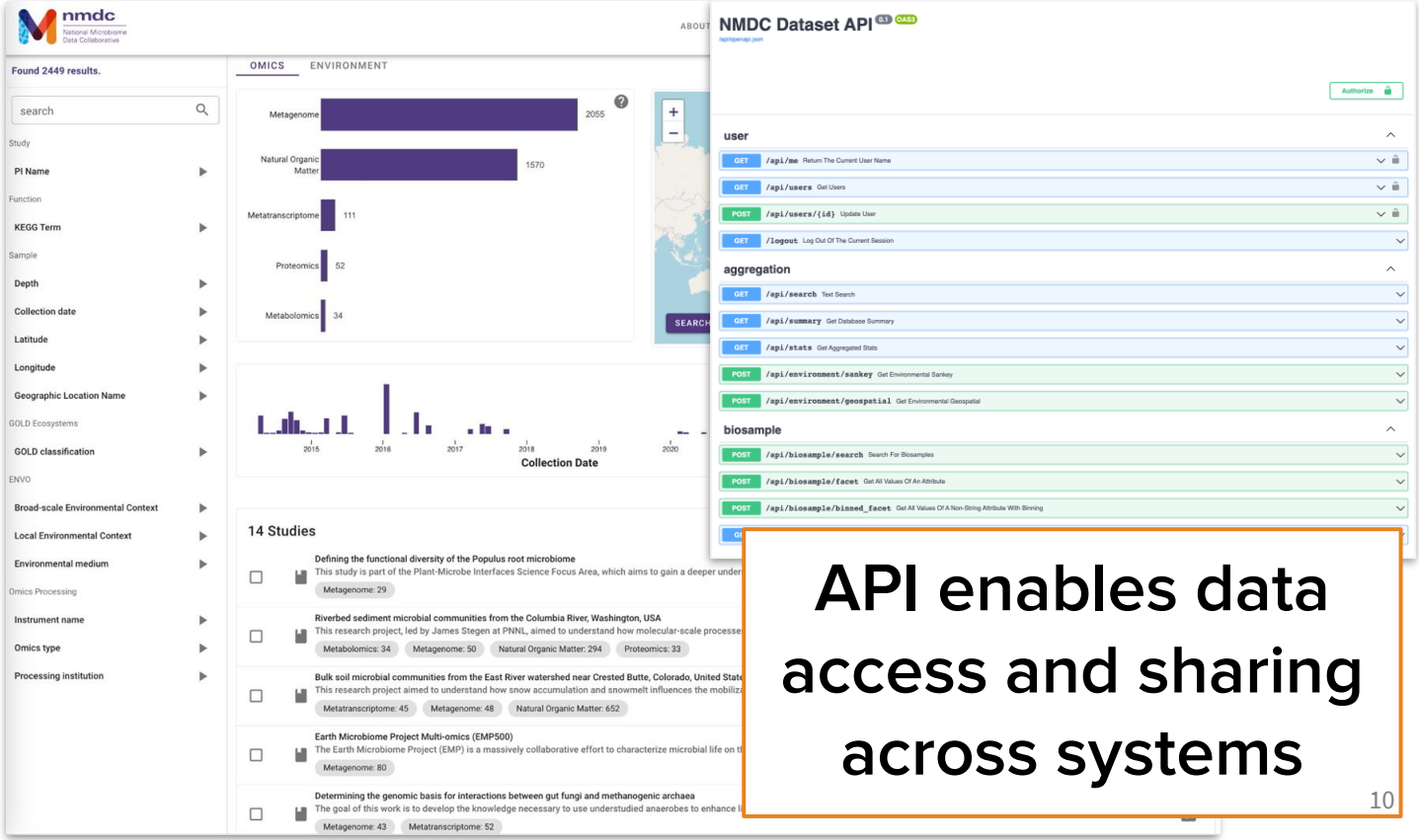
JAWS
JGI Analysis Workflow Service

Discovery with the Data Portal & API

nmdc
National Microbiome Data Collaborative

Data Portal & API

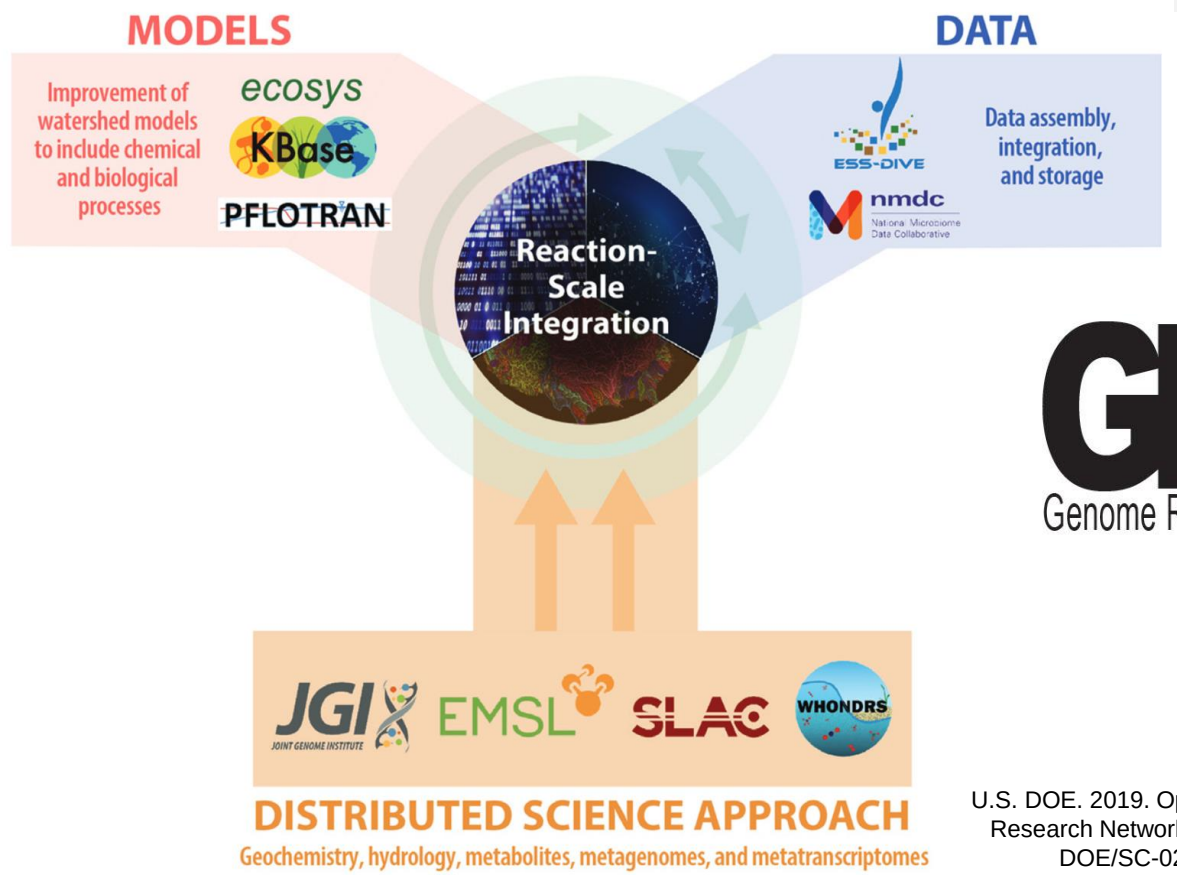
Access and discovery of microbiome information



API enables data access and sharing across systems

Supporting data integration & access

nmdc
National Microbiome Data Collaborative



GROW
Genome Resolved Open Watersheds

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

nmdc
National Microbiome Data Collaborative

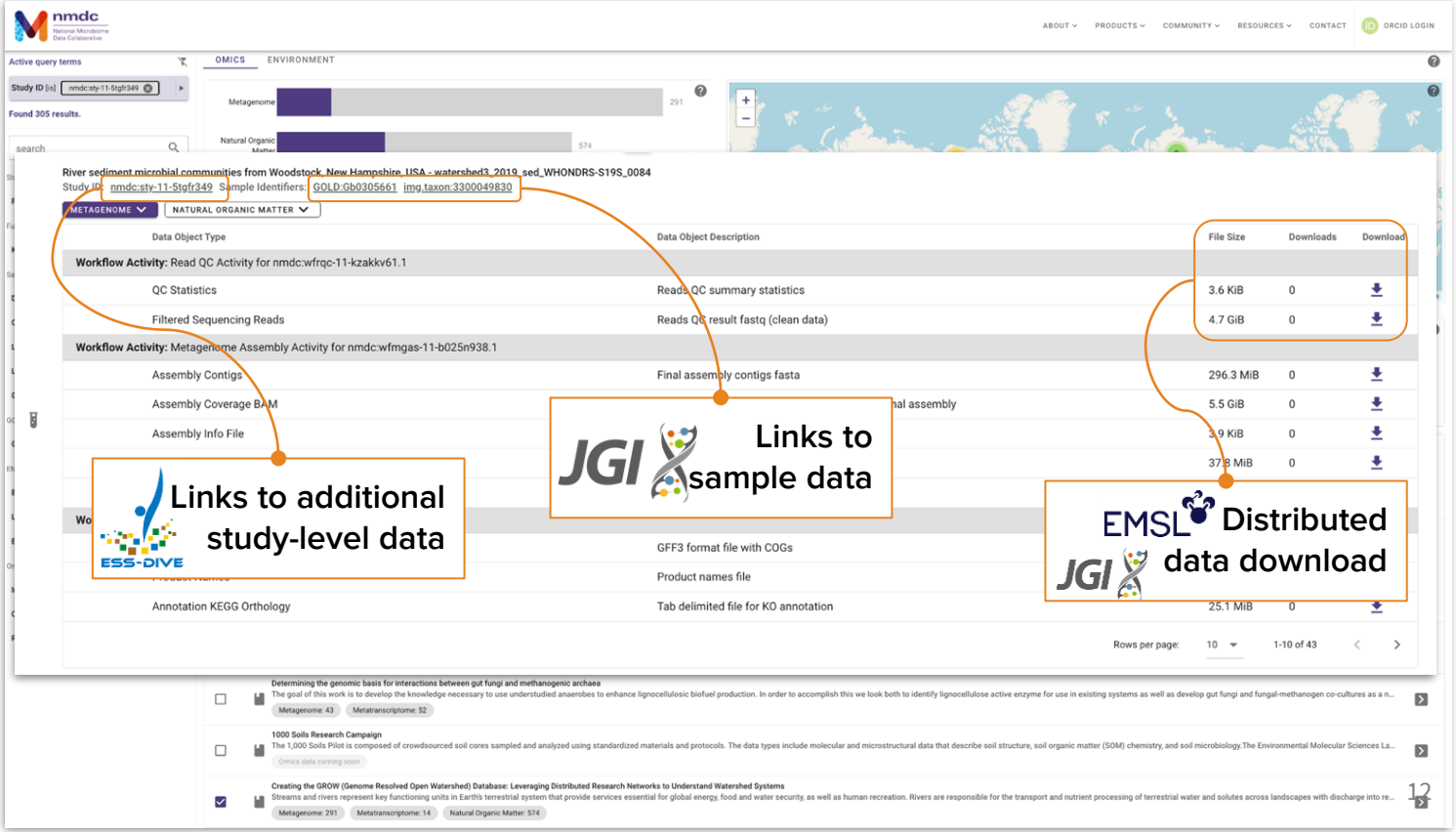
DATA

ESS-DIVE
nmdc
National Microbiome Data Collaborative

Data assembly, integration, and storage

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

Associated at the **biosample** level



Future opportunities

nmdc
National Microbiome Data Collaborative

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