



nmdc

National Microbiome
Data Collaborative

NMDC Sample Metadata Training
Changing Microbiomes Symposium NMDC workshop
May 31st, 2022

Objectives

- 1) Learn about metadata and data management best practices
- 2) Learn about NMDC Mission and explore the NMDC data portal
- 3) Identify metadata standards relevant to your studies

Resources



nmdc
National Microbiome
Data Collaborative

Workshop website:

https://github.com/erdavenport/PSUChangingMicrobiomes_NMDC

Agenda

Here's what's scheduled for the event.

May 31, 2022

➤ Registration 11:00 AM-5:00 PM

▽ NMDC Microbiome Metadata Standards Workshop 1:00 PM-3:00 PM

Garden Ballroom



Emily Davenport
Assistant Professor,
Department of
Biology
Penn State
University

Collecting, storing, and sharing data that are findable, accessible, interoperable, and reusable (FAIR) is an important challenge across microbiome fields. In this workshop hosted by the National Microbiome Data Collaborative (NMDC), participants will learn about metadata standards, how to plan data collection in their own projects to adhere to these standards, how to share their data with the community, and how to find and use publicly available microbiome data.

[Workshop Website](#)

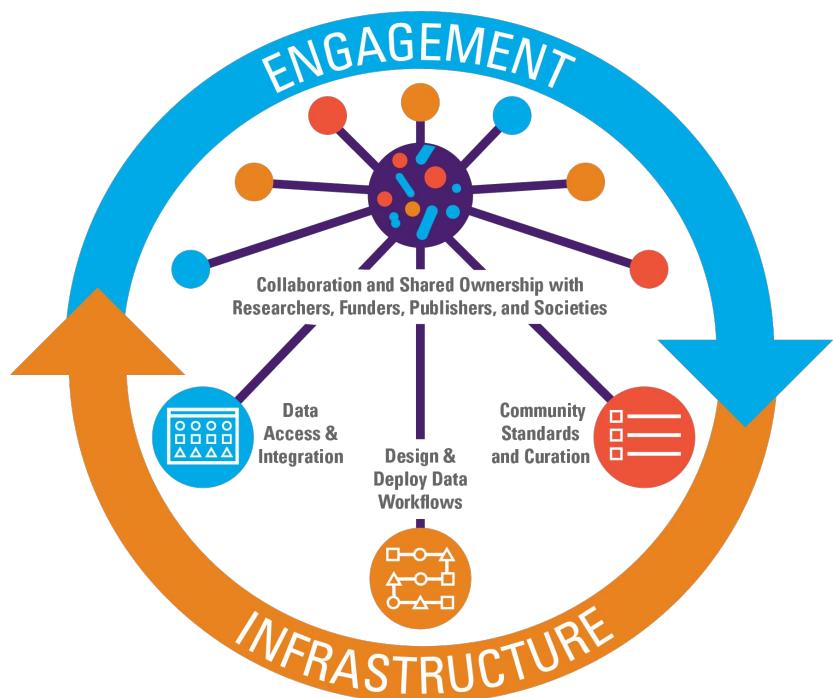


Who am I?

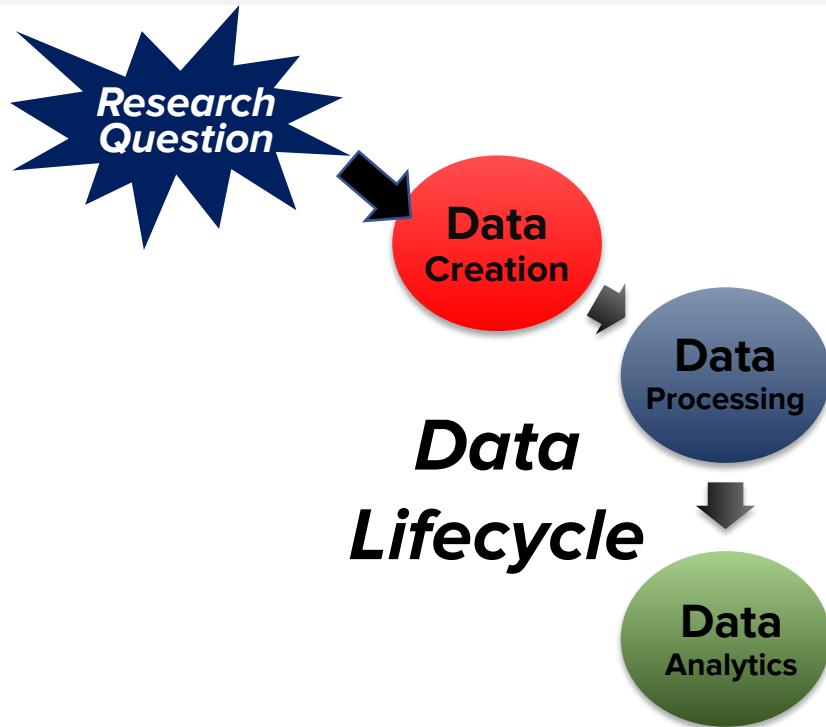


Overview of training

- Data Lifecycle and Metadata
- Data Management Best Practices
- The NMDC Mission
- Microbiome Metadata Community Standards
- Example Use Case – NMDC Metadata
- NMDC Data Portal



Data Lifecycle and Metadata



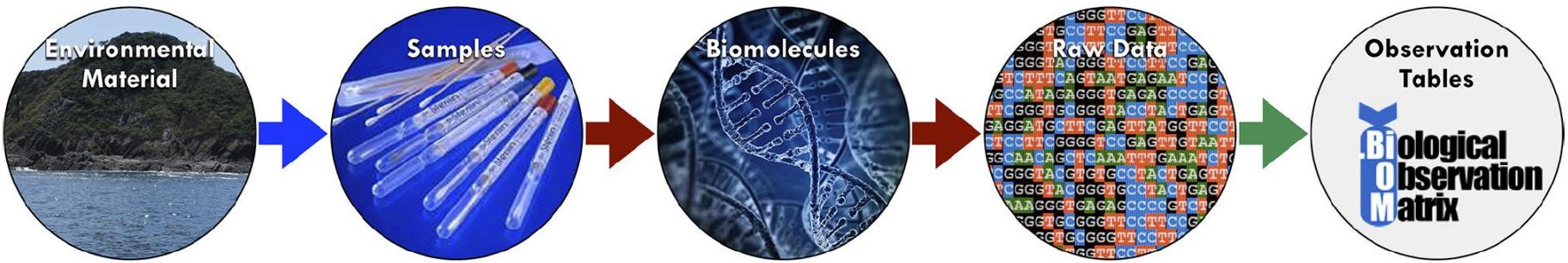
What are Metadata ?

Metadata are ...

- Contextual data about your data
- Vital for data
 - Preservation
 - Discovery
 - Access
 - Reuse



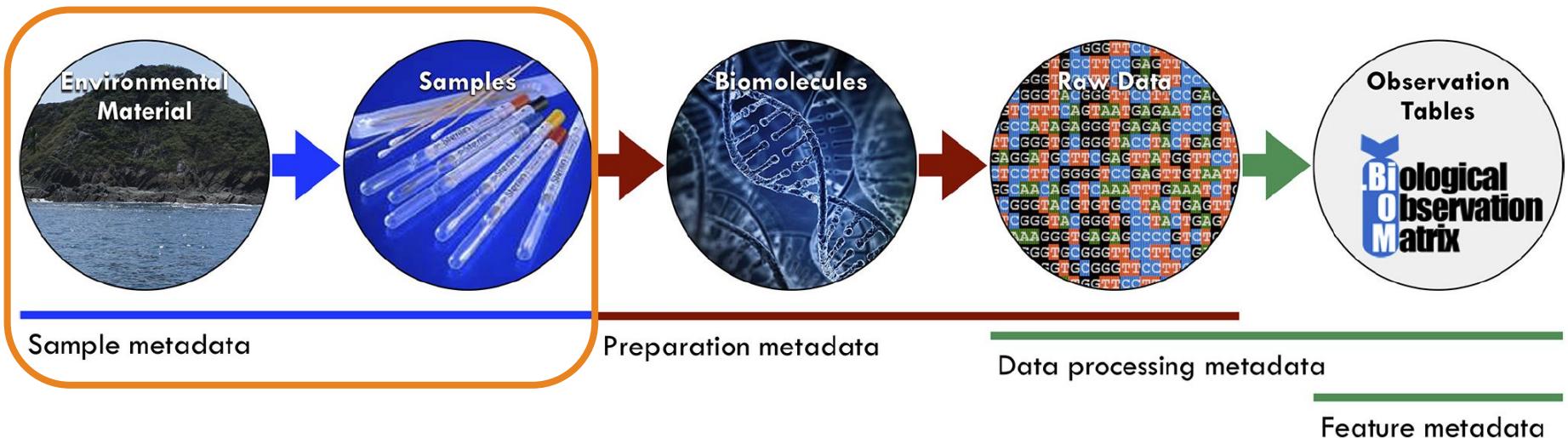
What are Metadata ?



From: **Introduction to Metadata and Ontologies:** Everything You Always Wanted to Know About Metadata and Ontologies (But Were Afraid to Ask) DOI: [10.25979/1607365](https://doi.org/10.25979/1607365)

What are Metadata ?

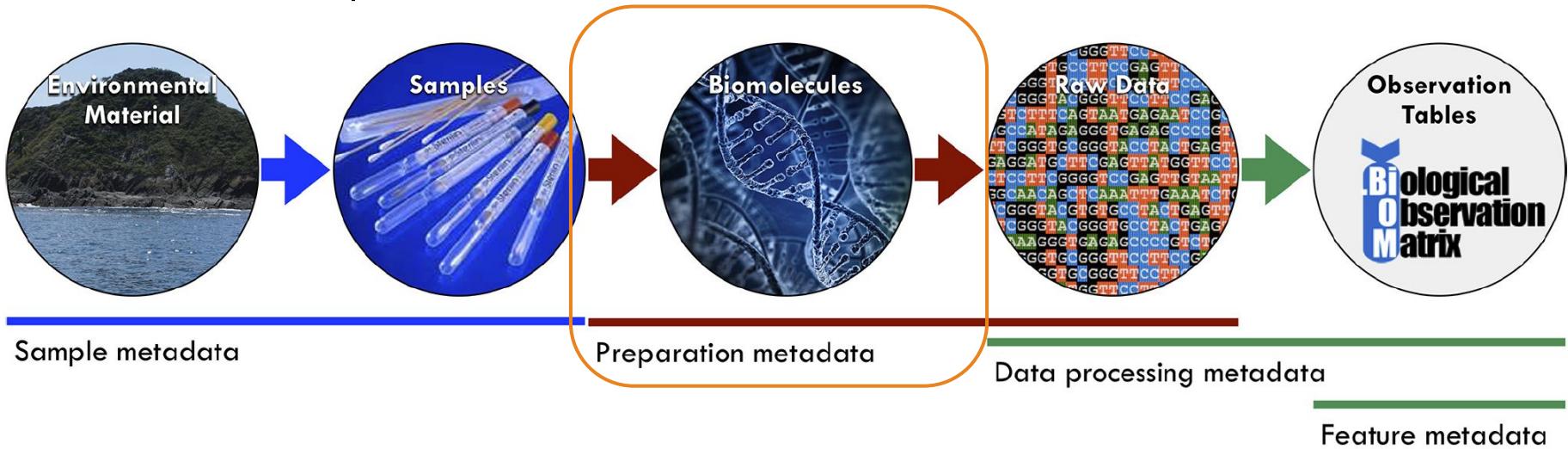
Microbiome Metadata - Sample metadata



From: **Introduction to Metadata and Ontologies**: Everything You Always Wanted to Know About Metadata and Ontologies (But Were Afraid to Ask) DOI: [10.25979/1607365](https://doi.org/10.25979/1607365)

What are Metadata ?

Not covered - Preparation metadata



Details on preparation and processing
should be made open and available
through online resources, like [Protocols.io](#)

What are Metadata ?

Sample metadata includes information about:

- **When** it was collected
- **Where** it was collected
- **What** kind of sample is it
- **Treatment** applied during experimentation
- **Environmental Properties** from which the sample was taken



What are Metadata ?

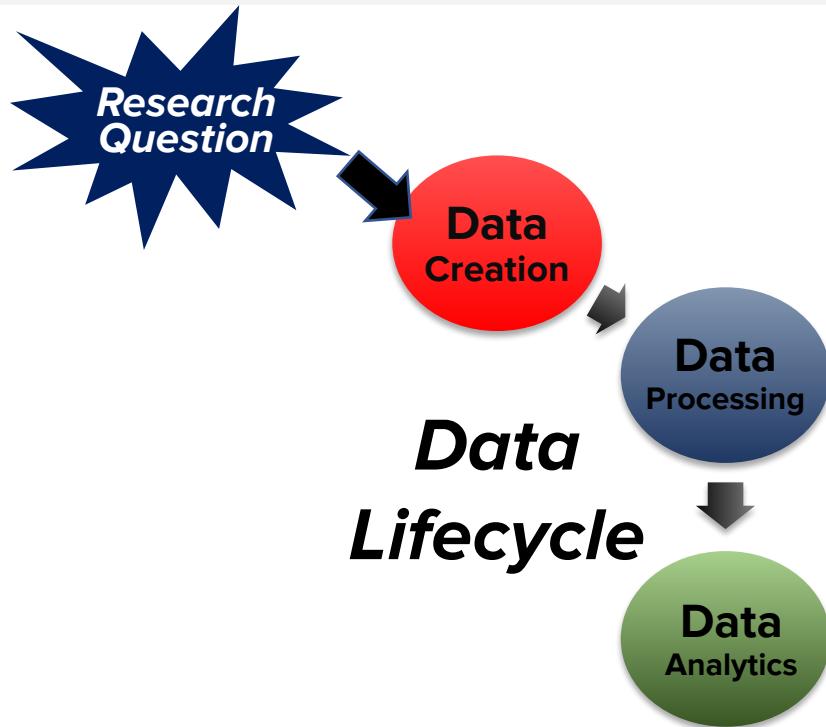


Sample metadata includes information about:

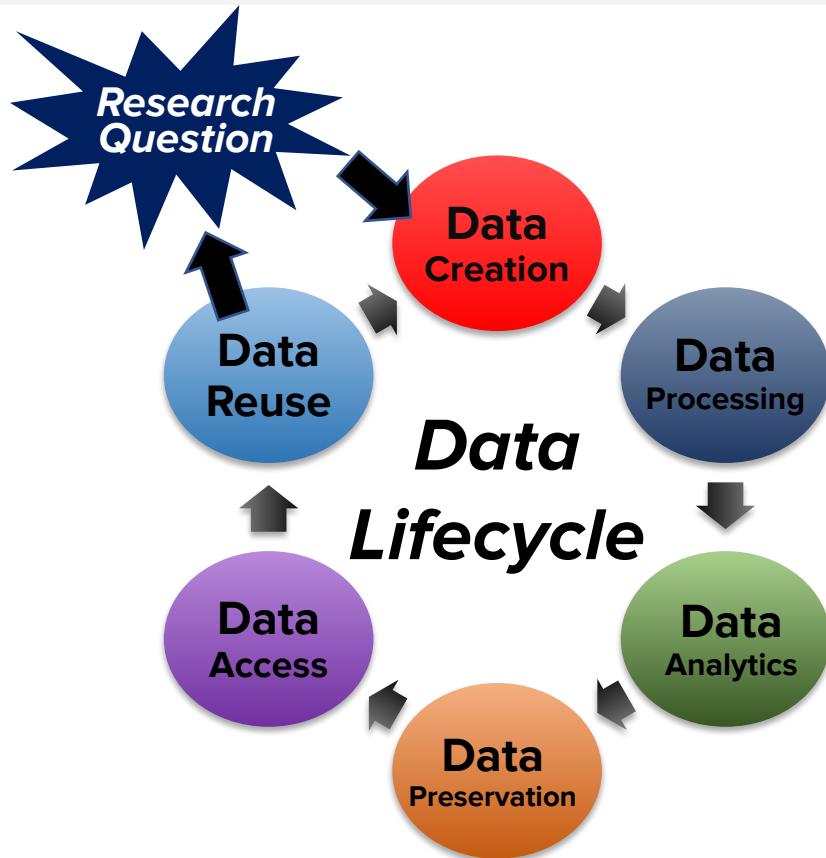
- **When** it was collected
- **Where** it was collected
- **What** kind of sample is it
- **Treatment** applied during experimentation
- **Environmental Properties** from which the sample was taken

What are some types of metadata that are relevant to your studies?

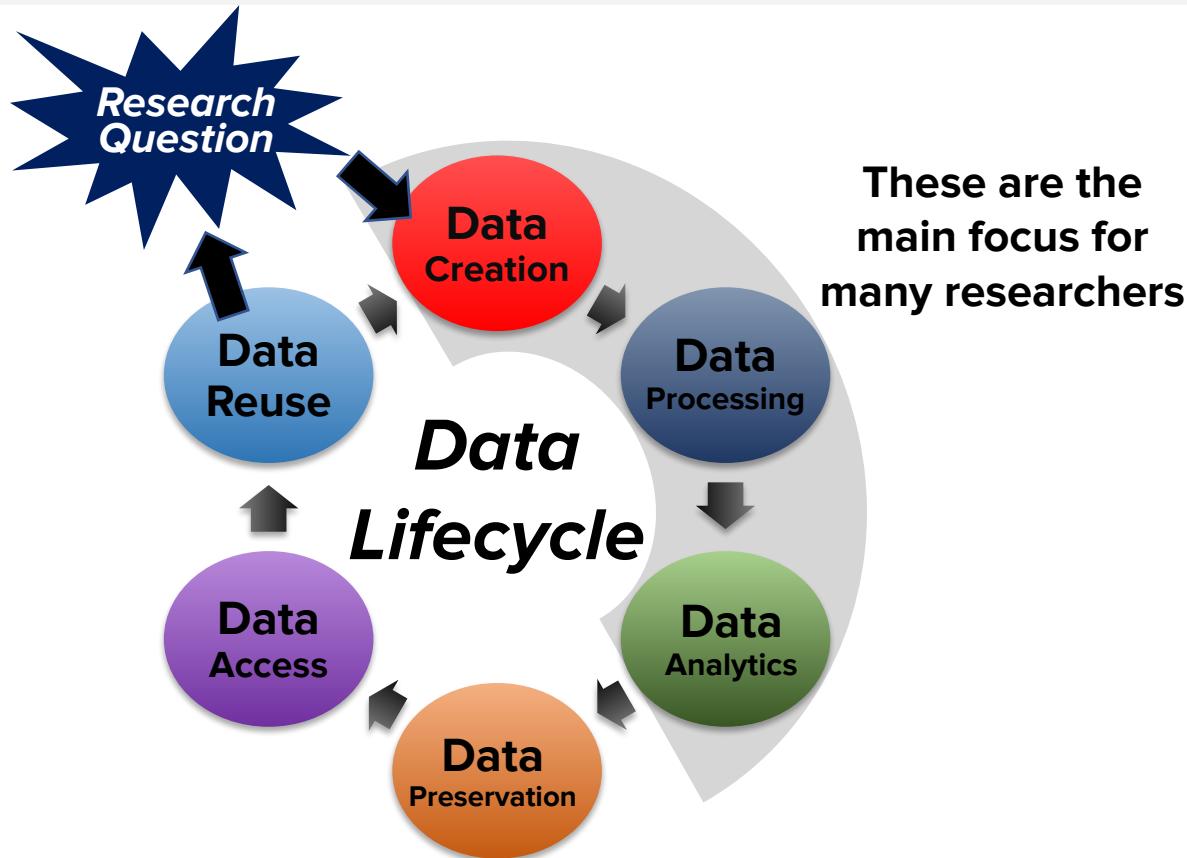
Data Lifecycle and Metadata



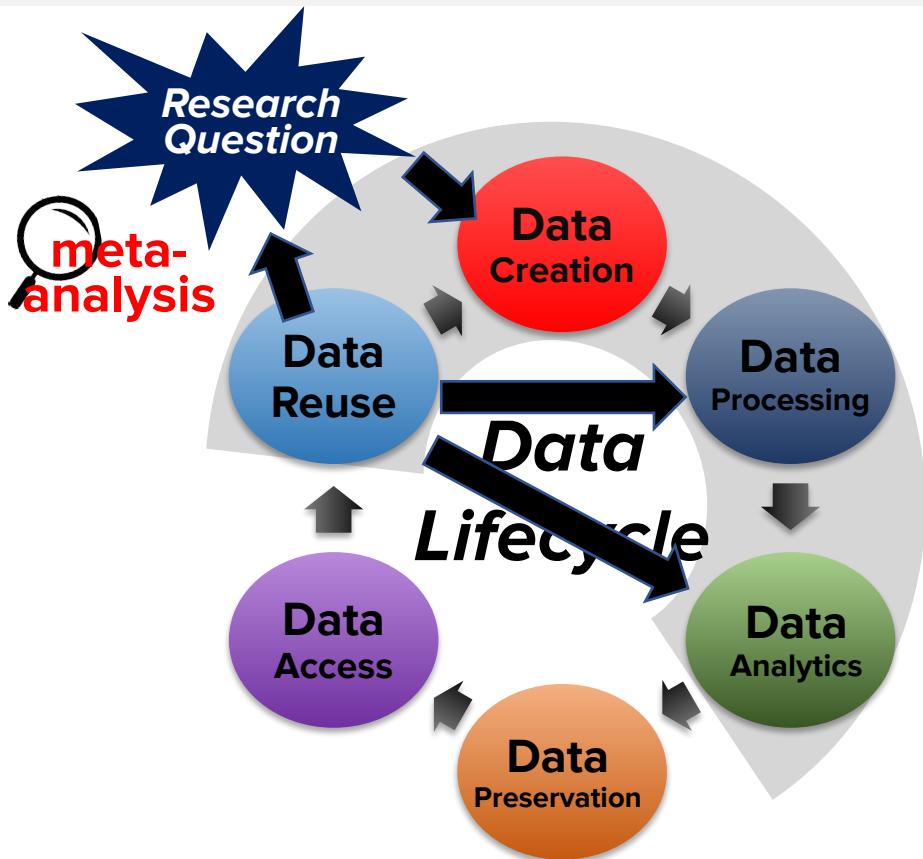
Data Lifecycle and Metadata



Data Lifecycle and Metadata



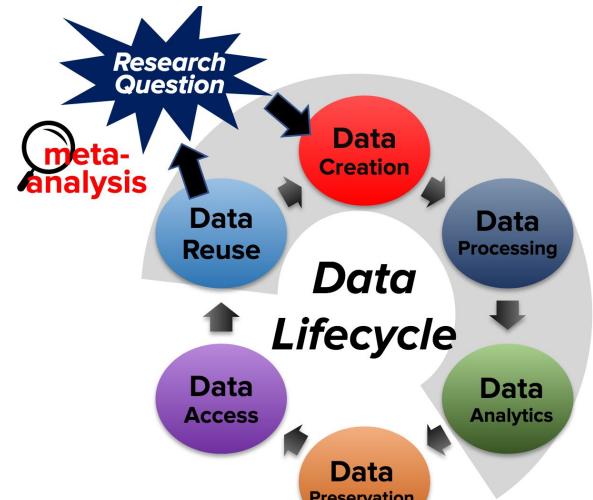
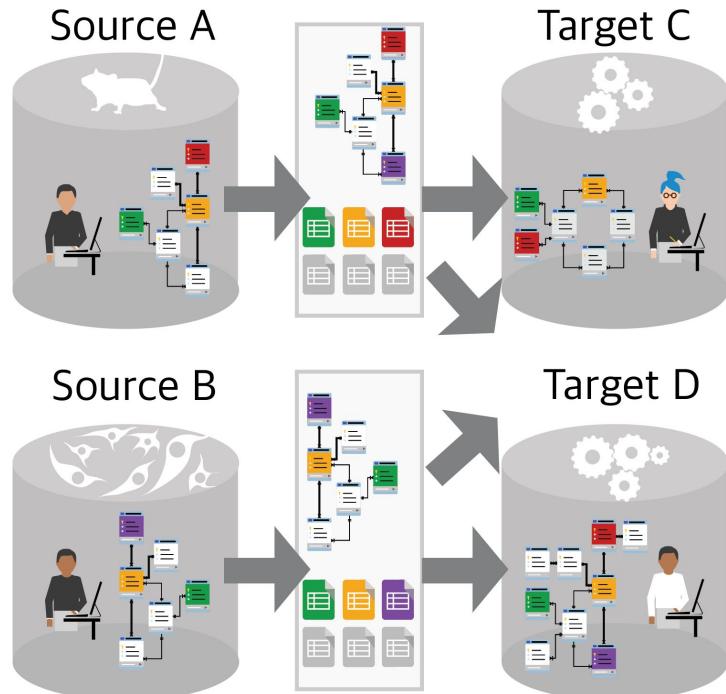
Data Lifecycle and Metadata



Data Lifecycle and Metadata

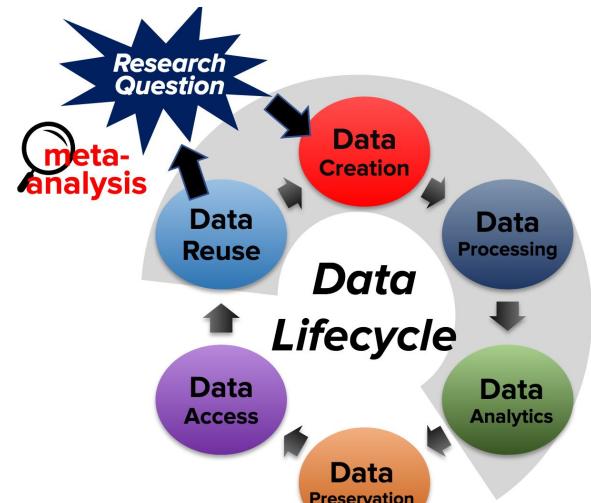
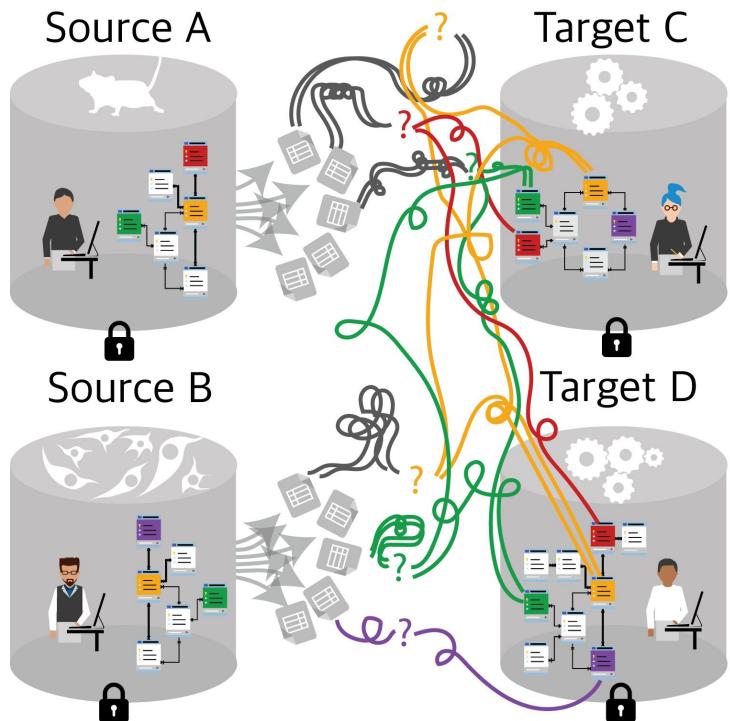


Data access and reuse - Goal



Data Lifecycle and Metadata

Data access and reuse - *Reality*

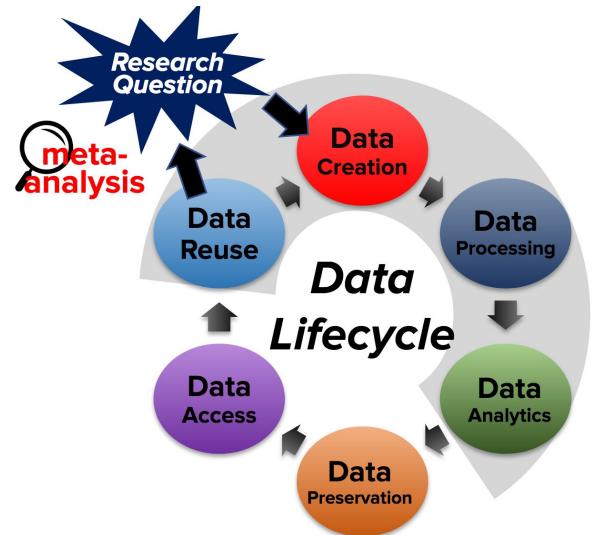


Data Lifecycle and Metadata

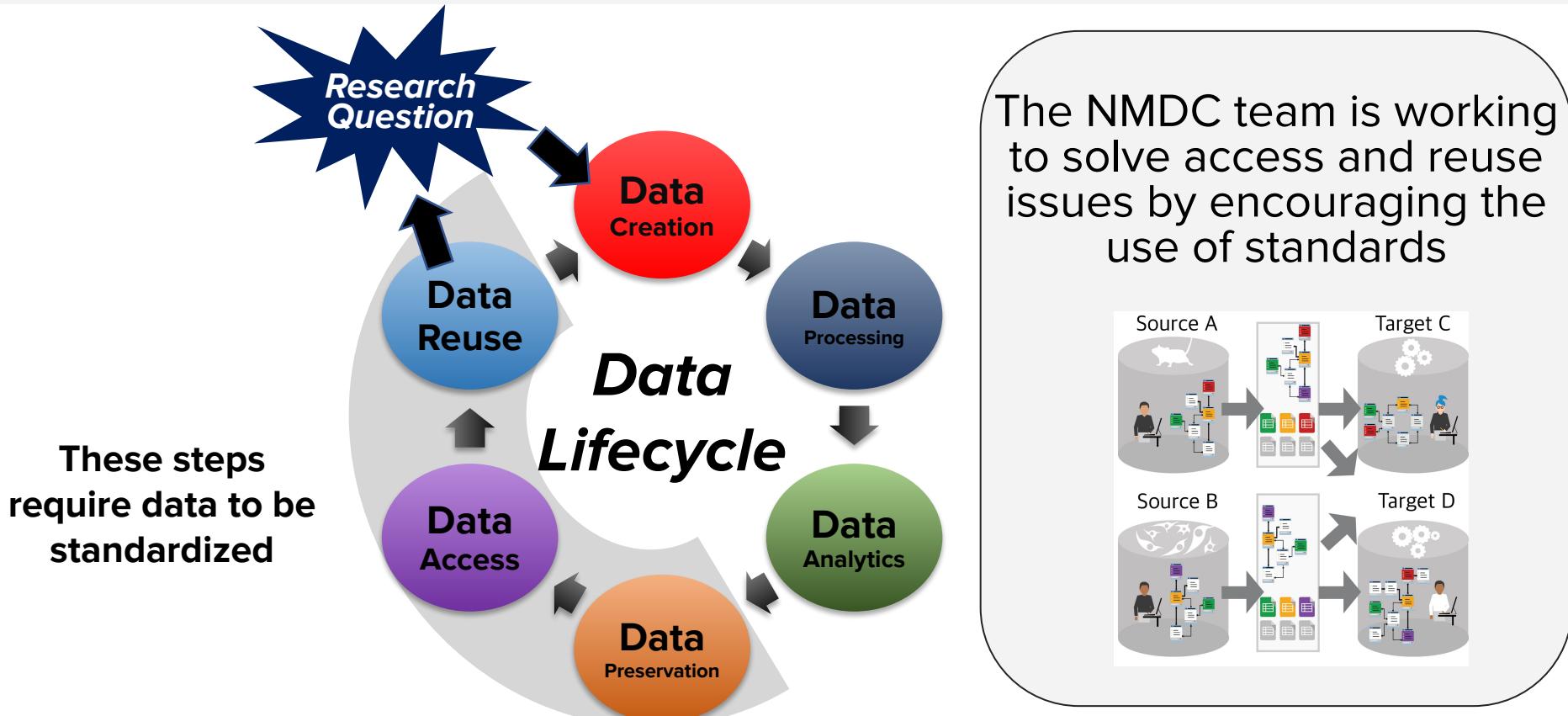


Data reuse

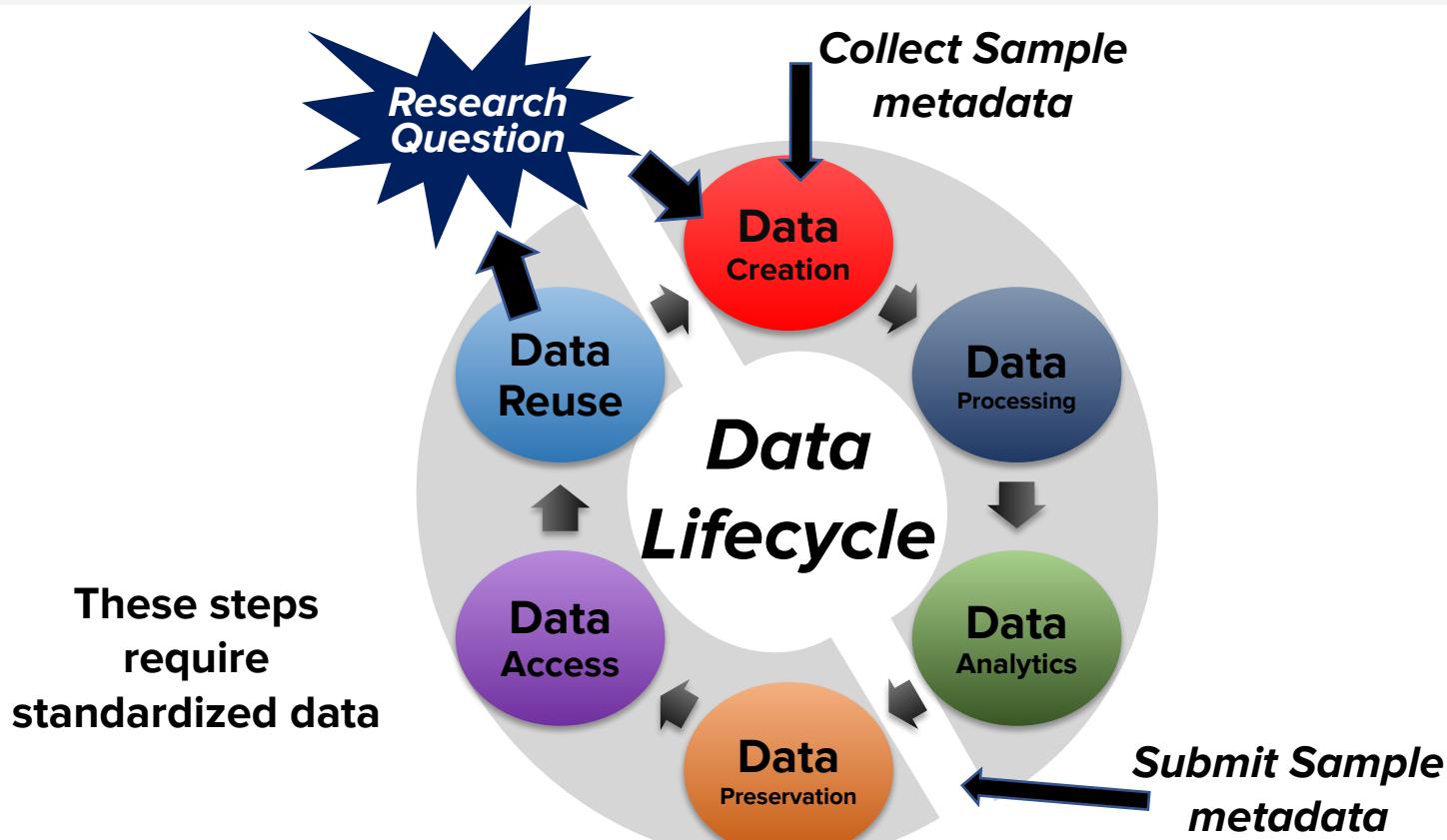
1. Have you reused microbiome datasets?
2. How was that process?



Data Lifecycle and Metadata



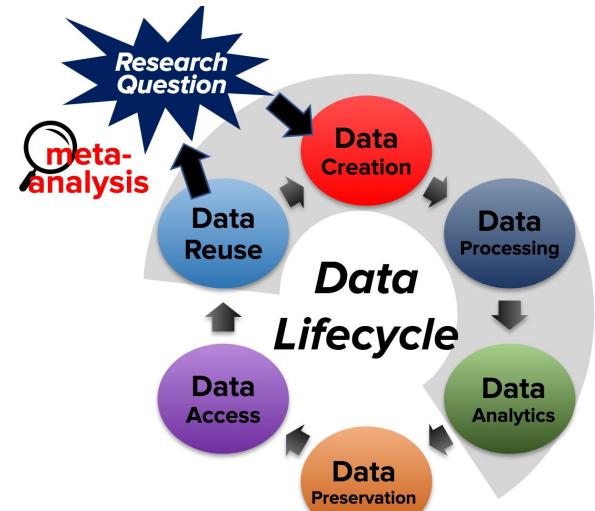
Data Lifecycle and Metadata



Data Lifecycle and Metadata

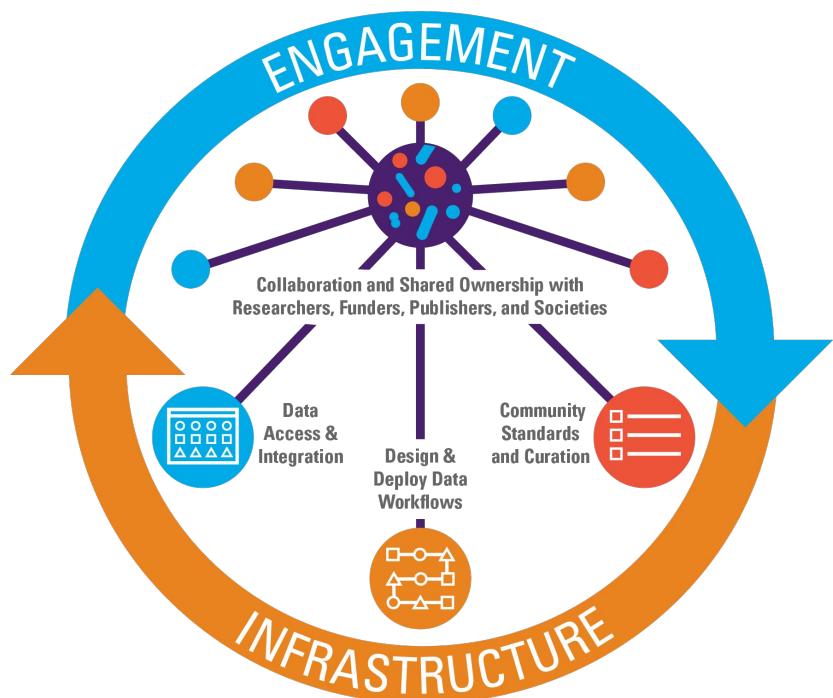
Considerations:

- During data collection:
 - be intentional about the process of collecting, storing, processing, and protecting data for research.
 - detail a data management plan for your samples
- During data processing:
 - organize data so that it can be easily shared
 - submit to relevant databases
- Outside of your own experiments:
 - Know what's out there
 - Know how to access it



Overview of training

- Data Lifecycle and Metadata
- Data Management Best Practices
- The NMDC Mission
- Microbiome Metadata Community Standards
- Example Use Case – NMDC Metadata
- NMDC Data Portal



Best Practices in Data Management



Benefits: Good data management allows for better science.

The effective reuse of data and integration of knowledge by the community after publication/release is *the key conduit leading to knowledge discovery and innovation.*

Goal: Enable great science through data sharing and reuse.

Best Practices in Data Management



For research: the intentional process of collecting, storing, processing, and protecting data

Best Practices in Data Management



For research: the *intentional* process of collecting, storing, processing, and protecting data

For data preservation: data are findable, accessible, interoperable, and reusable (FAIR) now and into the future. Good data management follows the **FAIR** data principles.

Wilkinson. et al. *The FAIR Guiding Principles for scientific data management and stewardship*. Sci Data **3**, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>

Best Practices in Data Management



● Findable

- Persistent ID
- Metadata online

● Accessible

- Data online
- Restrictions where needed

● Interoperable

- Use standards, controlled vocabularies
- Common (open) formats

● Reusable

- Rich documentation
- Clear usage license

How FAIR are your data?

Findable
It should be possible for others to discover your data. Rich metadata should be available online in a searchable resource, and the data should be assigned a persistent identifier.

- A persistent identifier is assigned to your data
- There are rich metadata, describing your data
- The metadata are online in a searchable resource e.g. a catalogue or data repository
- The metadata record specifies the persistent identifier

Accessible
It should be possible for humans and machines to gain access to your data, under specific conditions or restrictions where appropriate. FAIR does not mean that data need to be open! There should be metadata, even if the data aren't accessible.

- Following the persistent ID will take you to the data or associated metadata
- The protocol by which data can be retrieved follows recognised standards e.g. http
- The access procedure includes authentication and authorisation steps, if necessary
- Metadata are accessible, wherever possible, even if the data aren't

Interoperable
Data and metadata should conform to recognised formats and standards to allow them to be combined and exchanged.

- Data is provided in commonly understood and preferably open formats
- The metadata provided follows relevant standards
- Controlled vocabularies, keywords, thesauri or ontologies are used where possible
- Qualified references and links are provided to other related data

Reusable
Lots of documentation is needed to support data interpretation and reuse. The data should conform to community norms and be clearly licensed so others know what kinds of reuse are permitted.

- The data are accurate and well described with many relevant attributes
- The data have a clear and accessible data usage license
- It is clear how, why and by whom the data have been created and processed
- The data and metadata meet relevant domain standards

F A I R

'How FAIR are your data?' checklist, CC-BY by Sarah Jones & Marjan Grootveld, EUDAT; Image CC-BY-SA by SangyaHundir

How FAIR are your data?

https://zenodo.org/record/1065991#.YnQYu_PMIUE

Best Practices in Data Management



Benefits:

- Data are discoverable
- Data are accessible
- Data are comparable
- Data are persistent (e.g., DOI)

Impact:

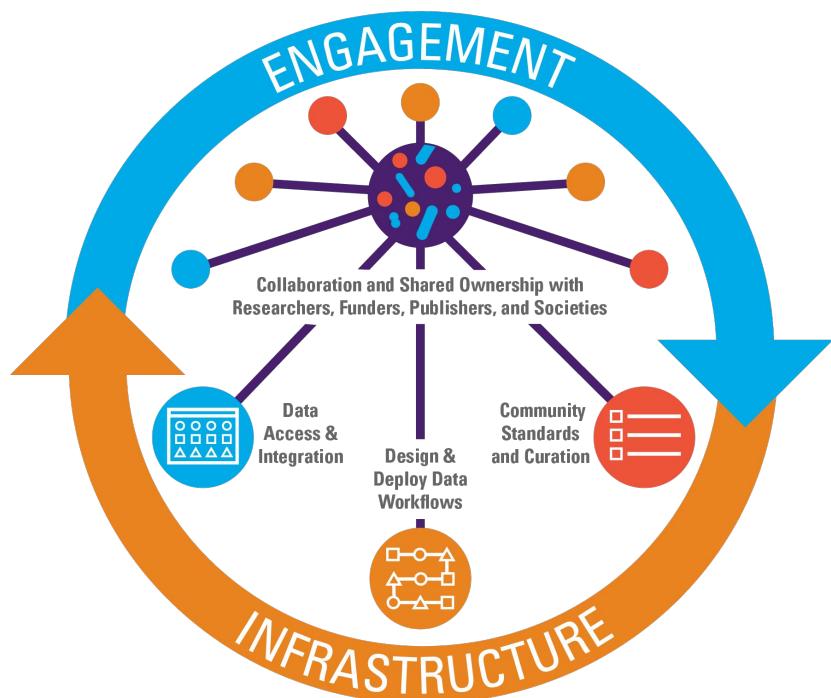
- Increased scientific exposure
- Work is appropriately credited
- Impact is measurable / quantifiable



From: Roche DG, Lanfear R, Binning SA, Haff TM, Schwanz LE, Cain KE, et al. (2014) Troubleshooting Public Data Archiving: Suggestions to Increase Participation. PLoS Biol 12(1): e1001779.
<https://doi.org/10.1371/journal.pbio.1001779>

Overview of training

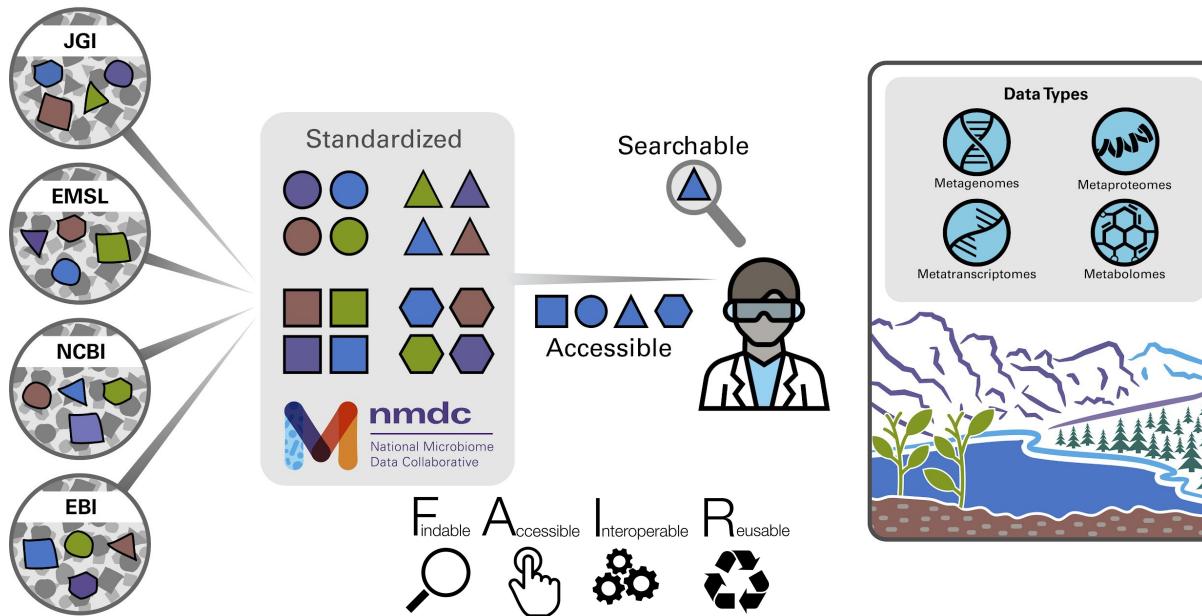
- Data Lifecycle and Metadata
- Data Management Best Practices
- The NMDC Mission
- Microbiome Metadata Community Standards
- Example Use Case – NMDC Metadata
- NMDC Data Portal



The NMDC Mission

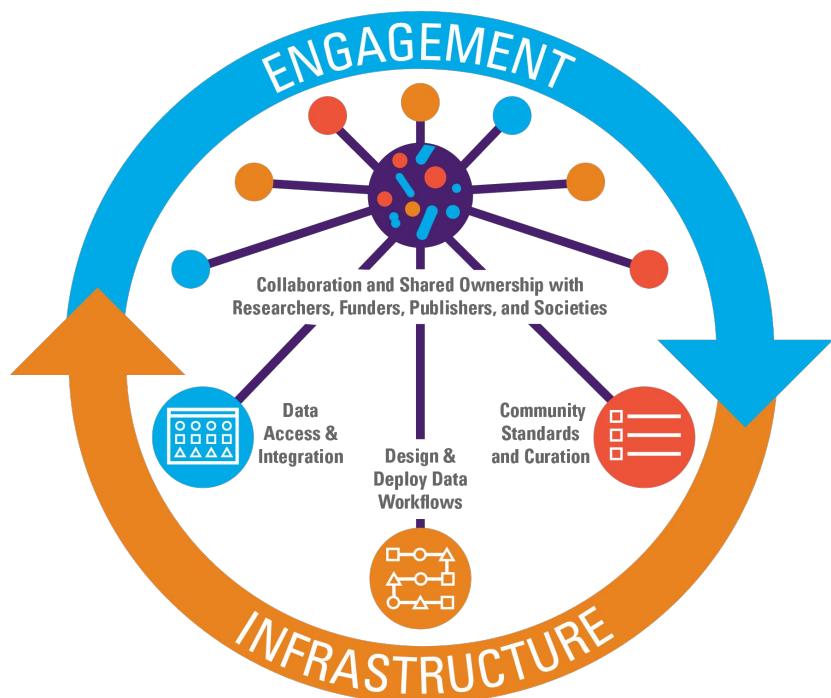


Provide a gateway to FAIR multi-omics microbiome data by leveraging best practices for data curation and processing



Overview of training

- Data Lifecycle and Metadata
- Data Management Best Practices
- The NMDC Mission
- Microbiome Metadata Community Standards
- Example Use Case – NMDC Metadata
- NMDC Data Portal



Why metadata standards?

Similar types of data can be difficult to reuse if they lack consistent language and formatting



| idNumber | material | temperature |
|-----------------|-----------------|--------------------|
| 3928 | soil | 23.2 °C |
| 3234 | groundwater | 9.02 °C |

| sampleNum | substance | temp |
|------------------|------------------|-------------|
| 8765 | dirt | 21.1 |
| 2312 | ground liquid | 7.0 |

Why metadata standards?

Adopting standards for reporting makes data human and machine readable.



| idNumber | material | temperature |
|-----------------|-----------------|--------------------|
| 3928 | soil | 23.2 °C |
| 3234 | groundwater | 9.02 °C |



| idNumber | material | temperature |
|-----------------|-----------------|--------------------|
| 8765 | soil | 21.1 °C |
| 2312 | groundwater | 7.0 °C |

Metadata standards support FAIR

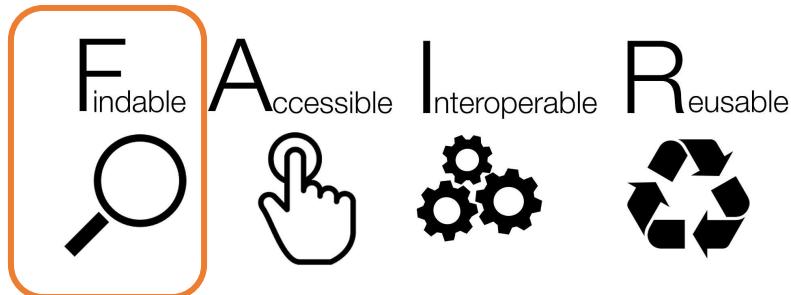
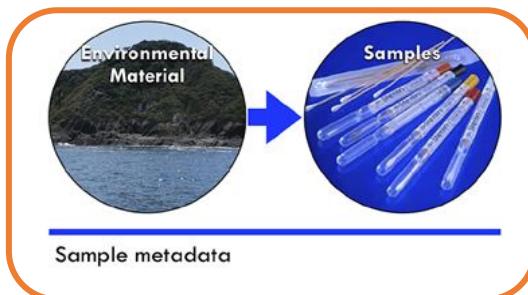


Standardized sample metadata

support search and discovery based on:

- **When** it was collected,
- **Where** it was collected,
- **What** kind of sample is it
- **Properties of the environment**

from which the sample was taken



Community Metadata Standards



**The NMDC metadata standards utilize and enhance
existing community-driven standards**



1. MIxS: Minimum Information about any (x) Sequence
Genomic Standards Consortium (GSC)



2. [GOLD: Genomes OnLine Database](#)
Joint Genome Institute (JGI)



3. [EnvO: Environment Ontology](#)
Open Biological and Biomedical Ontology (OBO) Foundry

The NMDC + Community Standards



NMDC – derive mandatory metadata from community standards

MixS

- Sample Identifiers : *sample_name* and *ID* (globally unique ID)
- Growth Facility : *growth_facil*
- Geographic Location (latitude longitude) : *lat_lon*
- Geographic Region (country and/or sea, region) : *geo_loc_name*
- Collection Date : *collection_date*
- Sample Material Processing : *samp_mat_process*
- Storage Conditions & Temperature : *store_cond* & *samp_store_temp*
- Broad-scale Environmental Context : *env_broad_scale*

EnvO

- Local Environmental Context : *env_local_scale*
- Environmental Medium : *env_medium*
- GOLD Environment Path (unique to the NMDC) : *gold_ecosystem*

GOLD

Genomic Standards Consortium



- **MIxS: Environmental Packages**
- System of unified standards for describing microbiome samples from diverse environments (e.g., plant, soil, sediment, water)
- Each package includes a list of ***metadata descriptors*** for each parameter/field
- NMDC leverages MIxS **Mandatory Metadata Fields** that are shared across all packages

MIxS Environmental Packages



17 packages (*currently available in NMDC)

New packages added with community input



MIxS
EnvO
GOLD

| MixS Environmental Packages | |
|-----------------------------|---|
| air | built environment |
| host-associated | human-associated |
| human-gut | human-oral |
| human-skin | human-vaginal |
| hydrocarbon resources-cores | hydrocarbon resources-fluids/swabs |
| microbial mat/biofilm | miscellaneous natural or artificial environment |
| *plant associated | *sediment |
| *soil | wastewater/sludge |
| *water | Pending: Ag, food, parasite, etc. |

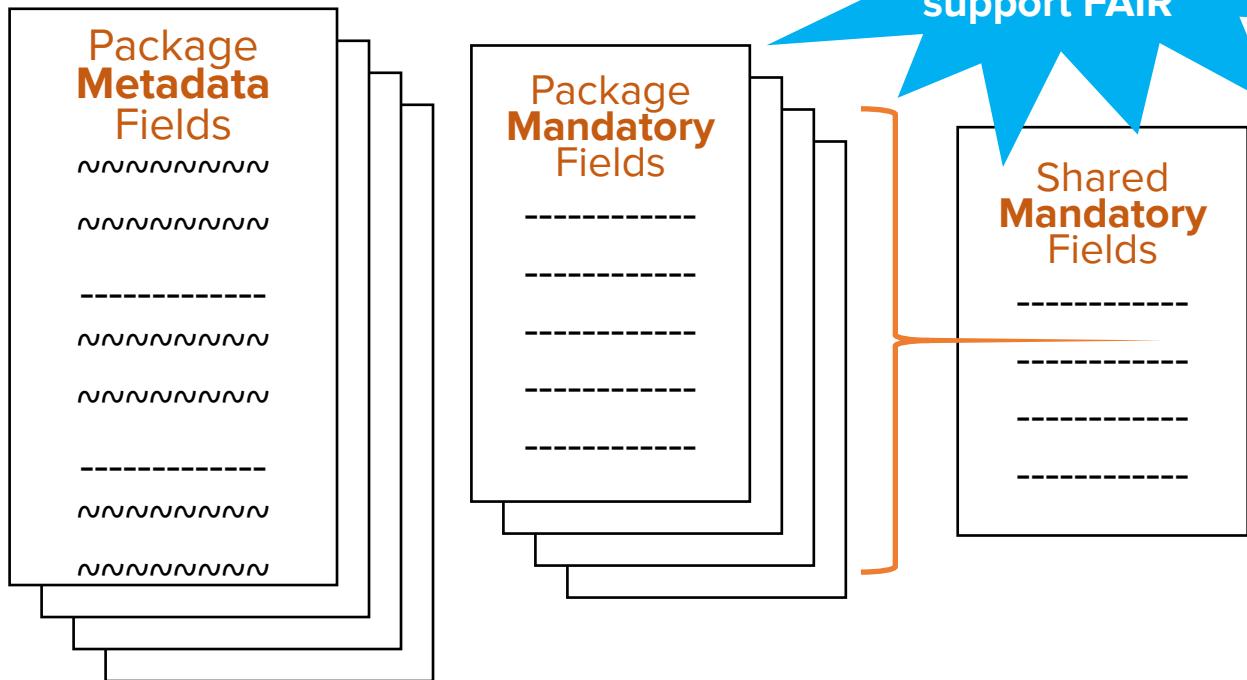
MIxS Mandatory Fields



Mandatory metadata fields are shared across all MIxS environmental packages and checklists



e.g., *plant*, *soil*,
sediment, *water*, ...



MIxS Checklists



GSC supports 5 data types as
“Minimum information about [checklist] sequence”



| MIxS Checklists |
|--|
| Genome (MIGS) |
| Metagenome (MIMS) |
| MARKer gene (MIMARKS) |
| Single Amplified Genome (MISAG) |
| Metagenome Amplified Genome (MIMAG) |

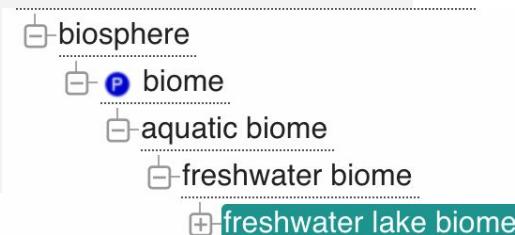


The NMDC supports
Metagenome
checklists (MIMS)

OBO EnvO (Environment Ontology)



- Dynamic, community resource
- Hierarchical classification of samples by environment
- Mandated by MIxS for environment fields
- Mapped to multiple environmental vocabularies (e.g., SWEET; Semantic Web Earth Environmental Sciences)
- **Community:**
 - Biosciences, Earth, environment



MIxS

EnvO

GOLD

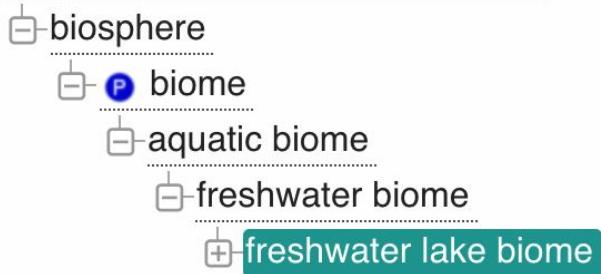
MIxS requires environment terms



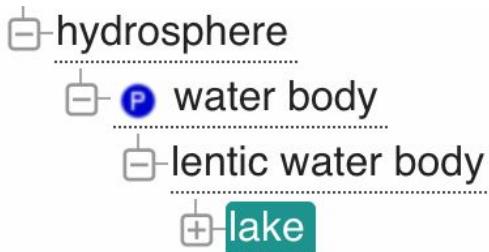
MIxS descriptors specify the ***sample environment*** with the Environment Ontology (EnvO)



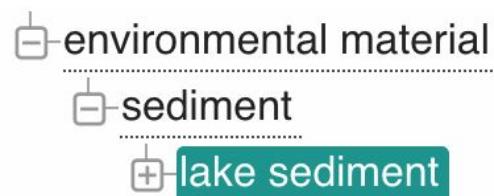
broad-scale environmental context



local-scale environmental context



environmental medium



MIxS

EnvO

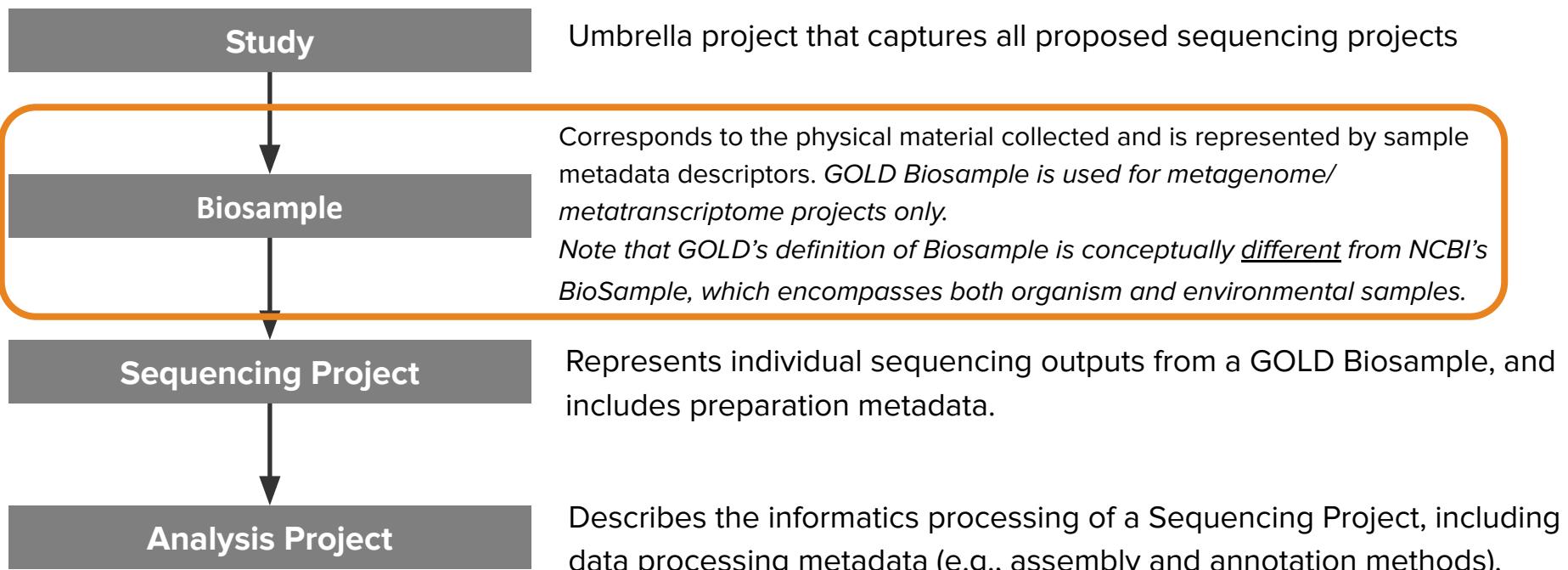
GOLD

The Genomes OnLine Database (GOLD)

- Manually curated, daily updated collection of genome projects and their metadata accumulated from around the world
- Metadata fields: ~600
- Controlled Vocabulary (CV) fields: 76 (3,873 terms)
- Currently contains > 130,000 microbiome biosamples



GOLD - Overview



GOLD - Five-level ecosystem path

GOLD Ecosystem classification

Ecosystem

Ecosystem Category

Ecosystem Type

Ecosystem Subtype

Specific Ecosystem

Example: Lake Sediment

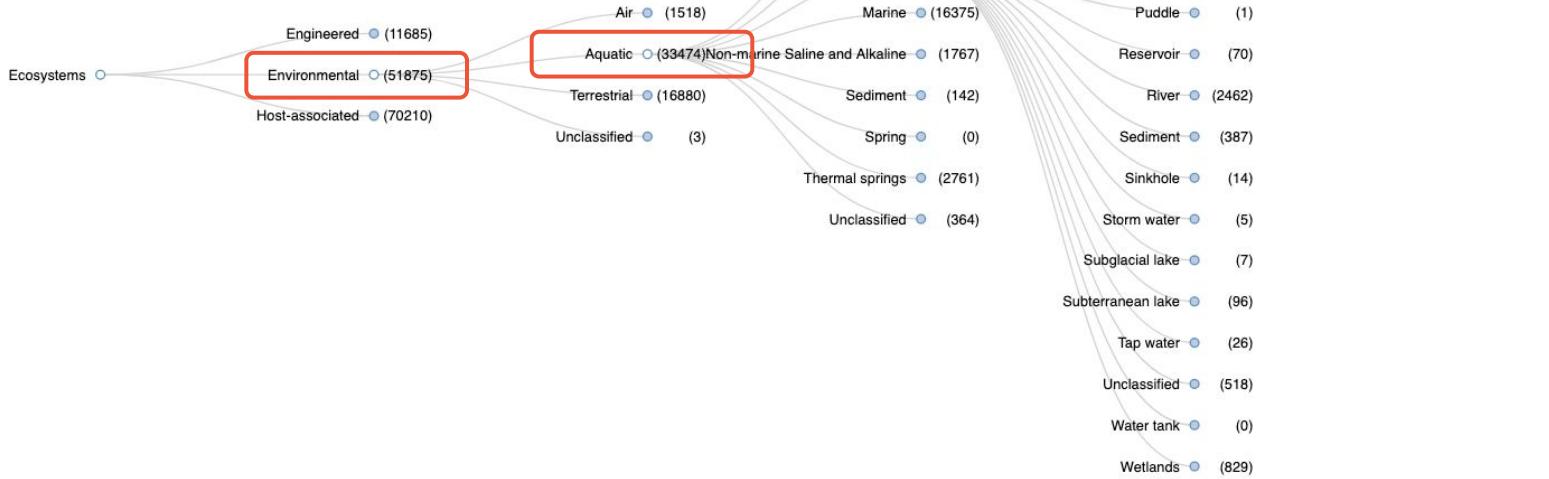
Environmental

Aquatic

Freshwater

Lake

Sediment



GOLD - Five-level ecosystem path



Activity: Explore the GOLD ecosystem tree viewer to identify your sample metadata path

<https://gold.jgi.doe.gov/ecosystemtree>

The NMDC + Community Standards



| MIxS / EnvO triad | | | GOLD Ecosystem classification | | | | |
|-------------------------|-------------------------|----------------------|-------------------------------|--------------------|----------------|--------------------|----------------|
| Broad-scale environment | Local-scale environment | Environmental Medium | Ecosystem | Ecosystem Category | Ecosystem Type | Specific Ecosystem | Ecosystem Tree |
| Freshwater lake biome | Lake Shore | Sediment | Environment | Aquatic | Freshwater | Lake | Sediment |
| Freshwater lake biome | Lake | Algal bloom | Environment | Aquatic | Freshwater | Lake | Algal bloom |

MIxS

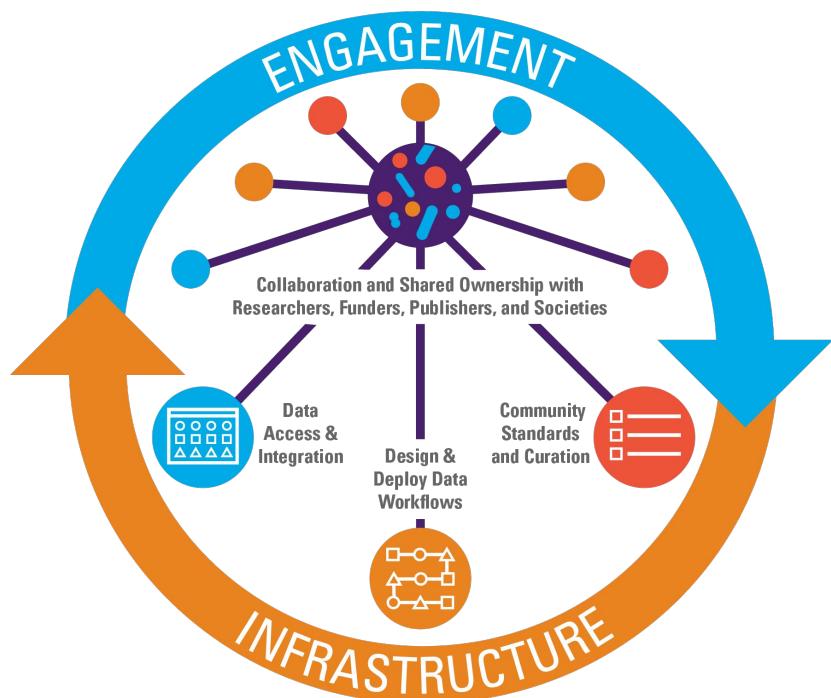
EnvO

GOLD

EnvO & GOLD terms together gives us improved environmental context to the microbiome biosamples!

Overview of training

- Data Lifecycle and Metadata
- Data Management Best Practices
- The NMDC Mission
- Microbiome Metadata Community Standards
- Example Use Case – NMDC Metadata
- NMDC Data Portal



Example: Human Microbiome MixS



Example: Human Microbiome MixS



nmdc

National Microbiome
Data Collaborative

<https://github.com/GenomicsStandardsConsortium/mixs/tree/main/release/excel>

A screenshot of a GitHub repository page for 'GenomicsStandardsConsortium / mixs'. The repository is public and has 16 watchers, 12 forks, and 14 stars. The main branch is 'main'. A list of files in the 'mixs / release / excel' directory is shown, all added by user 'ramonawalls' on March 23, 2023. The files include various microbiome samples: agriculture_mixs6.xlsx, air_mixs6.xlsx, built-environment_mixs6.xlsx, food-animal-and-feed_mixs6.xlsx, food-farm-environment_mixs6.xlsx, food-human-foods_mixs6.xlsx, food-production-facility_mixs6.xlsx, host-associated_mixs6.xlsx, human-associated_mixs6.xlsx, human-gut_mixs6.xlsx, human-oral_mixs6.xlsx, human-skin_mixs6.xlsx, and human-vaginal_mixs6.xlsx. An arrow points to the 'human-gut_mixs6.xlsx' file.

| File | Description | Last Commit |
|-------------------------------------|-----------------------------------|--------------|
| agriculture_mixs6.xlsx | add manually created excdel files | 2 months ago |
| air_mixs6.xlsx | add manually created excdel files | 2 months ago |
| built-environment_mixs6.xlsx | add manually created excdel files | 2 months ago |
| food-animal-and-feed_mixs6.xlsx | add manually created excdel files | 2 months ago |
| food-farm-environment_mixs6.xlsx | add manually created excdel files | 2 months ago |
| food-human-foods_mixs6.xlsx | add manually created excdel files | 2 months ago |
| food-production-facility_mixs6.xlsx | add manually created excdel files | 2 months ago |
| host-associated_mixs6.xlsx | add manually created excdel files | 2 months ago |
| human-associated_mixs6.xlsx | add manually created excdel files | 2 months ago |
| human-gut_mixs6.xlsx | add manually created excdel files | 2 months ago |
| human-oral_mixs6.xlsx | add manually created excdel files | 2 months ago |
| human-skin_mixs6.xlsx | add manually created excdel files | 2 months ago |
| human-vaginal_mixs6.xlsx | add manually created excdel files | 2 months ago |

Example: Human Microbiome MixS



<https://github.com/GenomicsStandardsConsortium/mixs/tree/main/release/excel>

A screenshot of a GitHub repository page for 'GenomicsStandardsConsortium / mixs'. The page shows a single file, 'human-gut_mixs6.xlsx', which is 77.1 KB in size and was added by 'ramonawalls'. A red circle highlights the 'Download' button at the bottom right of the file preview card.

Search or jump to... Pull requests Issues Marketplace Explore

GenomicsStandardsConsortium / mixs Public

Code Issues 222 Pull requests 1 Actions Projects 5 Wiki Security Insights

main mixs / release / excel / human-gut_mixs6.xlsx Go to file ...

ramonawalls add manually created excdel files Latest commit 04e2472 on Mar 23 History

1 contributor

77.1 KB Download View raw



© 2022 GitHub, Inc.

Terms

Privacy

Security

Status

Docs

Contact GitHub

Pricing

API

Training

Blog

About

Example: Human Microbiome MixS



Screenshot of Microsoft Excel showing a spreadsheet titled "human-gut_mixs6 (2)". The spreadsheet contains a table with columns for Environmental package, Structured comment name, Package item, Definition, Expected value, Value syntax, Example, Requirement, Preferred unit, Occurrence, and MIXS ID.

Table Headers:

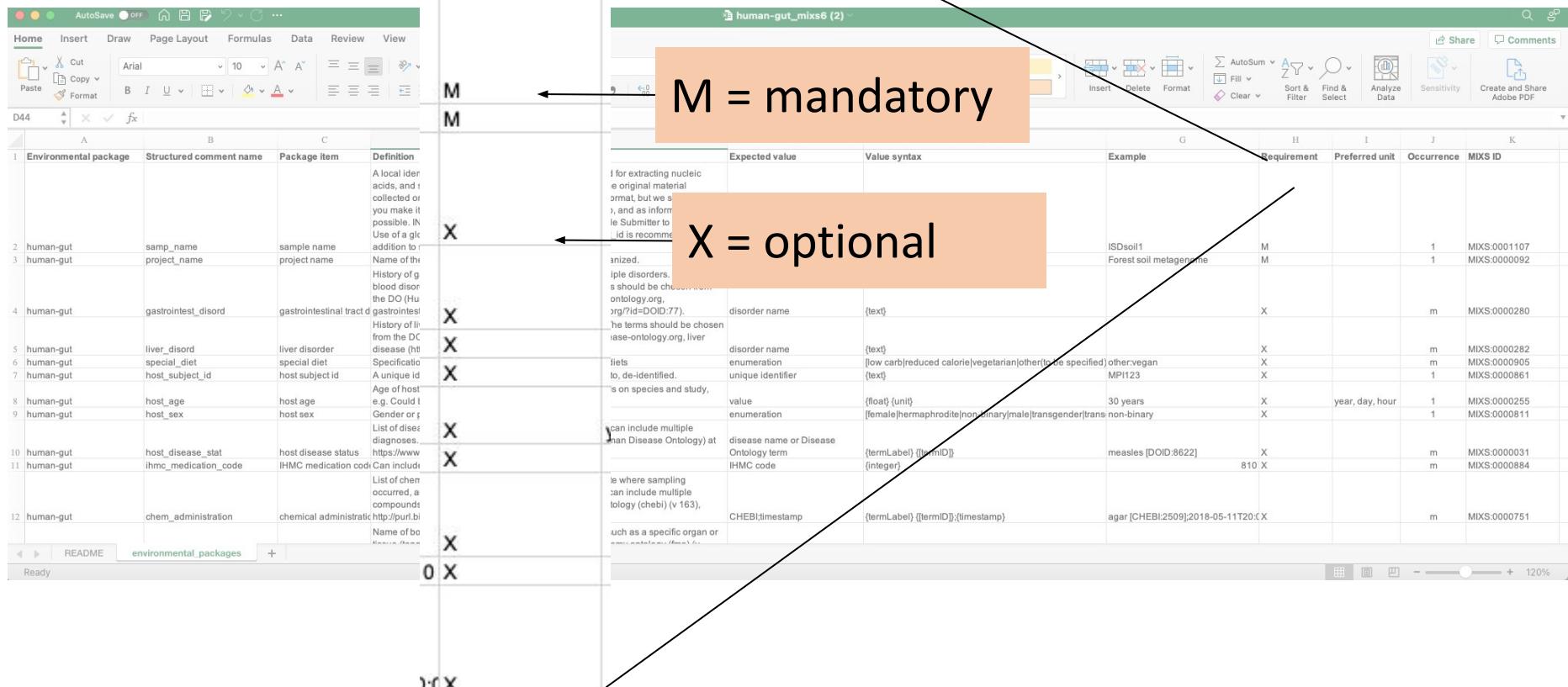
- A: Environmental package
- B: Structured comment name
- C: Package item
- D: Definition
- E: Expected value
- F: Value syntax
- G: Example
- H: Requirement
- I: Preferred unit
- J: Occurrence
- K: MIXS ID

Data Rows:

- Row 1: Environmental package, Structured comment name, Package item, Definition, Expected value, Value syntax, Example, Requirement, Preferred unit, Occurrence, MIXS ID
- Row 2: human-gut, samp_name, sample name, A local identifier or name that for the material sample used for extracting nucleic acids, and subsequent sequencing. It can refer either to the original material collector or to any derived sub-samples. It can have any format, but we suggest that you make it concise, unique and consistent within your lab, and as informative as possible. INSDC requires every sample name from a single Submitter to be unique. Use of a globally unique identifier for the field source_mat_id is recommended in addition to sample_name., text, {text}, ISDsoil1, M, 1, MIXS:0001107
- Row 3: human-gut, project_name, project name, Name of the project within which the sequencing was organized., text, {text}, Forest soil metagenome, M, 1, MIXS:000092
- Row 4: human-gut, gastrointest_disord, gastrointestinal tract disorder, History of gastrointestinal tract disorders; can include multiple disorders. History of blood disorders; can include multiple disorders. The terms should be chosen from the DO (Human Disease Ontology) at <http://www.disease-ontology.org>., disorder name, {text}, X, m, MIXS:0000280
- Row 5: human-gut, liver_disord, liver disorder, History of liver disorders; can include multiple disorders. The terms should be chosen from the DO (Human Disease Ontology) at <http://www.disease-ontology.org>, disorder name, {text}, X, m, MIXS:0000282
- Row 6: human-gut, special_diet, special diet, Specification of special diet; can include multiple special diets., enumeration, {low carb|reduced calorie|vegetarian|other(to be specified)}, X, m, MIXS:0000905
- Row 7: human-gut, host_subject_id, host subject id, A unique identifier by which each subject can be referred to, de-identified., unique identifier, {text}, MPI123, X, 1, MIXS:0000861
- Row 8: human-gut, host_age, host age, Age of host at the time of sampling; relevant scale depends on species and study, e.g. Could be seconds for amoebae or centuries for trees., value, {float} {unit}, 30 years, year, day, hour, 1, MIXS:0000255
- Row 9: human-gut, host_sex, host sex, Gender or physical sex of the host., enumeration, {female|hermaphrodite|non-binary|male|transgender|trans non-binary}, X, 1, MIXS:0000811
- Row 10: human-gut, host_disease_stat, host disease status, List of diseases with which the host has been diagnosed; can include multiple diagnoses. The terms should be chosen from the DO (Human Disease Ontology) at <https://www.disease-ontology.org>, disease name or Disease Ontology term, {termLabel} {[termID]}, measles [DOID:8622], X, m, MIXS:0000031
- Row 11: human-gut, ihmc_medication_code, IHMC medication code, List of chemical compounds administered to the host or site where sampling occurred, and when (e.g. Antibiotics, n fertilizer, air filter); can include multiple compounds. For chemical entities of biological interest ontology (chebi) (v 163), IHMC code, {integer}, 810 X, m, MIXS:0000884
- Row 12: human-gut, chem_administration, chemical administration, chemical administrative <http://purl.bioontology.org/ontology/chebi>, CHEBI:timestamp, {termLabel} {[termID]};{timestamp}, agar [CHEBI:2509];2018-05-11T20:X, X, m, MIXS:0000751

Bottom navigation bar: README, environmental_packages, +

Example: Human Gut Microbiome MIXS



The table illustrates the structure of the MIXS spreadsheet, showing columns for Environmental package, Structured comment name, Package item, Definition, Expected value, Value syntax, Example, Requirement, Preferred unit, Occurrence, and MIXS ID.

Mandatory Fields (Column H):

- Environmental package
- Structured comment name
- Package item
- Definition
- Expected value
- Value syntax
- Example
- Requirement
- Preferred unit
- Occurrence
- MIXS ID

Optional Fields (Column I):

- Host subject id
- Host sex
- Host age
- Gastrointestinal tract disorder
- Project name
- Samp name
- Special diet
- Liver disorder
- Disease Ontology term
- Chemical administration
- Chemical code
- Chemical timestamp
- Chemical name
- Chemical source
- Chemical quantity
- Chemical unit
- Chemical location
- Chemical context
- Chemical properties
- Chemical annotations
- Chemical references
- Chemical notes
- Chemical comments
- Chemical requirements
- Chemical occurrence
- Chemical MIXS ID

Example: Human Microbiome MIxS



Mandatory Fields:

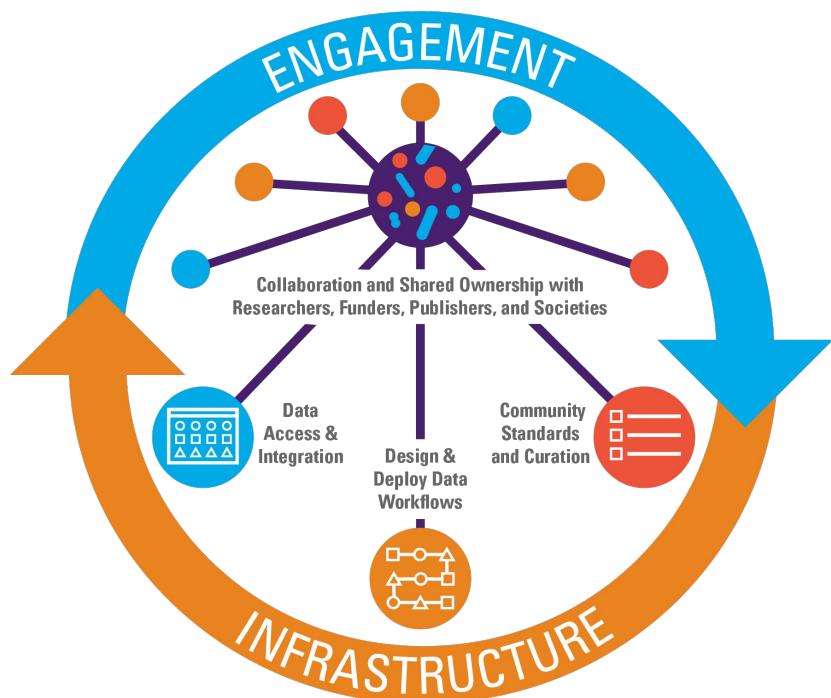
- sample_name
- project_name

Optional Field examples:

- host_age
- host_sex
- ihmc_medication_code
- host_height
- temp
- samp_store_loc
- organism_count
- misc_param

Overview of training

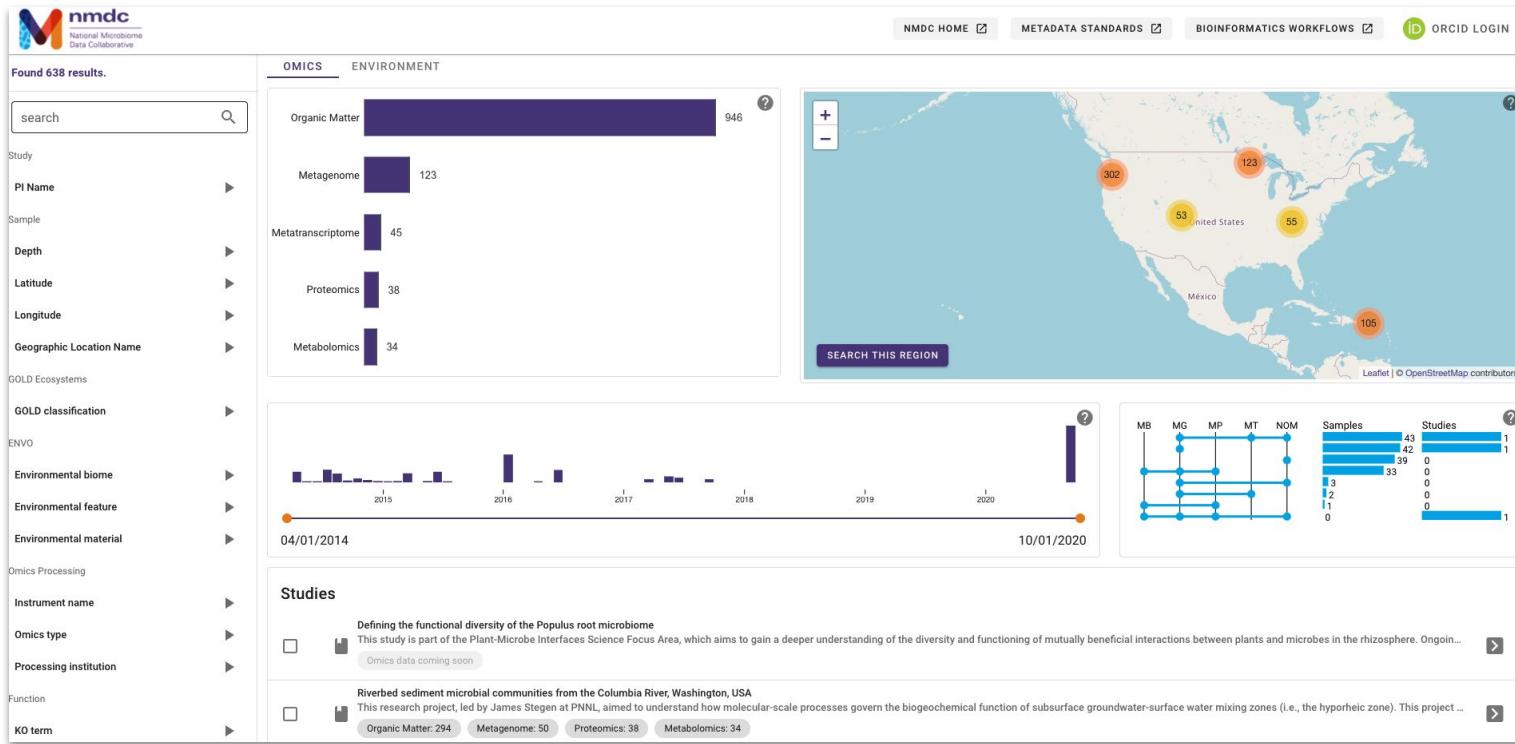
- Data Lifecycle and Metadata
- Data Management Best Practices
- The NMDC Mission
- Microbiome Metadata Community Standards
- Example Use Case – NMDC Metadata
- NMDC Data Portal



The NMDC Data Portal

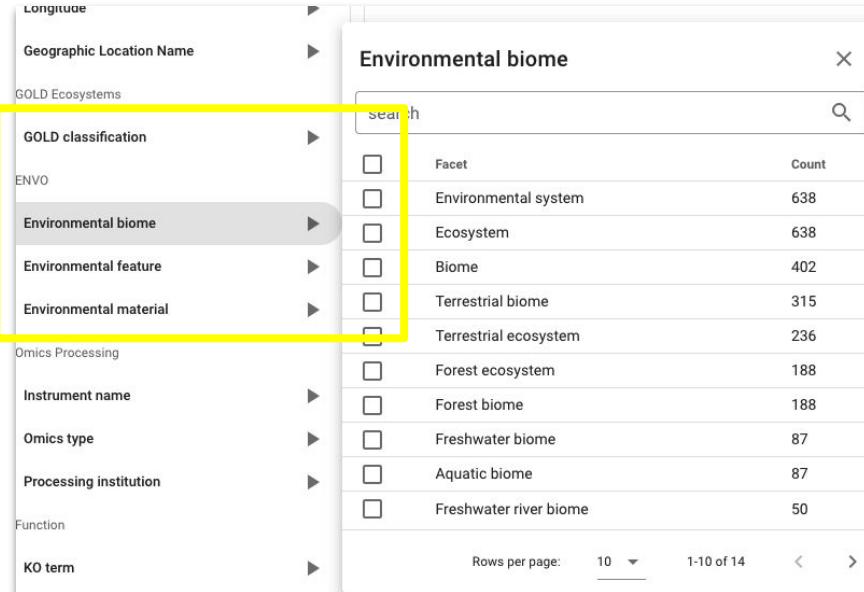


<https://data.microbiomedata.org>



The NMDC Data Portal

Mapped
environmental
Terms!



A screenshot of the NMDC Data Portal interface. On the left, a sidebar lists various categories: Longitude, Geographic Location Name, GOLD Ecosystems, GOLD classification, ENVO, Environmental biome, Environmental feature, Environmental material, Omics Processing, Instrument name, Omics type, Processing institution, Function, and KO term. The 'Environmental biome' category is highlighted with a yellow box. To the right of this sidebar is a detailed search interface titled 'Environmental biome'. It features a search bar at the top, followed by a table of results. The table has columns for 'Facet' (checkboxes), 'Count', and the term itself. The results are as follows:

| Facet | Count |
|------------------------|-------|
| Environmental system | 638 |
| Ecosystem | 638 |
| Biome | 402 |
| Terrestrial biome | 315 |
| Terrestrial ecosystem | 236 |
| Forest ecosystem | 188 |
| Forest biome | 188 |
| Freshwater biome | 87 |
| Aquatic biome | 87 |
| Freshwater river biome | 50 |

At the bottom of the search interface, there are buttons for 'Rows per page:' (set to 10), a page number indicator '1-10 of 14', and navigation arrows.

The NMDC Data Portal



Study ID [is] gold:Gs0135149

Collection date [between] (04/01/2014, 01/28/2019)

Environmental material [is] Soil

Found 53 results.

search

Study

PI Name

Sample

Depth

Latitude

Longitude

Geographic Location Name

GOLD Ecosystems

04/01/2014

01/28/2019

ENV

Environmental biome

Environmental feature

Environmental material

OMICS ENVIRONMENT

| Category | Count |
|-------------------|----------|
| Organic Matter | 652 |
| Metagenome | 48 |
| Metatranscriptome | 45 |
| Proteomics | No match |
| Metabolomics | No match |

SEARCH THIS REGION

Leaflet | © OpenStreetMap contributors

MG MT NOM

| Category | Samples | Studies |
|----------|---------|---------|
| MG | 43 | 1 |
| MT | 0 | 0 |
| NOM | 0 | 0 |

Defining the functional diversity of the Populus root microbiome

This study is part of the Plant-Microbe Interfaces Science Focus Area, which aims to gain a deeper understanding of the diversity and functioning of mutually beneficial interactions between plants and microbes in the rhizosphere. Ongoing...

Omic data coming soon

The NMDC Data Portal



Study ID [is] gold:Gs0135149

Collection date [between] (04/01/2014, 01/28/2019)

Environmental material [is] Soil

Found 53 results.

search

Study
PI Name
Sample
Depth
Latitude
Longitude

Bulk soil microbial communities from the East River watershed near Crested Butte, Colorado, United States - ER_357
Study ID: gold:Gs0135149

Bulk soil microbial communities from the East River watershed near Crested Butte, Colorado, United States - ER_150
Study ID: gold:Gs0135149

Bulk soil microbial communities from the East River watershed near Crested Butte, Colorado, United States - ER_149
Study ID: gold:Gs0135149

Samples

Bulk Download ? Log in to bulk download

Log in with OrcID to use bulk download.

1 2 3 4 >

The NMDC Data Portal



Bulk soil microbial communities from the East River watershed near Crested Butte, Colorado, United States
This research project aimed to understand how snow accumulation and snowmelt influences the mobilization of organic matter in the soil. The study was conducted in a high-elevation, semi-arid environment where snowmelt is a primary source of water and nutrients. The research team collected soil samples at various depths and analyzed them for their microbial community structure and function.

Study ID [is] gold:Gs0135149 ▶

Collection date [between] (04/01/2014, 01/28/2019) ▶

Environmental material [is] Soil ▶

Found 53 results.

search

Samples

Bulk Download ? Log in to bulk download

Log in with ORCID to use bulk download.

Bulk soil microbial communities from the East River watershed near Crested Butte, Colorado, United States - ER_357
Study ID: gold:Gs0135149 ▶
METAGENOME ▾ METATRANSCRIPTOME ▾ ORGANIC MATTER ▾

Bulk soil microbial communities from the East River watershed near Crested Butte, Colorado, United States - ER_150
Study ID: gold:Gs0135149 ▶
METAGENOME ▾ METATRANSCRIPTOME ▾ ORGANIC MATTER ▾

Bulk soil microbial communities from the East River watershed near Crested Butte, Colorado, United States - ER_149
Study ID: gold:Gs0135149 ▶
METAGENOME ▾ METATRANSCRIPTOME ▾

1 2 3 4 >

The NMDC Data Portal



 Bulk soil microbial communities from the East River watershed near Crested Butte, Colorado, United States

Description This research project aimed to understand how snow accumulation and snowmelt influences the mobilization of nitrogen through the soil microbiome in a mountainous catchment at the East River Watershed in Colorado. This project sought to identify bacteria, archaea, and fungi that were associated with the microbial biomass bloom that occurs during winter and the biomass crash following snowmelt. This project also sought to understand whether the traits that govern microbial community assembly during and after snowmelt were phylogenetically conserved. Samples were collected during winter, the snowmelt period, and after snowmelt in spring, from an area that transitioned from an upland hillslope to a riparian floodplain. This project is part of the Watershed Function Science Focus Area: <https://watershed.lbl.gov/>.

Metagenome: 48 Metatranscriptome: 45 Organic Matter: 652



Eoin Brodie
Principal investigator

Item Attributes

 ID
gold:Gs0135149

 DOI
<https://doi.org/10.25585/1488224>

 Open in GOLD
<https://gold.jgi.doe.gov/study?id=Gs0135149>

NMDC HOME  METADATA STANDARDS  BIOINFORMATICS WORKFLOWS  ORCID LOG 

Dataset Citation

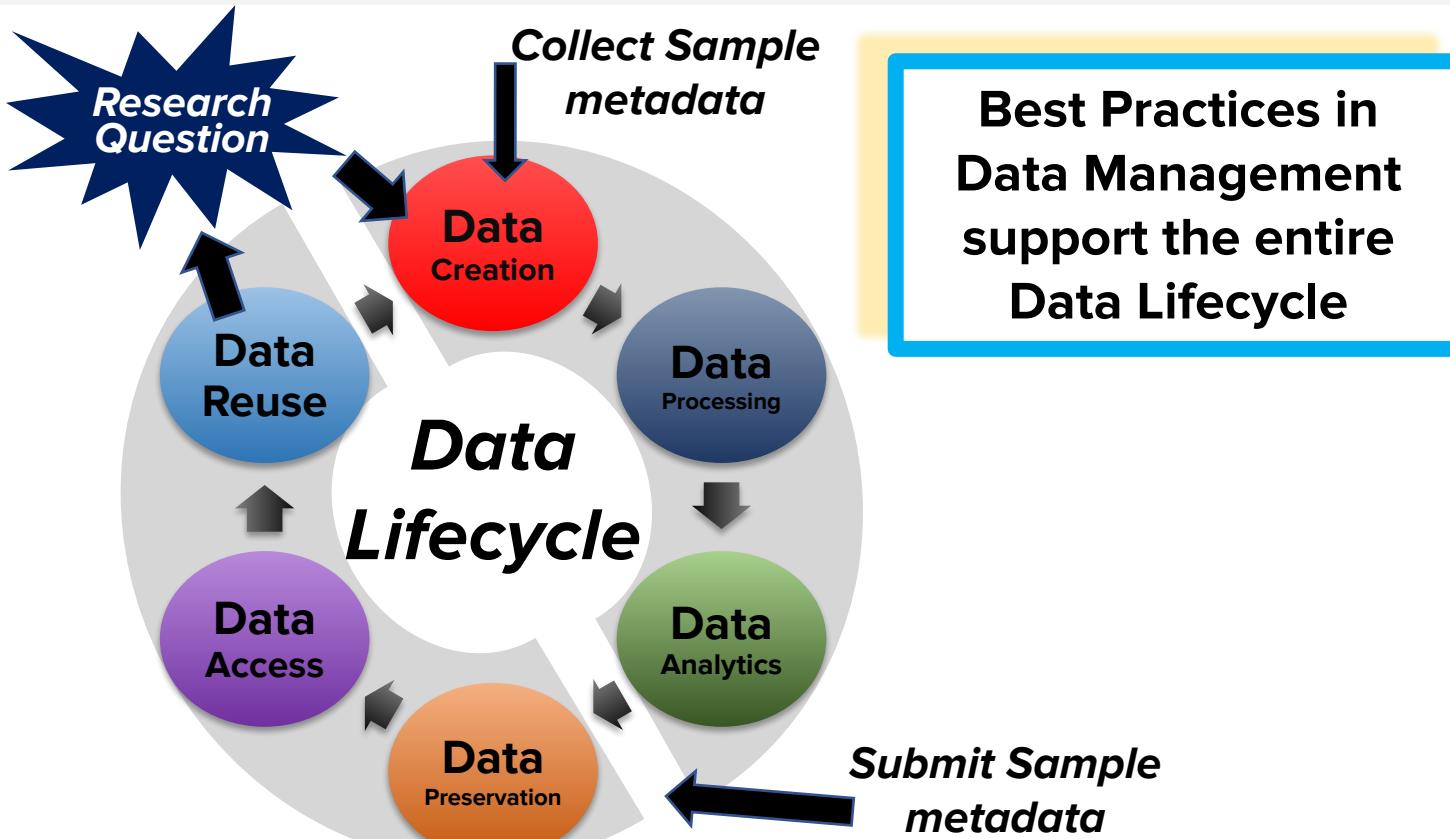
Sorensen, P., Brodie, E., Beller, H., Wang, S., Bill, M., & Bouskill, N. (2019). Sample Collection Metadata for Soil Cores from the East River Watershed, Colorado collected in 2017 [Data set]. Lawrence Berkeley National Laboratory (LBNL), Berkeley, CA (United States). <https://doi.org/10.21952/WTR/1573029>

Supporting discovery through good data management*

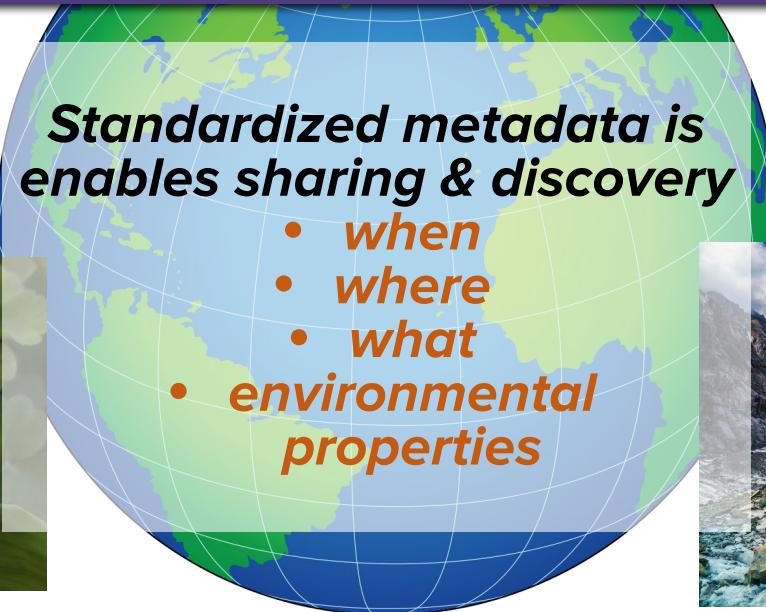
“Good data management is not a goal in itself, but rather is the key conduit leading to *knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse by the community* after the data publication process.”

*Wilkinson. et al. *The FAIR Guiding Principles for scientific data management and stewardship*. Sci Data **3**, 160018 (2016).
<https://doi.org/10.1038/sdata.2016.18>

Data Lifecycle and Metadata



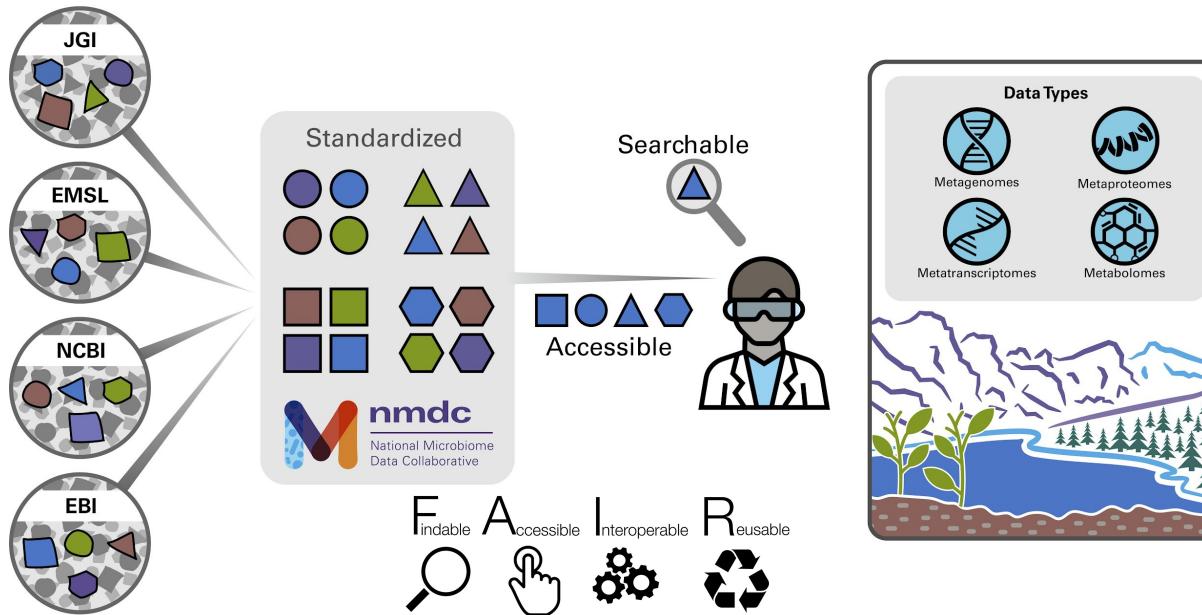
The NMDC Data Portal enables data discovery and access for data access, comparison, and meta-analysis
What's your next research question?



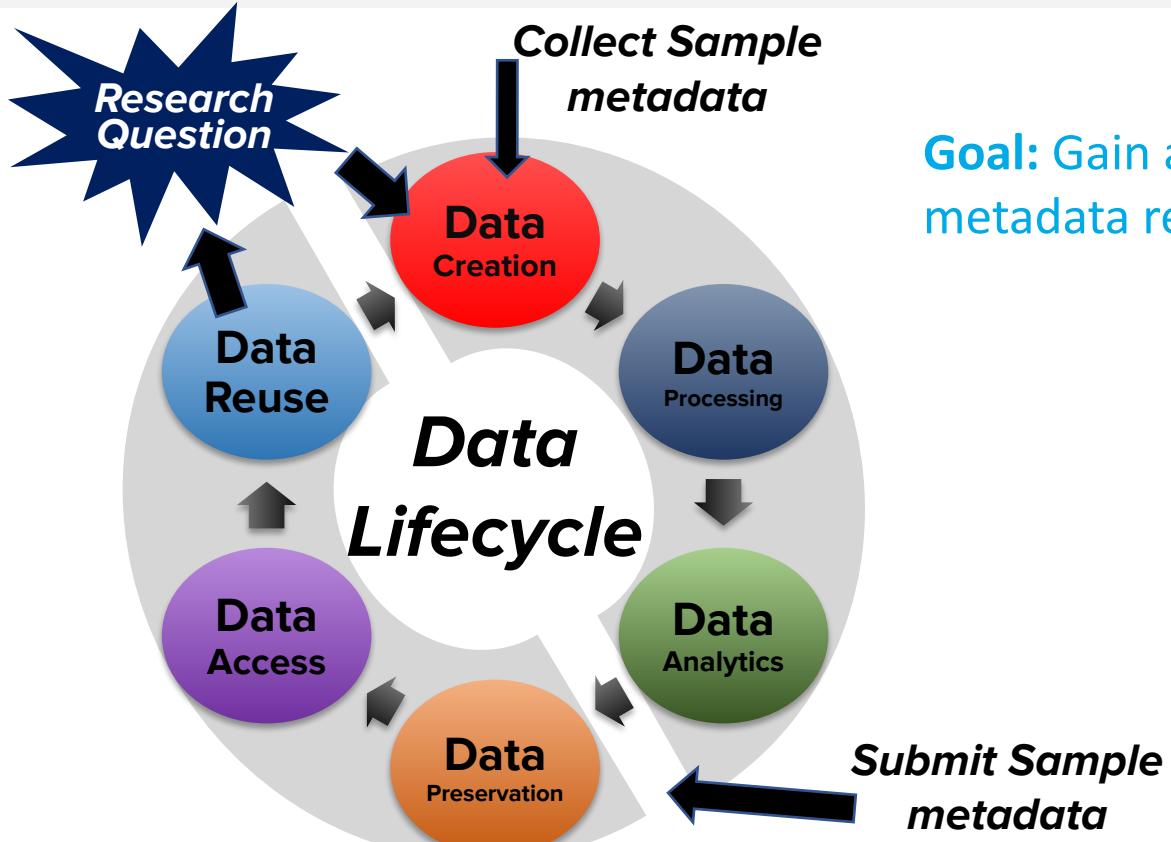
The NMDC Mission



Provide a gateway to FAIR multi-omics microbiome data
leveraging best practices for data curation and processing



Activity!



Goal: Gain an understanding of the metadata required for your studies

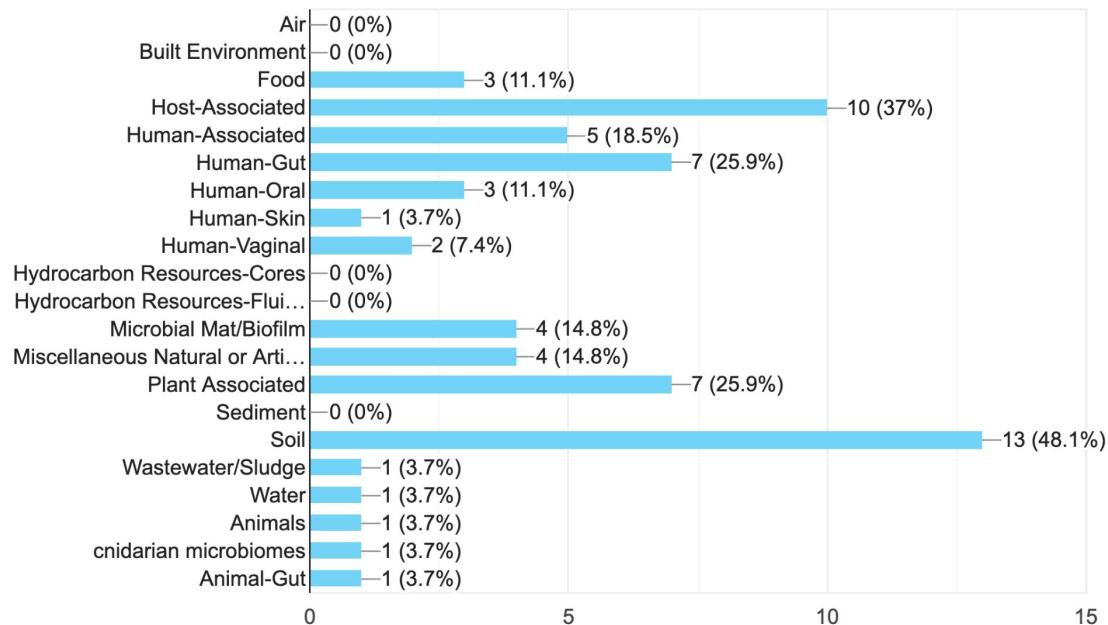
Activity!

1. Break out into small groups (~4-5 people) by environment type.

Activity!

What microbiomes do you study? Please check all that apply

27 responses



Activity!

1. Break out into small groups (~4-5 people) by environment type.
2. Introduce yourselves and elevator pitch of what you study.
3. Select one person's study to go into more depth for the activity.
4. What sample metadata should be collected for this study?
5. What MIxS package is relevant for this study?
6. What are the mandatory fields required for your MIxS package?
7. What are some optional fields that could be useful for your study?
8. Are there any fields that would be important to include for your study but aren't included in your MIxS package?
9. Do you feel any of the optional fields should be mandatory in your package?

Example: Human Microbiome MixS



nmdc
National Microbiome
Data Collaborative

<https://github.com/GenomicsStandardsConsortium/mixs/tree/main/release/excel>

A screenshot of a GitHub repository page for 'GenomicsStandardsConsortium / mixs'. The repository is public and has 16 watchers, 12 forks, and 14 stars. The main branch is 'main'. A list of files under the 'mixs / release / excel' directory is shown, all added by user 'ramonawalls' on March 23, 2023. The files include various microbiome samples: agriculture_mixs6.xlsx, air_mixs6.xlsx, built-environment_mixs6.xlsx, food-animal-and-feed_mixs6.xlsx, food-farm-environment_mixs6.xlsx, food-human-foods_mixs6.xlsx, food-production-facility_mixs6.xlsx, host-associated_mixs6.xlsx, human-associated_mixs6.xlsx, human-gut_mixs6.xlsx, human-oral_mixs6.xlsx, human-skin_mixs6.xlsx, and human-vaginal_mixs6.xlsx. An arrow points to the 'human-gut_mixs6.xlsx' file.

| File | Description | Last Commit |
|-------------------------------------|-----------------------------------|--------------|
| agriculture_mixs6.xlsx | add manually created excdel files | 2 months ago |
| air_mixs6.xlsx | add manually created excdel files | 2 months ago |
| built-environment_mixs6.xlsx | add manually created excdel files | 2 months ago |
| food-animal-and-feed_mixs6.xlsx | add manually created excdel files | 2 months ago |
| food-farm-environment_mixs6.xlsx | add manually created excdel files | 2 months ago |
| food-human-foods_mixs6.xlsx | add manually created excdel files | 2 months ago |
| food-production-facility_mixs6.xlsx | add manually created excdel files | 2 months ago |
| host-associated_mixs6.xlsx | add manually created excdel files | 2 months ago |
| human-associated_mixs6.xlsx | add manually created excdel files | 2 months ago |
| human-gut_mixs6.xlsx | add manually created excdel files | 2 months ago |
| human-oral_mixs6.xlsx | add manually created excdel files | 2 months ago |
| human-skin_mixs6.xlsx | add manually created excdel files | 2 months ago |
| human-vaginal_mixs6.xlsx | add manually created excdel files | 2 months ago |

Example: Human Microbiome MixS



<https://github.com/GenomicsStandardsConsortium/mixs/tree/main/release/excel>

A screenshot of a GitHub repository page for 'GenomicsStandardsConsortium / mixs'. The page shows a single file, 'human-gut_mixs6.xlsx', which is 77.1 KB in size and was added by 'ramonawalls'. A red circle highlights the 'Download' button in the file's preview card. The GitHub interface includes a navigation bar with links for Code, Issues, Pull requests, Actions, Projects, Wiki, Security, and Insights. The repository has 16 watchers, 12 forks, and 14 stars.

Example: Human Microbiome MixS



Screenshot of Microsoft Excel showing the 'human-gut_mixs6 (2)' spreadsheet. The table defines environmental packages and their structured comment names, package items, definitions, expected values, value syntax, examples, requirements, preferred units, occurrences, and MixS IDs.

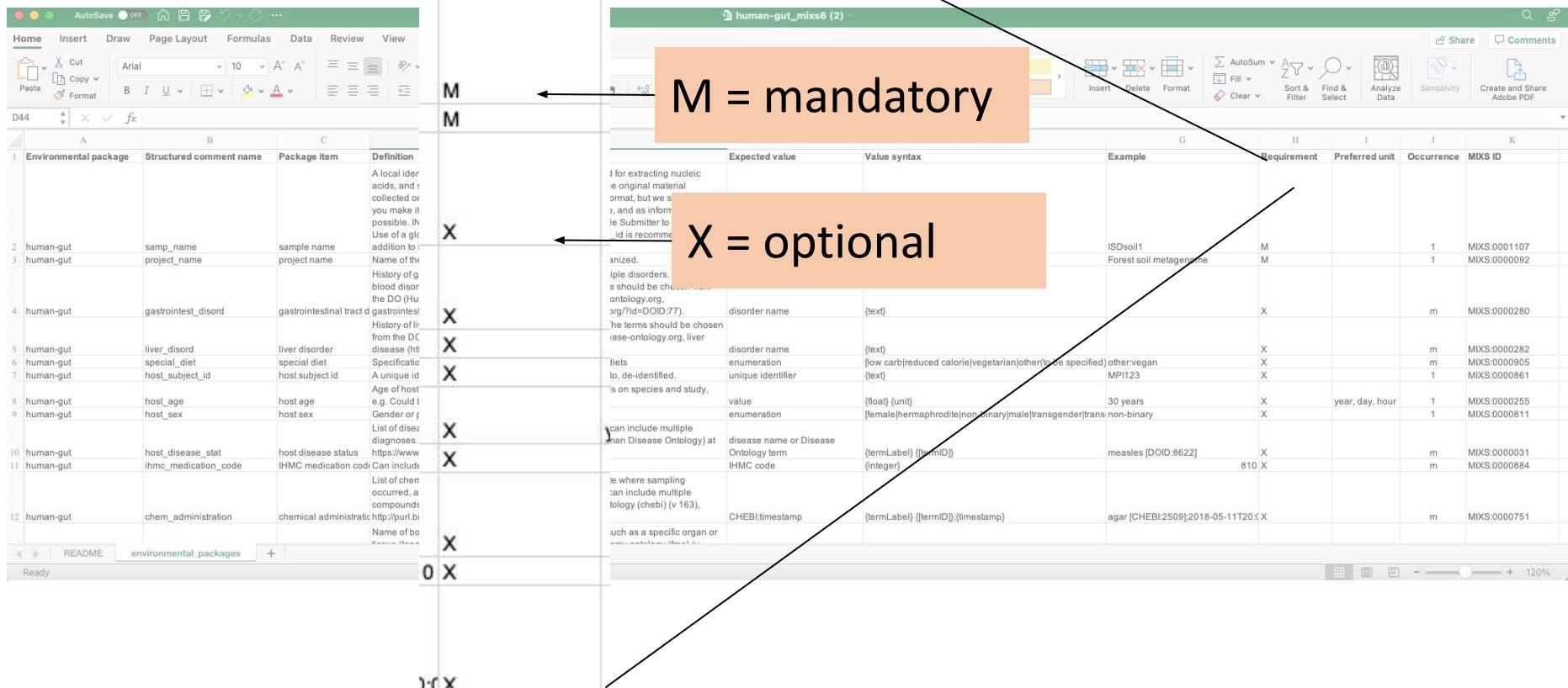
| | A | B | C | D | E | F | G | H | I | J | K |
|----|-----------------------|-------------------------|---------------------------------|---|---|--------------|------------------------------------|-----------------|----------------|------------|--------------|
| | Environmental package | Structured comment name | Package item | Definition | Expected value | Value syntax | Example | Requirement | Preferred unit | Occurrence | MIXS ID |
| 1 | | | | A local identifier or name that for the material sample used for extracting nucleic acids, and subsequent sequencing. It can refer either to the original material collector or to any derived sub-samples. It can have any format, but we suggest that you make it concise, unique and consistent within your lab, and as informative as possible. INSDC requires every sample name from a single Submitter to be unique. Use of a globally unique identifier for the field source_mat_id is recommended in addition to sample_name. | | | | | | | |
| 2 | human-gut | samp_name | sample name | text | {text} | | ISDsoil1 | M | | 1 | MIXS:0001107 |
| 3 | human-gut | project_name | project name | text | {text} | | Forest soil metagenome | M | | 1 | MIXS:000092 |
| 4 | human-gut | gastrointest_disord | gastrointestinal tract disorder | disorder name | {text} | | | X | | m | MIXS:0000280 |
| 5 | human-gut | liver_disord | liver disorder | disorder name | {text} | | | X | | m | MIXS:0000282 |
| 6 | human-gut | special_diet | special diet | enumeration | {low carb reduced calorie vegetarian other(to be specified)} | other:vegan | | X | | m | MIXS:0000905 |
| 7 | human-gut | host_subject_id | host subject id | unique identifier | {text} | | MP1f23 | X | | 1 | MIXS:0000861 |
| 8 | human-gut | host_age | host age | value | {float} {unit} | 30 years | X | year, day, hour | 1 | | MIXS:0000255 |
| 9 | human-gut | host_sex | host sex | enumeration | {female hermaphrodite non-binary male transgender trans non-binary} | | X | | 1 | | MIXS:0000811 |
| 10 | human-gut | host_disease_stat | host disease status | disease name or Disease Ontology term | {termLabel} [{termID}] | | measles [DOID:8622] | X | | m | MIXS:0000031 |
| 11 | human-gut | ihmc_medication_code | IHMC medication code | IHMC code | {integer} | | 810 | X | | m | MIXS:0000884 |
| 12 | human-gut | chem_administration | chemical administration | CHEBI:timestamp | {termLabel} [{termID}];{timestamp} | | agar [CHEBI:2509];2018-05-11T20:XX | | | m | MIXS:0000751 |

READY environmental_packages +

Ready

120%

Example: Human Gut Microbiome MixS



Requirement

| | A | B | C | D | E | F | G | H | I | J | K | |
|----|-----------------------|-------------------------|---------------------------------|--|---|---------------------------------------|---|------------------------------------|-------------|-----------------|------------|--------------|
| | Environmental package | Structured comment name | Package item | Definition | | Expected value | Value syntax | Example | Requirement | Preferred unit | Occurrence | MIXS ID |
| 1 | human-gut | samp_name | sample name | A local identifier for extracting nucleic acids, and is collected or you make it possible. In addition to the sample name, the History of blood disorders, the DO (Human Disease Ontology) at the DC level should be chosen. The terms should be chosen from the Disease Ontology, liver disorder. A unique identifier for the sample name, project name, and host subject id. | X | disorder name | {text} | ISDsoil1 | M | m | 1 | MIXS:0001107 |
| 2 | human-gut | project_name | project name | | X | disorder name | {text} | Forest soil metagenome | M | m | 1 | MIXS:000092 |
| 3 | human-gut | gastrointest_disord | gastrointestinal tract disorder | The History of blood disorders, the DO (Human Disease Ontology) at the DC level should be chosen. The terms should be chosen from the Disease Ontology, liver disorder. A unique identifier for the sample name, project name, and host subject id. | X | disorder name | {text} | | X | m | | MIXS:000280 |
| 4 | human-gut | liver_disord | liver disorder | | X | disorder name | {text} | | X | m | | MIXS:000282 |
| 5 | human-gut | special_diet | special diet | | X | enumeration | {low carb reduced calorie vegetarian other} | | X | m | | MIXS:000905 |
| 6 | human-gut | host_subject_id | host subject id | A unique identifier for the sample name, project name, and host subject id. | X | unique identifier | {text} | MPI123 | X | m | 1 | MIXS:000861 |
| 7 | human-gut | host_age | host age | Age of host, e.g. Could I be pregnant? | X | value | {float} {unit} | 30 years | X | year, day, hour | 1 | MIXS:000255 |
| 8 | human-gut | host_sex | host sex | Gender or sex | X | enumeration | {female hermaphrodite non-binary male transgender trans non-binary} | | X | m | 1 | MIXS:000811 |
| 9 | human-gut | host_disease_stat | host disease status | List of disease diagnoses, e.g. Could I have a cold? | X | disease name or Disease Ontology term | {termLabel} [{termID}] | measles [DOID:8622] | X | m | | MIXS:000031 |
| 10 | human-gut | ihmc_medication_code | ihmc medication code | Code that can include multiple IHMC codes | X | IHMC code | {integer} | 810 | X | m | | MIXS:000084 |
| 11 | human-gut | chem_administration | chemical administration | List of chemicals that can occur, a compound | X | CHEBI:timestamp | {termLabel} [{termID}];{timestamp} | agar [CHEBI:2509];2018-05-11T20:XX | X | m | | MIXS:000751 |
| 12 | human-gut | | | Name of both | X | | | | | | | |

READY environment_packages + 0 X

Take home messages



- 1) Understanding about importance of collecting and sharing metadata
- 2) Identify metadata standards relevant to your studies
- 3) Gained familiarity with relevant MIxS packages
- 4) Gained familiarity with the NMDC mission
- 5) Small improvements will lead to big change!

Post-event survey



<https://bit.ly/PSUNMDC>

Contributors:

- Michele Thornton, Geospatial Analyst/Data Curation at ORNL DAAC
- Emily Davenport, Penn State University

The NMDC team:
<https://microbiomedata.org/team>

*Questions?
Thank you!*