



nmcdc

National Microbiome
Data Collaborative

2020 BETO Leveraging Existing Bioenergy Data

Kjiersten Fagnan

NMDC Infrastructure Lead, CIO DOE Joint Genome Institute

July 21, 2020

“Cost of data”

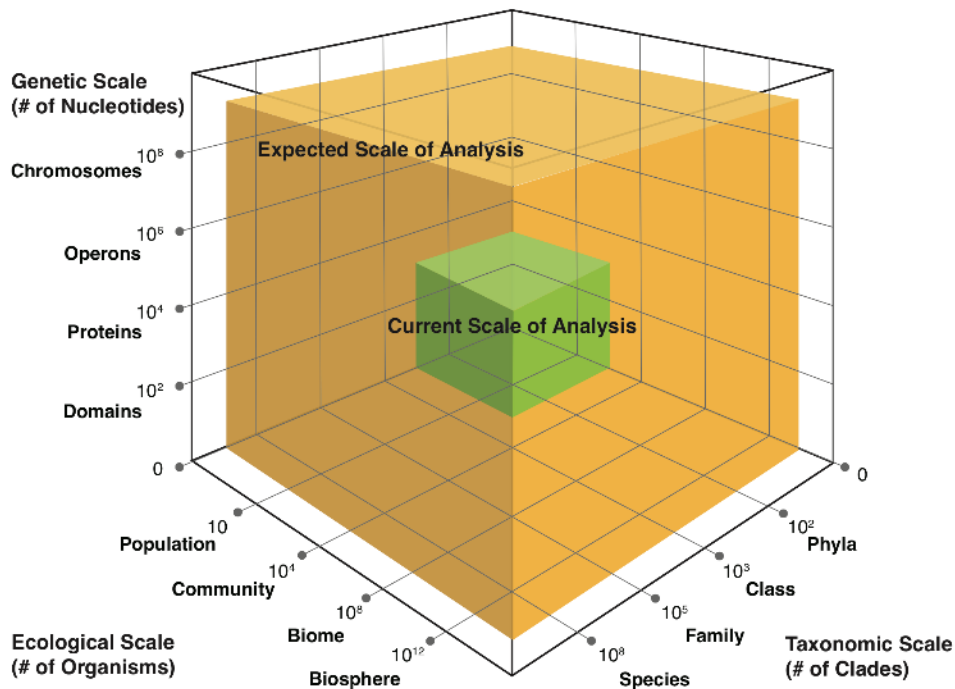


How much should we be “willing to pay for data”?

- Data are worthless without good ***contextual information*** (i.e. metadata) - it’s expensive (and potentially impossible) to curate data after the fact!
- There is a cost associated with building the ***software and hardware infrastructure*** needed to facilitate data access and sharing

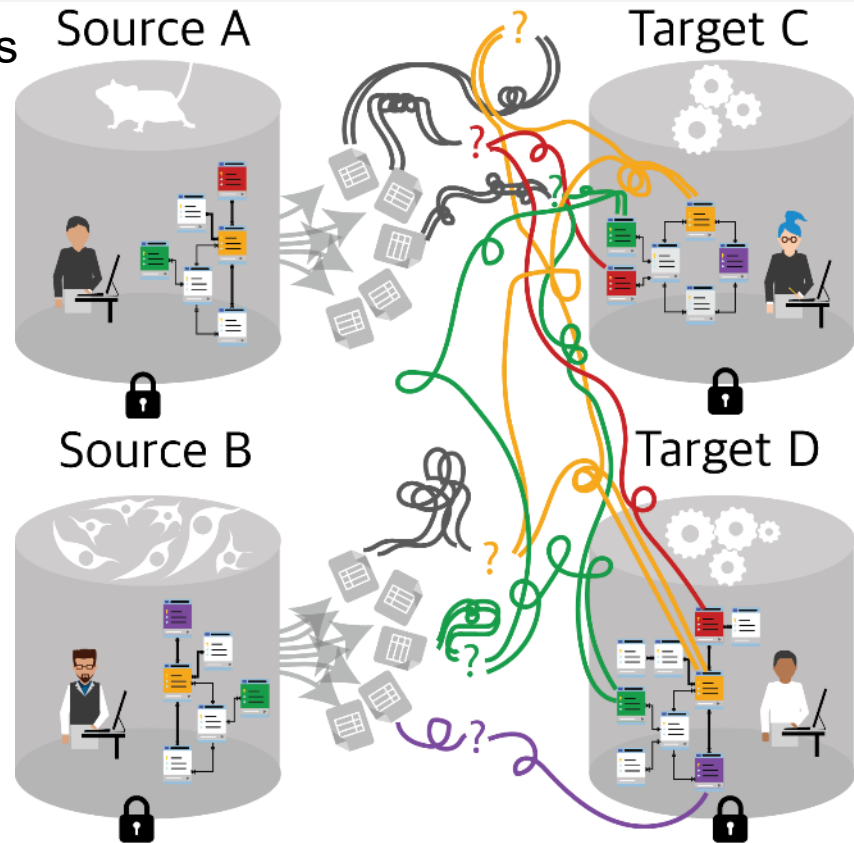
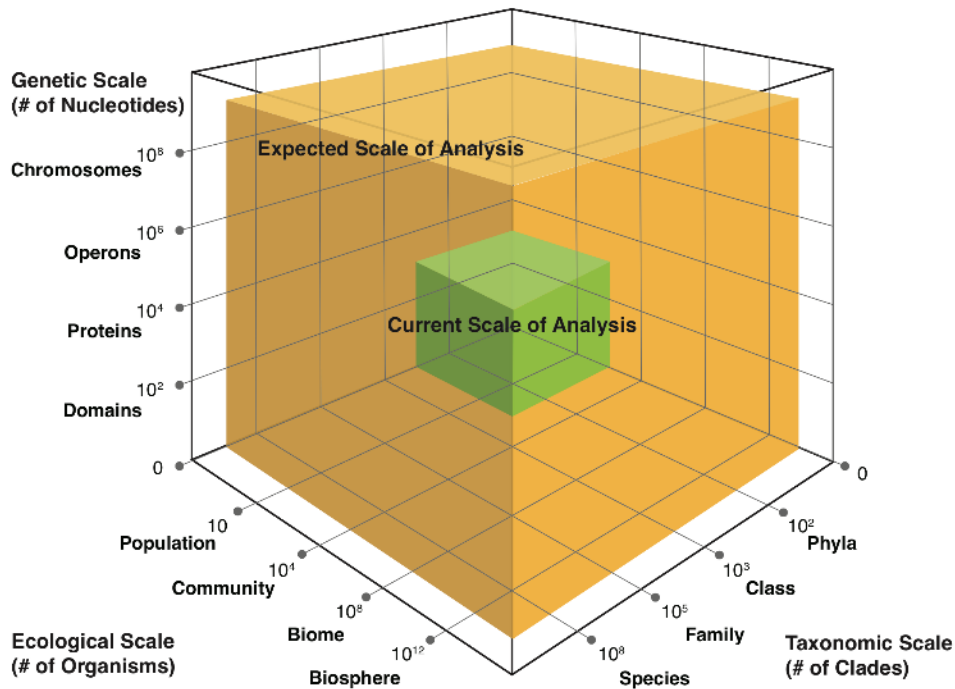
The immense scale of omics data

Advances in sequencing and omics technologies have **far outpaced** data infrastructure



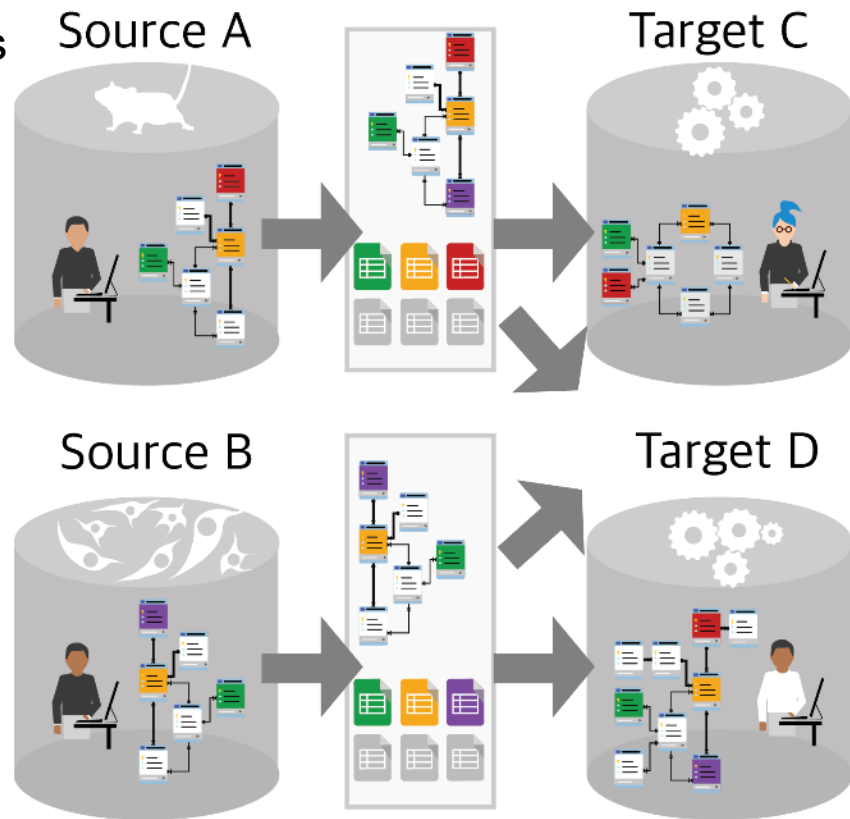
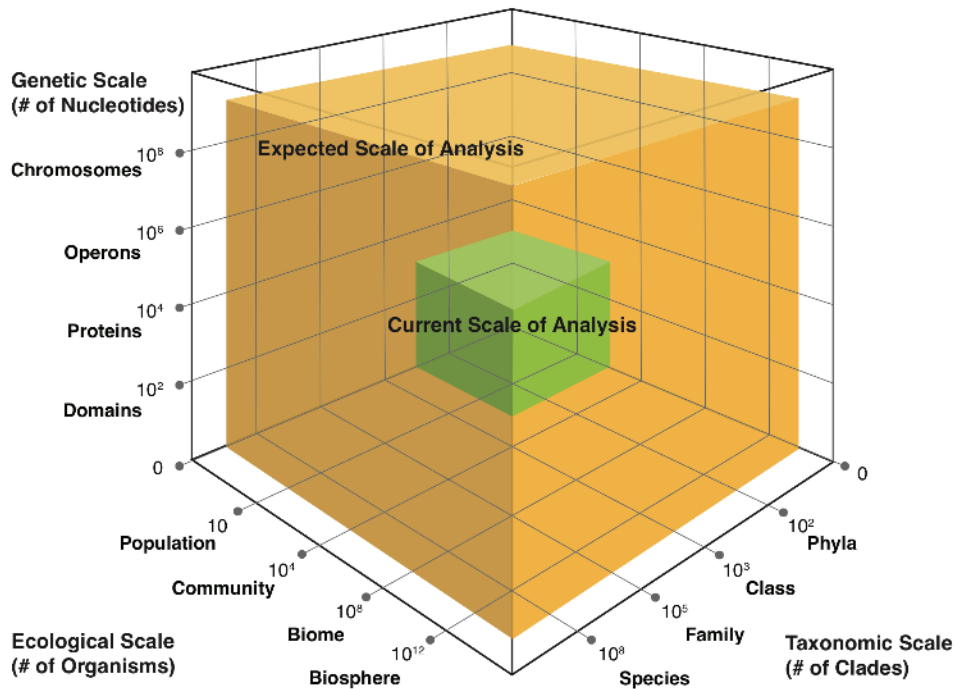
The immense scale of omics data

Advances in sequencing and omics technologies have **far outpaced** data infrastructure



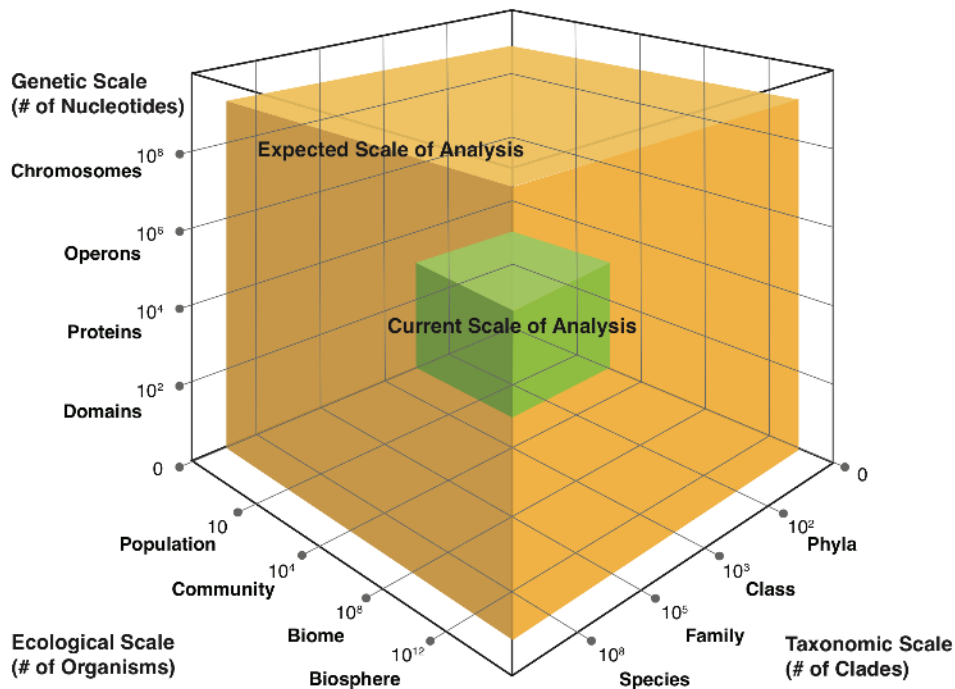
The immense scale of omics data

Advances in sequencing and omics technologies have **far outpaced** data infrastructure

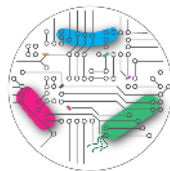


The immense scale of omics data

Advances in sequencing and omics technologies have **far outpaced** data infrastructure



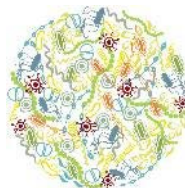
Imagine the possibilities



Uncover biological paradigms and ecosystem phenomena derived from integrated omics surveys



Link genotype to phenotype



Microbiome dynamics and ecosystem processes through integrating molecular and process measurements

What can you do with all that data?



nmdc
National Microbiome
Data Collaborative

A Gordon Bell Prize (Supercomputing) winner in 2018 used all the well-characterized publicly available data to look at genetic underpinnings of opioid addiction.

Employing Supercomputers to Combat the Opioid Epidemic

Paper Title: “Attacking the Opioid Epidemic: Determining the Epistatic and Pleiotropic Genetic Architectures for Chronic Pain and Opioid Addiction”

Prize Category: Sustained Performance Prize

Wayne Joubert, et al. 2018. Attacking the opioid epidemic: determining the epistatic and pleiotropic genetic architectures for chronic pain and opioid addiction. In Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC '18). IEEE Press, Article 57, 1–14.

What can you do with all that data?



nmdc

National Microbiome
Data Collaborative

A Gor
used a
look a

Access to large amounts of 'omics data enables scientists to explore a broad range of hypotheses!

2018
data to
n.

Employ

Paper T

Architec

opic Genetic

Prize Category: Sustained Performance Prize

Wayne Joubert, et al. 2018. Attacking the opioid epidemic: determining the epistatic and pleiotropic genetic architectures for chronic pain and opioid addiction. In Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC '18). IEEE Press, Article 57, 1–14.

Envisioning the NMDC



Broad and inclusive program development activities over two years

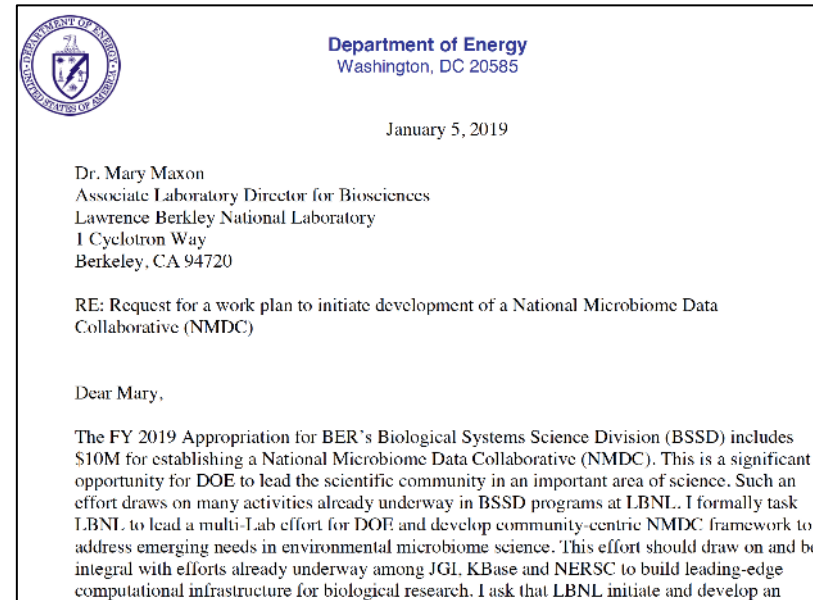


2019 DOE Task Letter



LBL to lead a multi-Lab effort for DOE to develop a community -centric NMDC framework

- Provide a framework with sufficient agility to **enable complete data access, advanced analyses, and tool development** for addressing microbiome research
- Adhere to **FAIR principles**
- Empower the broader scientific community to **access, analyze, share and reproduce** microbiome data to promote reproducibility and enhance cross-study comparison
- Allow **integration of empirical, computational, and mechanistic modeling tools** for prediction and management of microbial communities' dynamics and activities
- Take advantage of **HPC systems** available within the DOE complex
- Facilitate **incorporation of new or updated information** (i.e. annotation) as it becomes available
- **Maintain contact** with the larger microbiome research community to assess changing needs and/or capabilities



2019 DOE Task Letter

LBNL to lead

- Provide a framework for **data access** addressing
- Adhere to **F**
- Empower the **share and r** reproducibility
- Allow **integ** **mechanistic** microbial community
- Take advantage of **ITC systems** available within the DOE complex
- Facilitate **incorporation of new or updated information** (i.e. annotation) as it becomes available
- **Maintain contact** with the larger microbiome research community to assess changing needs and/or capabilities

The U.S. Department of Energy has invested **\$20M** in a pilot to improve the **quality of and access to microbiome data.**

NMDC framework

Energy
20585
2019
of a National Microbiome Data
Biological Systems Science Division (BSSD) includes \$10M for establishing a National Microbiome Data Collaborative (NMDC). This is a significant opportunity for DOE to lead the scientific community in an important area of science. Such an effort draws on many activities already underway in BSSD programs at LBNL. I formally task LBNL to lead a multi-Lab effort for DOE and develop community-centric NMDC framework to address emerging needs in environmental microbiome science. This effort should draw on and be integral with efforts already underway among JGL, KBase and NERSC to build leading-edge computational infrastructure for biological research. I ask that LBNL initiate and develop an

Vision

Empower the research community to harness microbiome data exploration and discovery through a collaborative and integrative data science ecosystem



NMDC Guiding Principles



Standards

Community-driven and
accepted

Continued development to
address future needs



Quality

Curation and quality control
to ensure data adheres to
those standards



Integration

Standardized, reproducible
analytical pipelines across
heterogeneous data sets

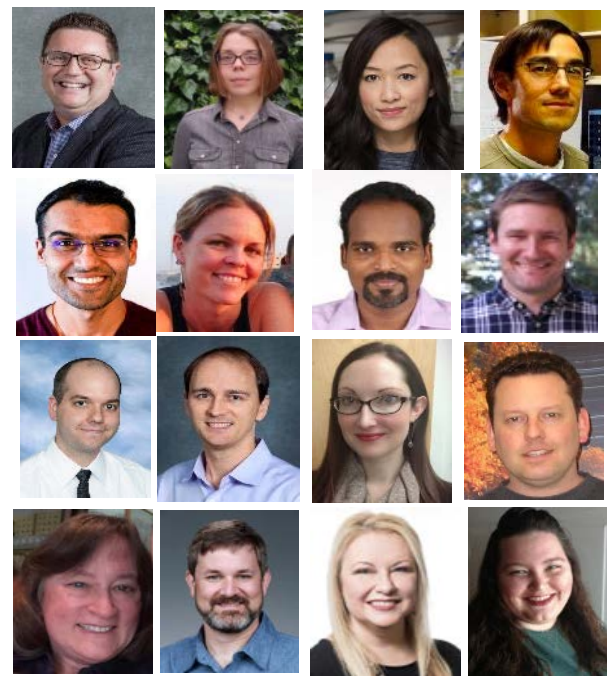
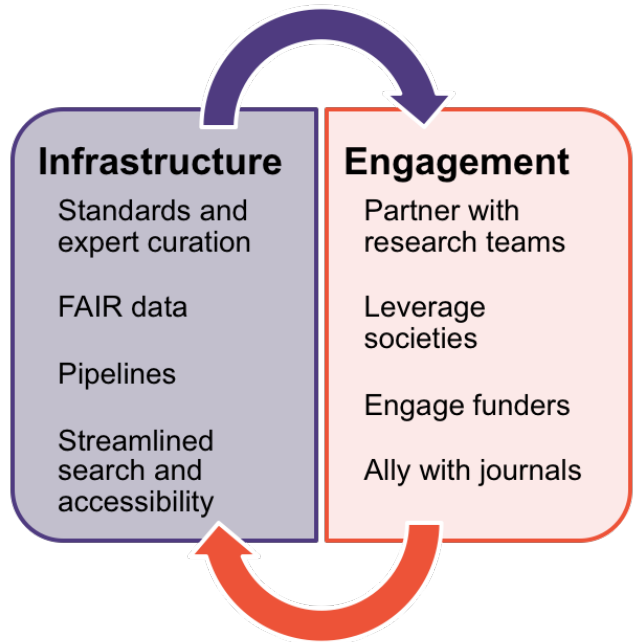


Access

Discovery based on scientific
inquiry

Search using existing data

One vision, two strategic priorities



Inclusive engagement strategy



Engagement

Partner with
research teams

Leverage
societies

Engage funders

Ally with journals

Inclusive engagement strategy



Collaborate with research teams
to support individual projects

Work across research
programs and initiatives

Engagement

Partner with
research teams

Leverage
societies

Engage funders

Ally with journals

Inclusive engagement strategy



Collaborate with research teams
to support individual projects

Work across research
programs and initiatives

Network with broad
stakeholder groups

Engagement

Partner with
research teams

Leverage
societies

Engage funders

Ally with journals

Inclusive engagement strategy



Collaborate with research teams
to support individual projects

Work across research
programs and initiatives

Network with broad
stakeholder groups

Support data management
plans across agencies

Improve links across data
and publications

Engagement

Partner with
research teams

Leverage
societies

Engage funders

Ally with journals

An integrative infrastructure



Infrastructure

Standards and
expert curation

FAIR data

Workflows

Streamlined
search and
accessibility

An integrative infrastructure

Minimal Information about
any (x) Sequence (**MiXS**)



Map ontologies and
harmonize sample metadata



Infrastructure

Standards and
expert curation

FAIR data

Workflows

Streamlined
search and
accessibility

An integrative infrastructure

Minimal Information about
any (x) Sequence (**MIxS**)



Map ontologies and
harmonize sample metadata



**FAIR: Findable, Accessible,
Interoperable, Reusable**



Infrastructure

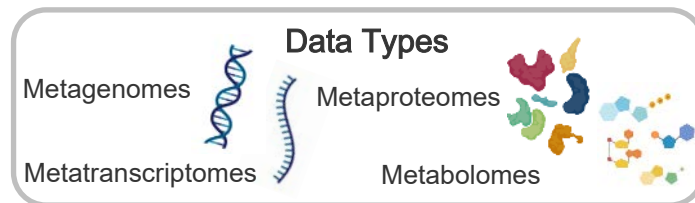
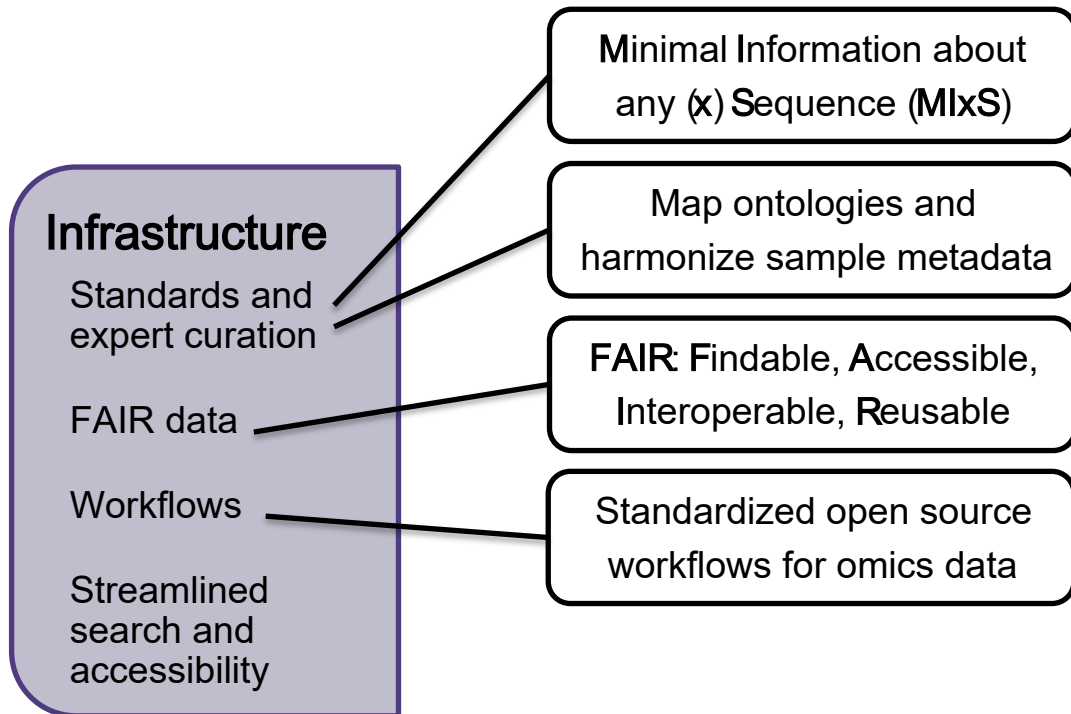
Standards and
expert curation

FAIR data

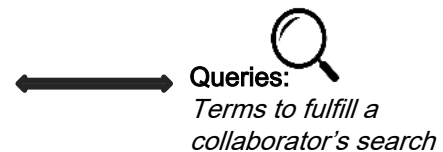
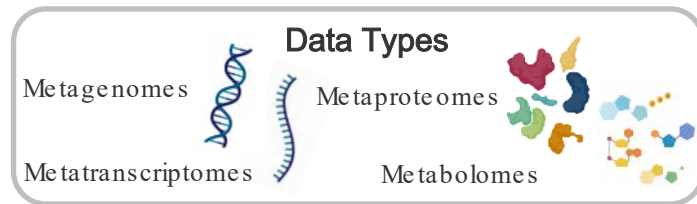
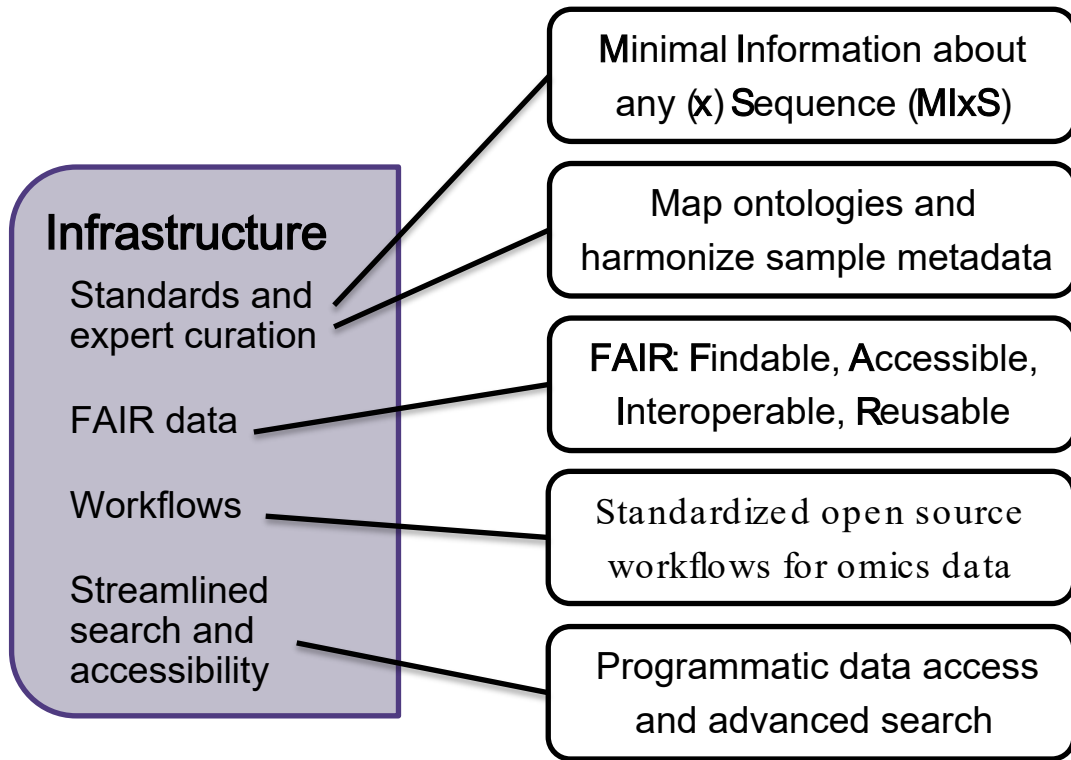
Workflows

Streamlined
search and
accessibility

An integrative infrastructure



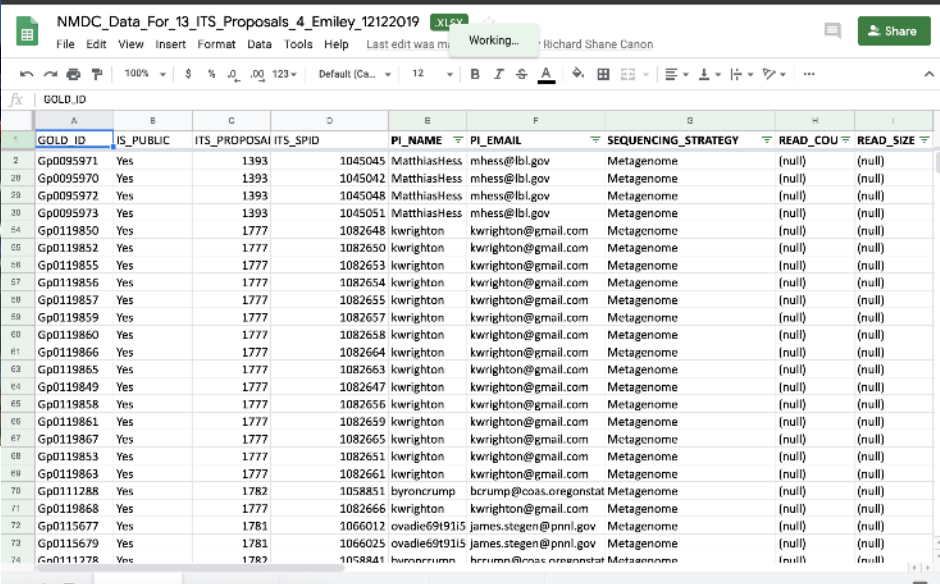

An integrative infrastructure



The Metadata Challenge



The Metadata Challenge



Excel spreadsheet titled "NMDC_Data_For_13_ITS_Proposals_4_Emiley_12122019" showing a table of metadata for ITS proposals. The table has columns for GOLD_ID, IS_PUBLIC, ITS_PROPOSAL, ITS_SPID, PI_NAME, PI_EMAIL, SEQUENCING_STRATEGY, READ_COU, and READ_SIZE.

GOLD_ID	IS_PUBLIC	ITS_PROPOSAL	ITS_SPID	PI_NAME	PI_EMAIL	SEQUENCING_STRATEGY	READ_COU	READ_SIZE
Gp0095971	Yes		1393	1045045 MatthiasHess	mhess@lbl.gov	Metagenome	(null)	(null)
Gp0095970	Yes		1393	1045042 MatthiasHess	mhess@lbl.gov	Metagenome	(null)	(null)
Gp0095972	Yes		1393	1045048 MatthiasHess	mhess@lbl.gov	Metagenome	(null)	(null)
Gp0095973	Yes		1393	1045051 MatthiasHess	mhess@lbl.gov	Metagenome	(null)	(null)
Gp0119850	Yes		1777	1082648 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
Gp0119852	Yes		1777	1082650 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
Gp0119855	Yes		1777	1082653 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
Gp0119856	Yes		1777	1082654 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
Gp0119857	Yes		1777	1082655 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
Gp0119859	Yes		1777	1082657 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
Gp0119860	Yes		1777	1082658 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
Gp0119866	Yes		1777	1082664 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
Gp0119865	Yes		1777	1082663 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
Gp0119849	Yes		1777	1082647 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
Gp0119858	Yes		1777	1082656 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
Gp0119861	Yes		1777	1082659 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
Gp0119867	Yes		1777	1082665 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
Gp0119853	Yes		1777	1082651 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
Gp0119863	Yes		1777	1082661 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
Gp0111288	Yes		1782	1058851 byroncrump	bcrump@coas.oregonstat	Metagenome	(null)	(null)
Gp0119868	Yes		1777	1082666 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
Gp0115677	Yes		1781	1066012 ovdie691915	james.stegen@pnml.gov	Metagenome	(null)	(null)
Gp0115679	Yes		1781	1066025 ovdie691915	james.stegen@pnml.gov	Metagenome	(null)	(null)
Gp0111778	Yes		1782	1058841 byroncrump	bcrump@coas.oregonstat	Metagenome	(null)	(null)

The Metadata Challenge

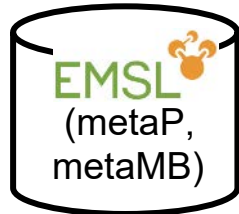
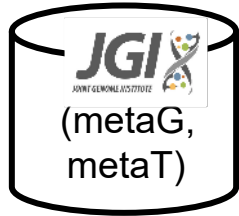


	B	C	D	E	F	G	H	I
	IS_PUBLIC	ITS_PROPOSAL	ITS_SPID	PI_NAME	PI_EMAIL	SEQUENCING_STRATEGY	READ_COU	READ_SIZE
	Yes		1393	1045045 MatthiasHess	mhess@lbl.gov	Metagenome	(null)	(null)
	Yes		1393	1045042 MatthiasHess	mhess@lbl.gov	Metagenome	(null)	(null)
	Yes		1393	1045048 MatthiasHess	mhess@lbl.gov	Metagenome	(null)	(null)
	Yes		1393	1045051 MatthiasHess	mhess@lbl.gov	Metagenome	(null)	(null)
	Yes		1777	1082648 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
	Yes		1777	1082650 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
	Yes		1777	1082653 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
	Yes		1777	1082654 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
	Yes		1777	1082655 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
	Yes		1777	1082657 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
	Yes		1777	1082658 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
	Yes		1777	1082664 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
	Yes		1777	1082663 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
	Yes		1777	1082647 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
	Yes		1777	1082656 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
66	Gp0119861	Yes	1777	1082659 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
67	Gp0119867	Yes	1777	1082665 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
68	Gp0119853	Yes	1777	1082651 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
69	Gp0119863	Yes	1777	1082661 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
70	Gp0111288	Yes	1782	1058851 byroncrump	bcrump@coas.oregonstat	Metagenome	(null)	(null)
71	Gp0119868	Yes	1777	1082666 kwrighton	kwrighton@gmail.com	Metagenome	(null)	(null)
72	Gp0115677	Yes	1781	1066012 ovdie691915	james.stegen@pnml.gov	Metagenome	(null)	(null)
73	Gp0115679	Yes	1781	1066025 ovdie691915	james.stegen@pnml.gov	Metagenome	(null)	(null)
74	Gp0111778	Yes	1782	1058841 byroncrump	bcrump@coas.oregonstat	Metagenome	(null)	(null)

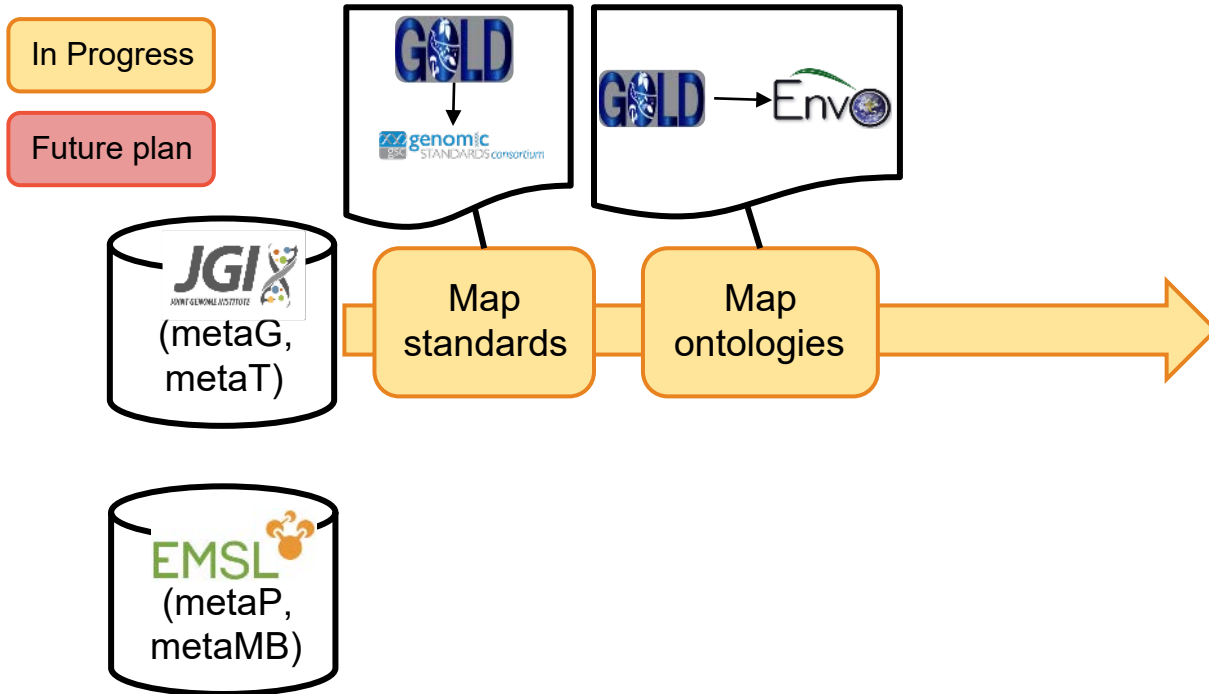
Community-driven standards & ontologies

In Progress

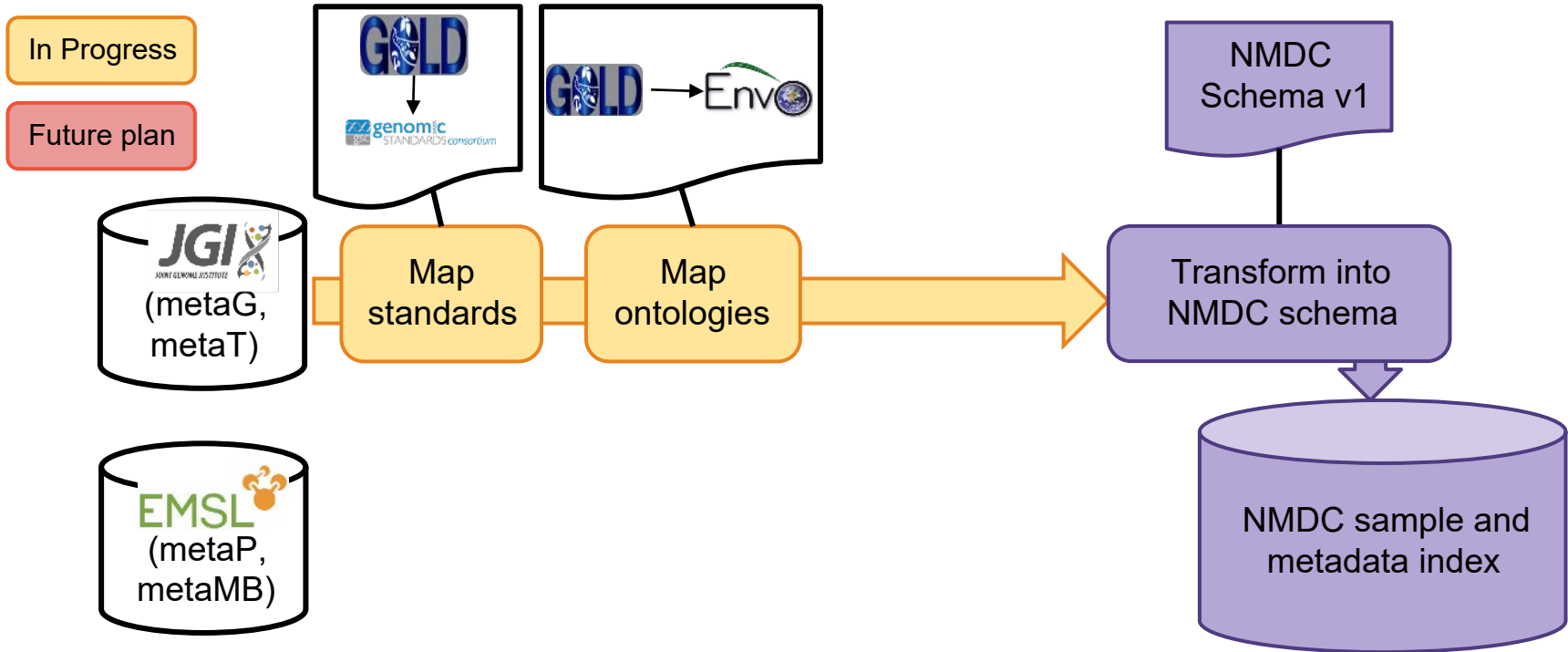
Future plan



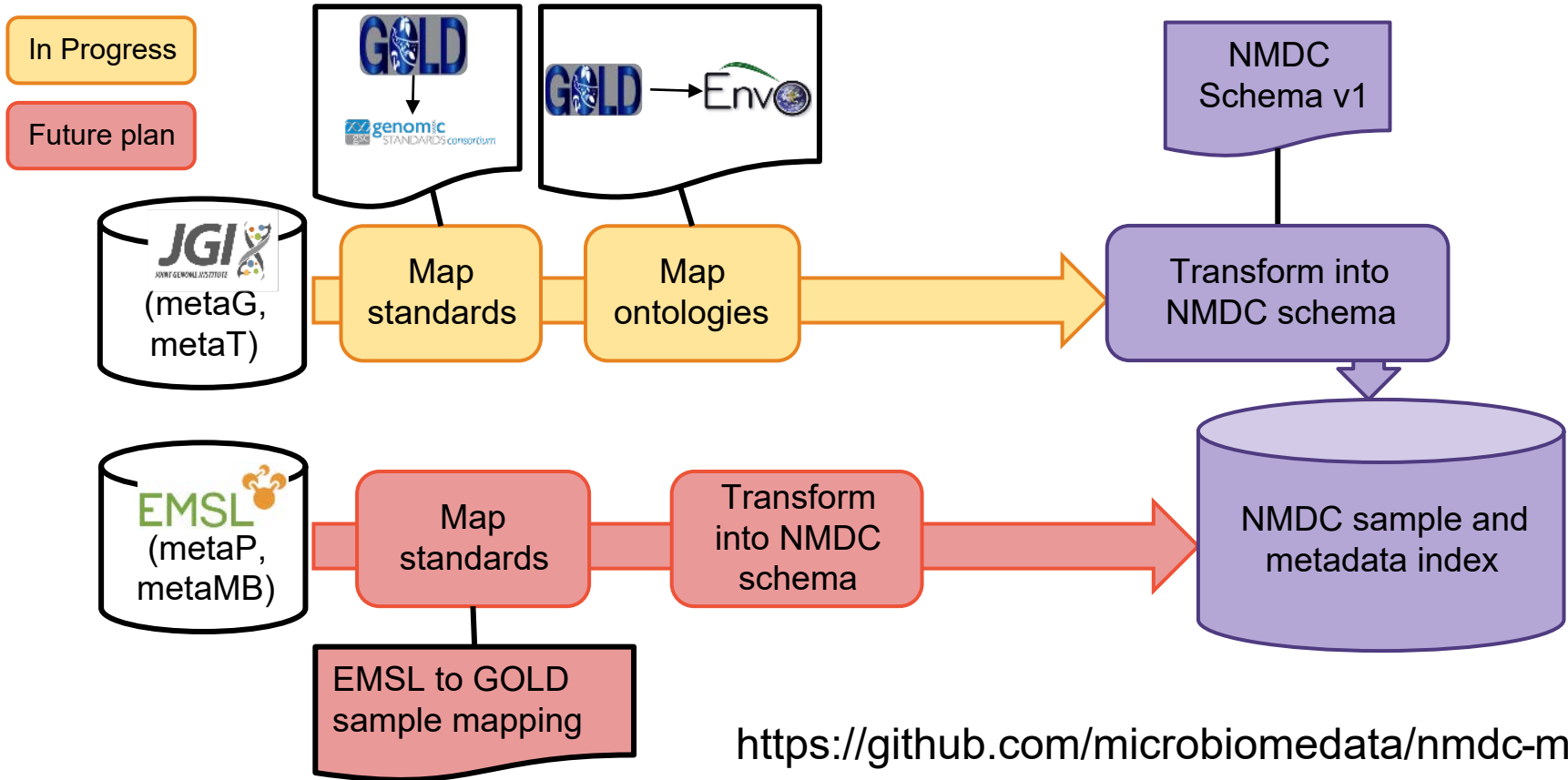
Community-driven standards & ontologies



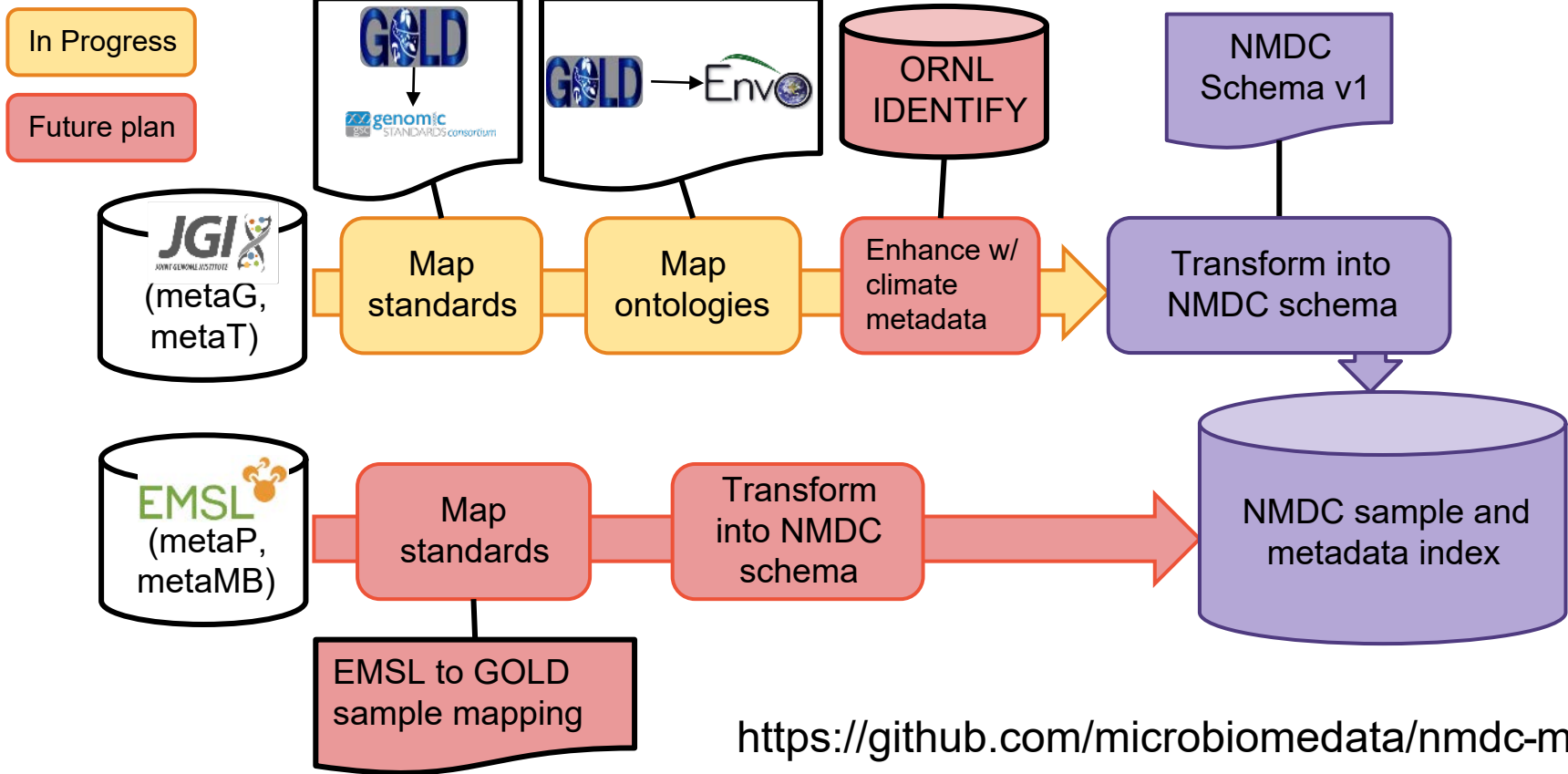
Community-driven standards & ontologies



Community-driven standards & ontologies



Community-driven standards & ontologies



Standardized pipelines & workflows



nmdc

National Microbiome
Data Collaborative

Data Types

Metagenomes



Metatranscriptomes



Metaproteomes



Metabolomes



Standardized pipelines & workflows





nmdc
National Microbiome
Data Collaborative

Data Types

Metagenomes 

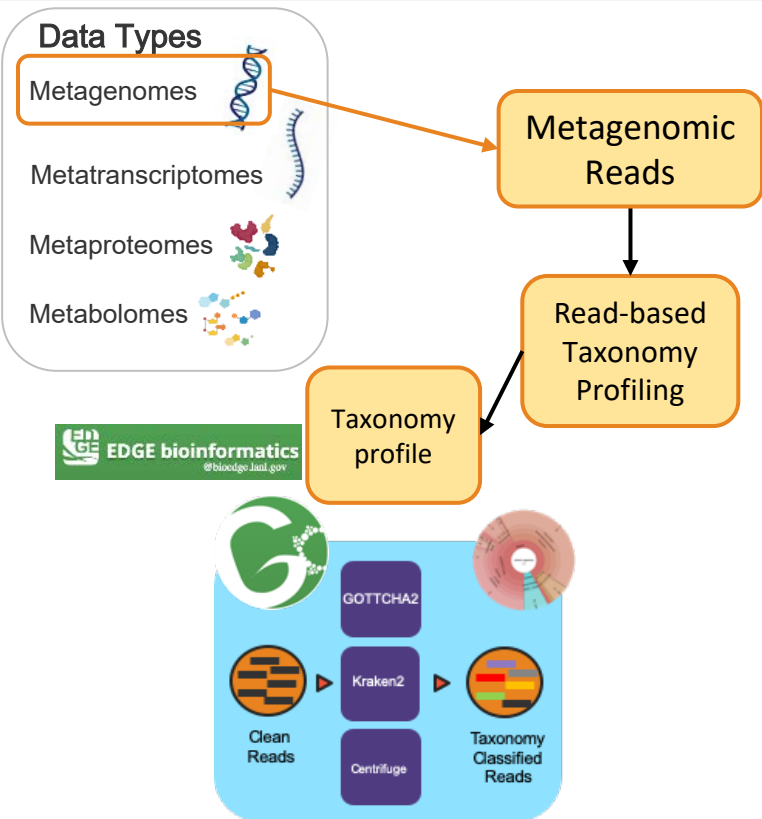
Metatranscriptomes 

Metaproteomes 

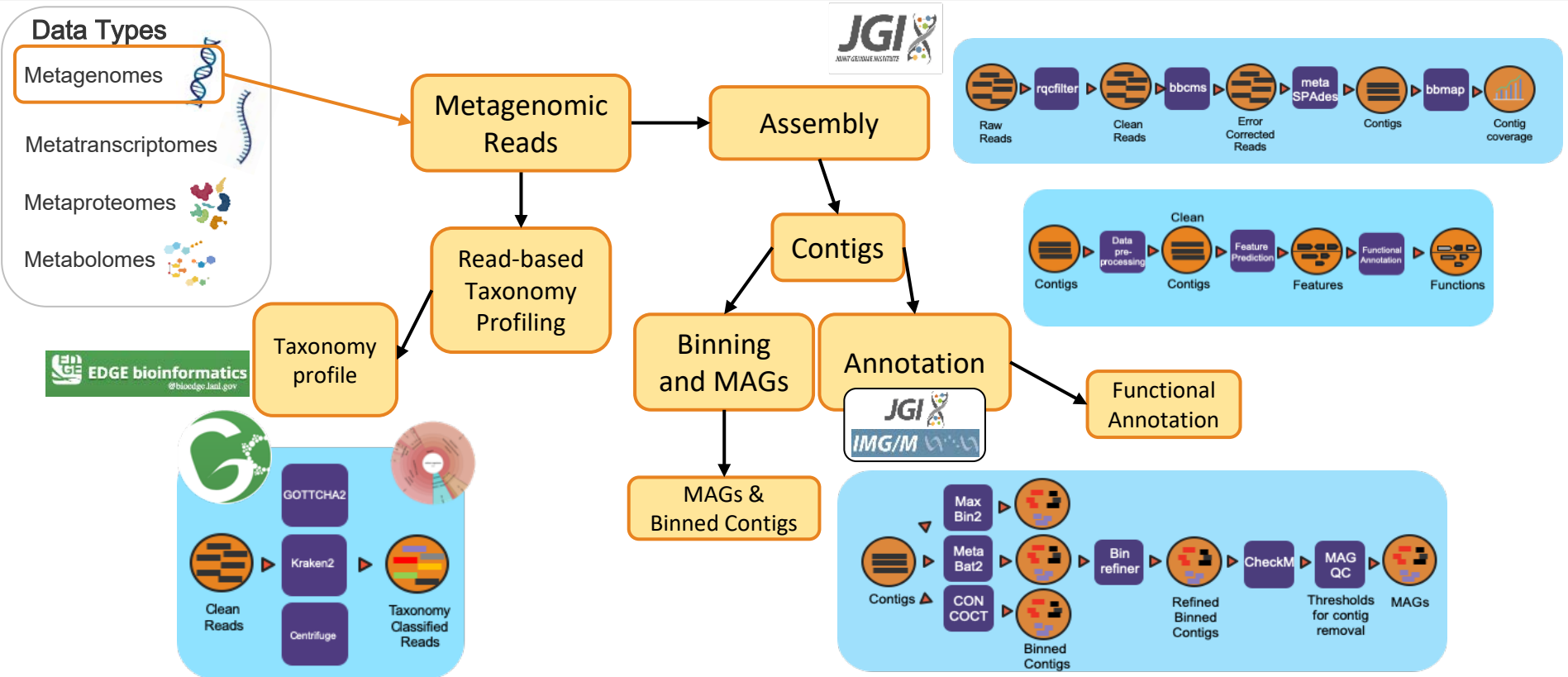
Metabolomes 

Metagenomic
Reads

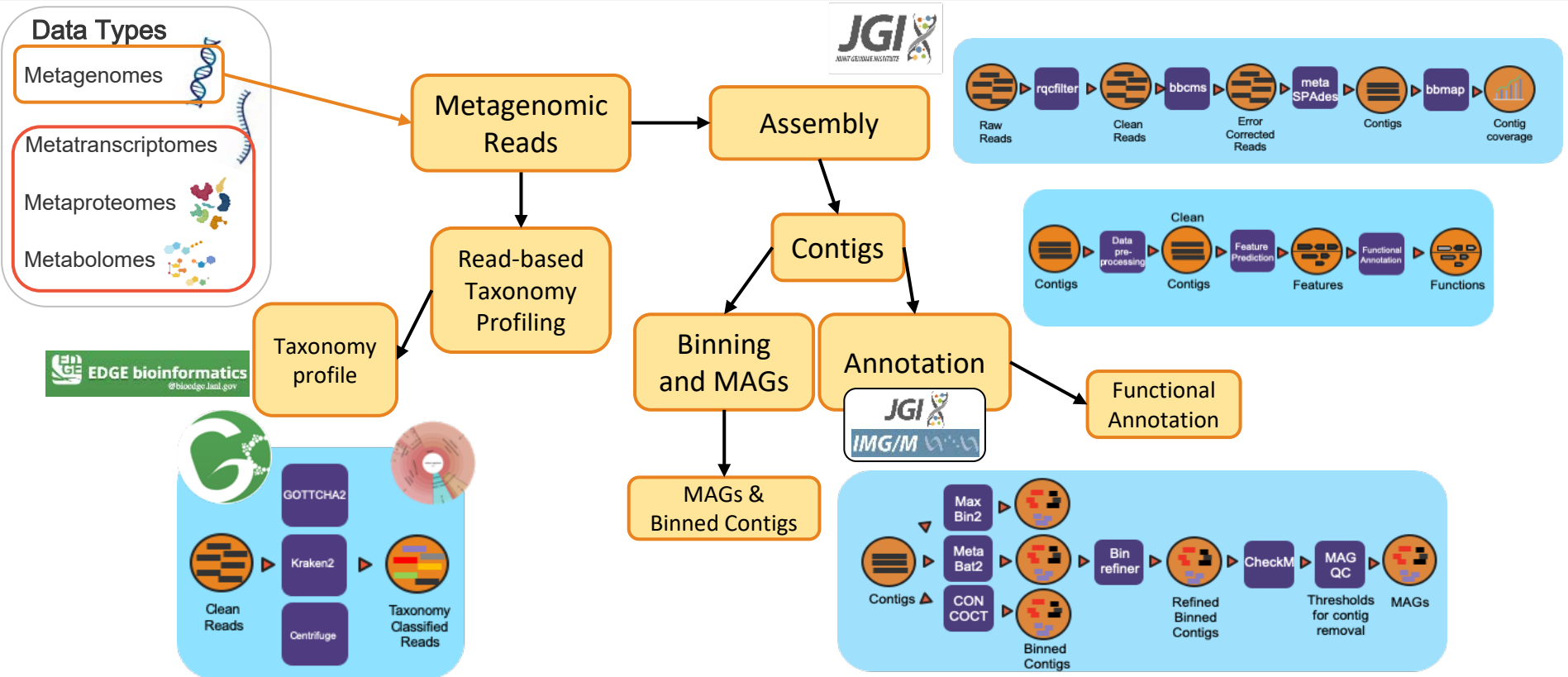
Standardized pipelines & workflows



Standardized pipelines & workflows



Standardized pipelines & workflows

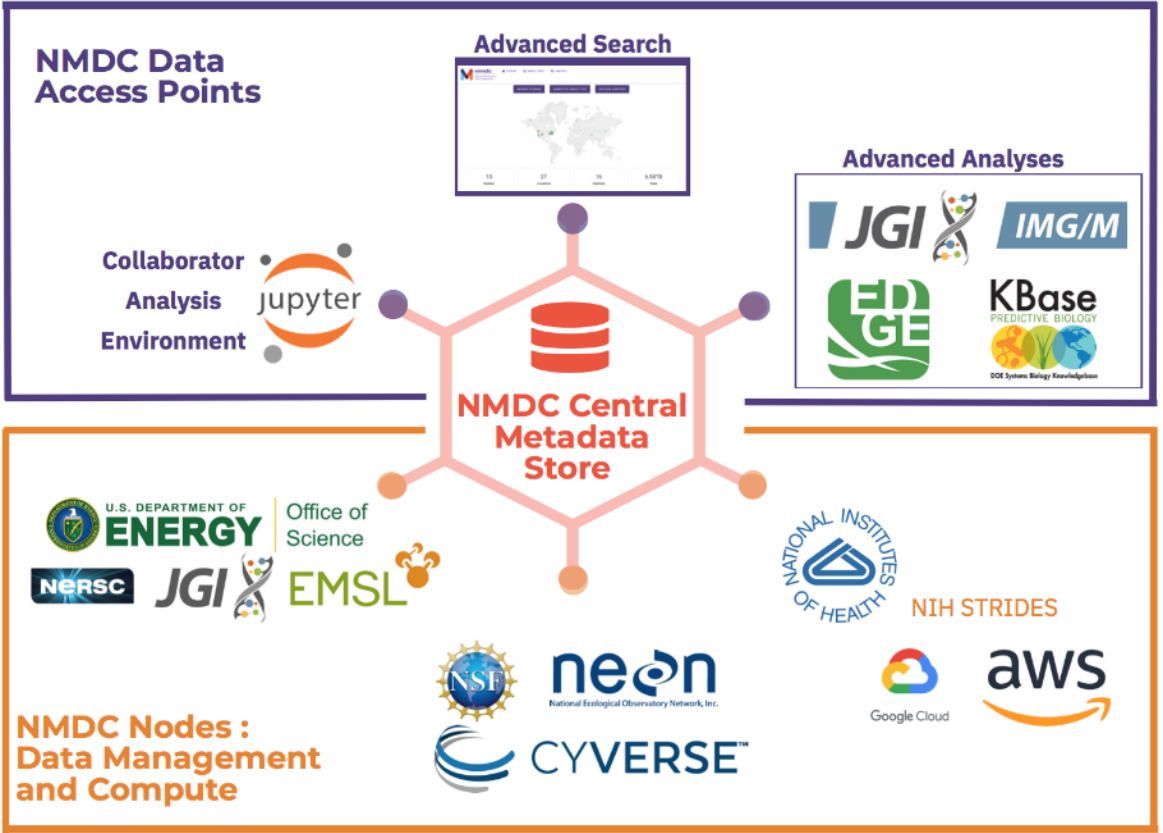


Distributed Infrastructure

Linked by a Central Metadata Store



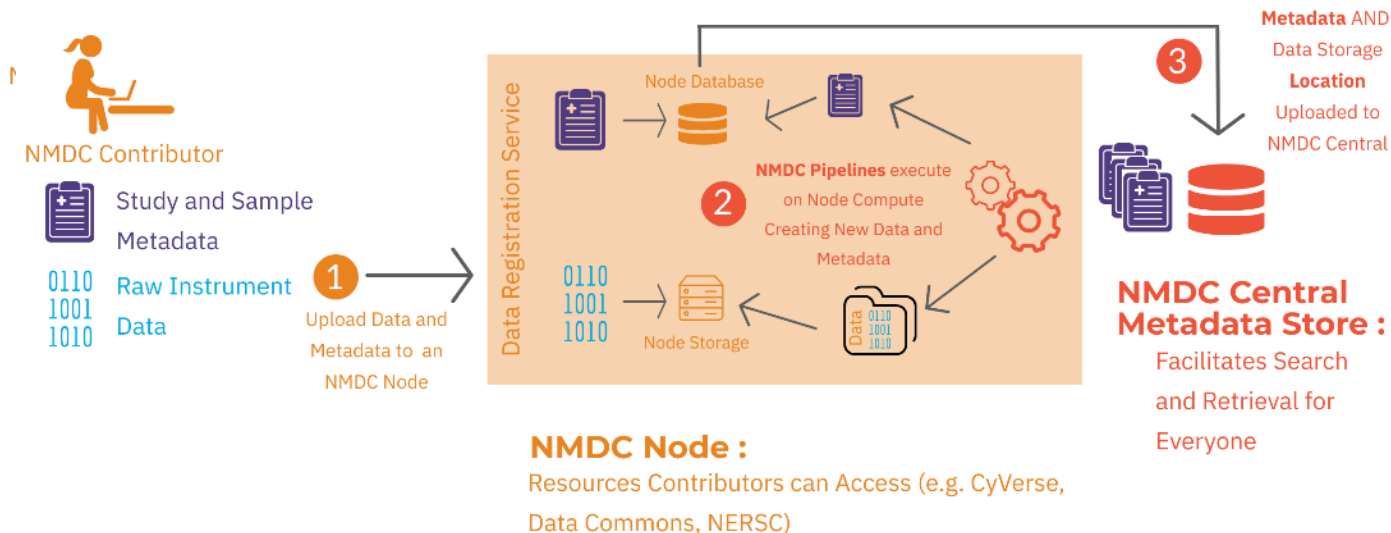
nmdc
National Microbiome
Data Collaborative



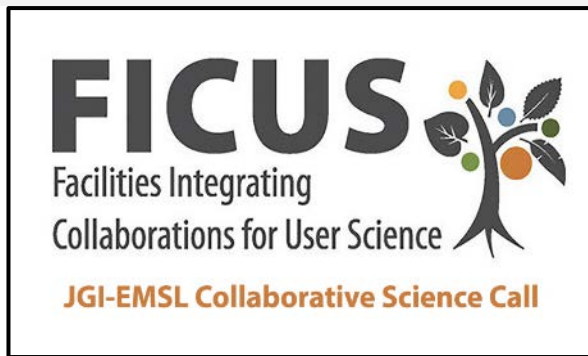


DATA FLOW IN THE NMDC

Data is at the heart of the NMDC - Findable Accessible Interoperable and Reusable (FAIR) data. So how do you contribute data to NMDC and what happens when you do?



NMDC Pilot- DOE Node



Data is at the heart of the NMDC - Findable Accessible Interoperable and Reusable (FAIR) data. So how do you contribute data to NMDC and what happens when you do?

NMDC Contributor

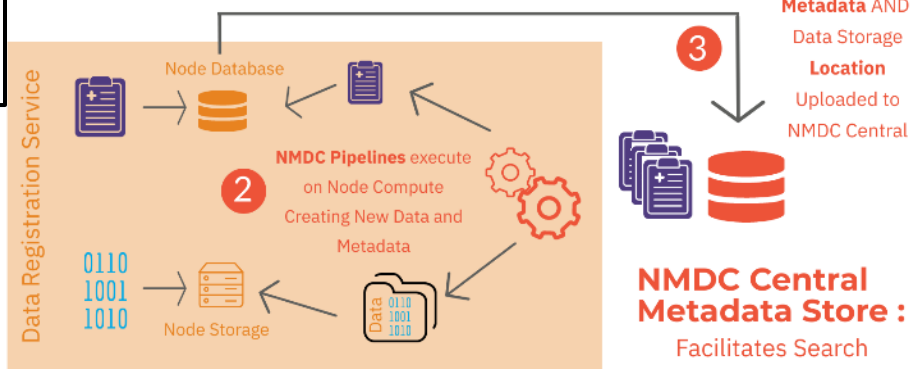


Study and Sample
Metadata



Raw Instrument
Data

1 Upload Data and
Metadata to an
NMDC Node



NMDC is a DOE-sponsored *pilot* leveraging DOE computing facilities for distributed computing, data storage, and web services infrastructure



and Retrieval for
Everyone

Data Use Policy- Open Access



NMDC policy modeled after other public data resources

- [Creative Commons 4.0 with Attribution](#)
- *NMDC data collaborators can download and use or transform the data freely, provided they cite the data appropriately*
- NMDC data contributors can see who has downloaded their data

Study info pages

Description of the study - maybe pull from the executive summary of the proposals where appropriate

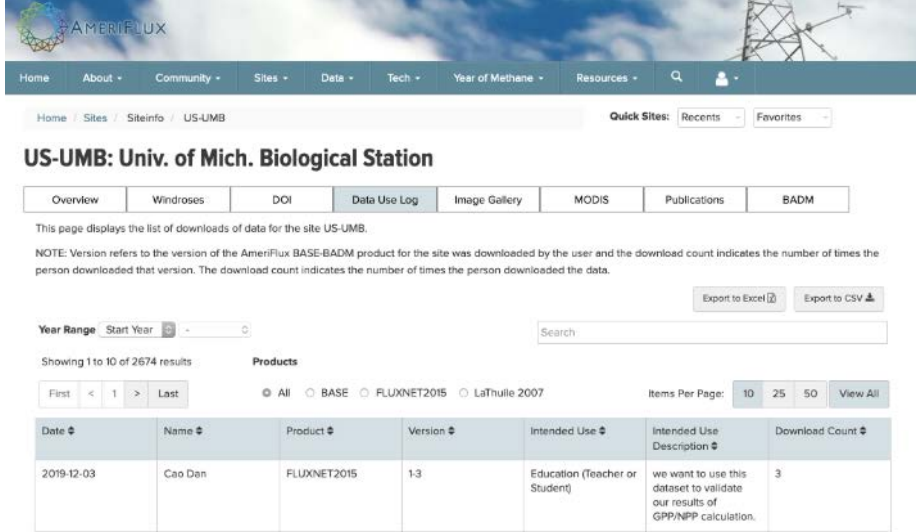
How to cite this study - call DOI to citation service (<https://citation.crosscite.org/>)

Publications associated with this study
Possibly use N4L work or ORCID -- likely to need manual curation

Data Use Policy- Open Access

NMDC policy modeled after other public data resources

- [Creative Commons 4.0 with Attribution](#)
- *NMDC data collaborators can download and use or transform the data freely, provided they cite the data appropriately*
- NMDC data contributors can see who has downloaded their data



The screenshot shows the AmeriFlux website interface. The main heading is "US-UMB: Univ. of Mich. Biological Station". Below this, there are tabs for "Overview", "Windroses", "DOI", "Data Use Log", "Image Gallery", "MODIS", "Publications", and "BADM". The "Data Use Log" tab is active, displaying a list of data downloads for the site US-UMB. A note explains that the version refers to the AmeriFlux BASE-BADM product and the download count indicates the number of times the data was downloaded. There are options to "Export to Excel" and "Export to CSV". A search bar and a "Year Range" dropdown are also visible. The table below shows the download records.

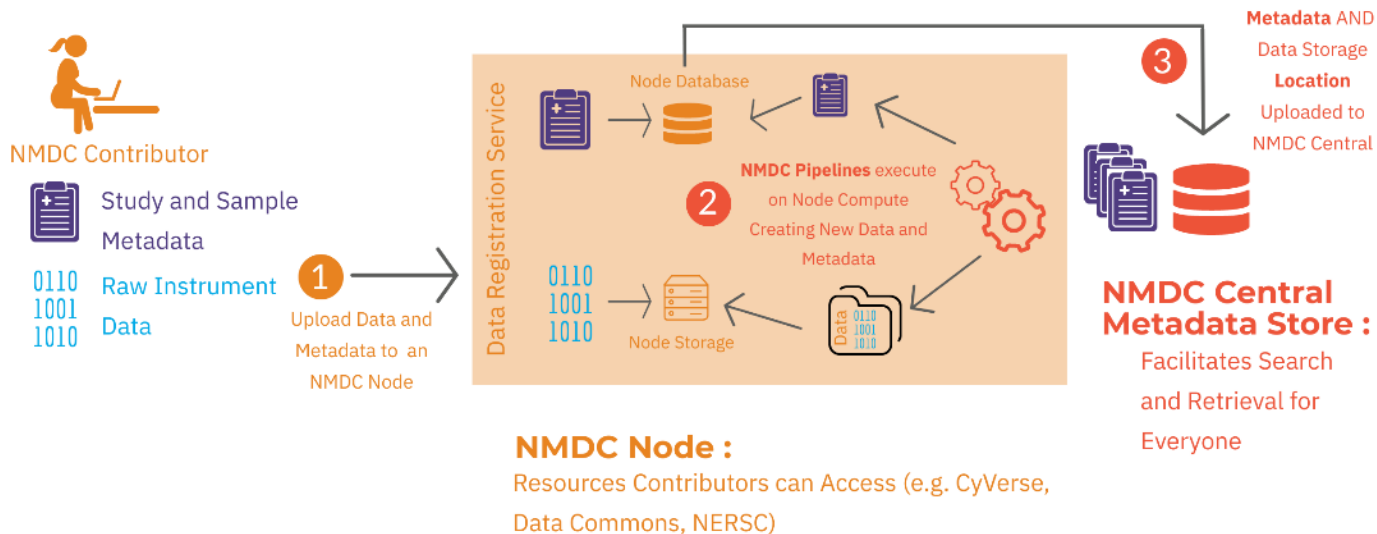
Date	Name	Product	Version	Intended Use	Intended Use Description	Download Count
2019-12-03	Cao Dan	FLUXNET2015	1-3	Education (Teacher or Student)	we want to use this dataset to validate our results of GPP/NPP calculation.	3

Interested in Contributing Data?

Contact us!

DATA FLOW IN THE NMDC

Data is at the heart of the NMDC - Findable Accessible Interoperable and Reusable (FAIR) data. So how do you contribute data to NMDC and what happens when you do?



www.microbiomedata.org
[@microbiomedata](https://twitter.com/microbiomedata)

Thank you!
kmfagnan@lbl.gov



nmdc

National Microbiome
Data Collaborative