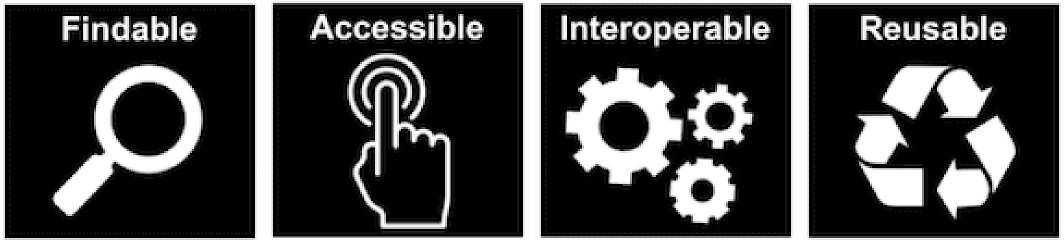
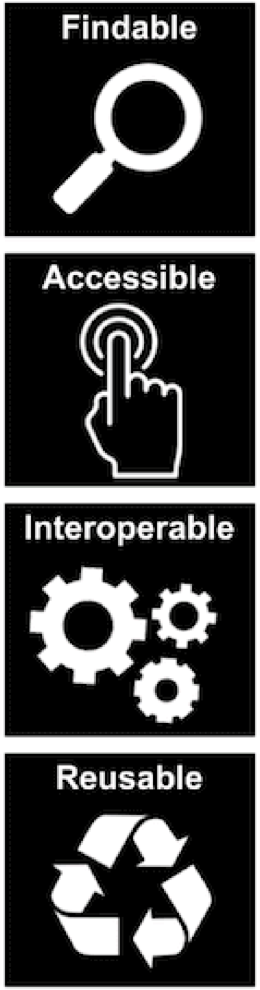
FAIR principles for ‘omic sequence datasets



A.Ponsero - 24.06.25

QIB FAIR WORKSHOPS

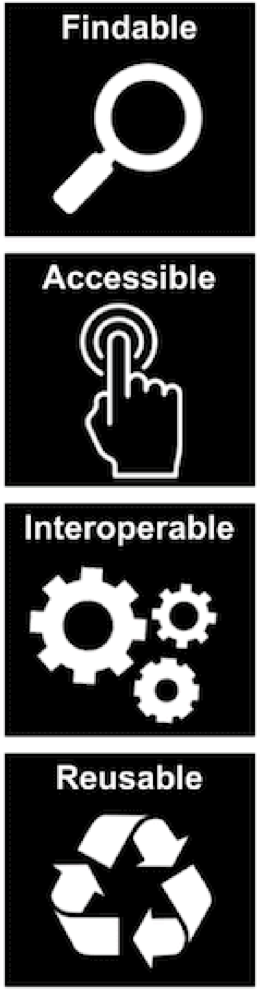
FAIR framework to enhance scientific data reusability



Since the publication of the FAIR data principles in 2016 :

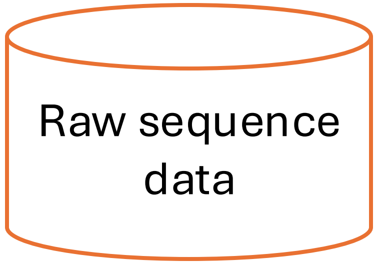
- Growing recognition of the importance of data management
- Push from institutions and funders to promote FAIR data
- Development of tools and methods for biological FAIR data

FAIR framework to enhance scientific data reusability



Since the publication of the FAIR data principles in 2016 :

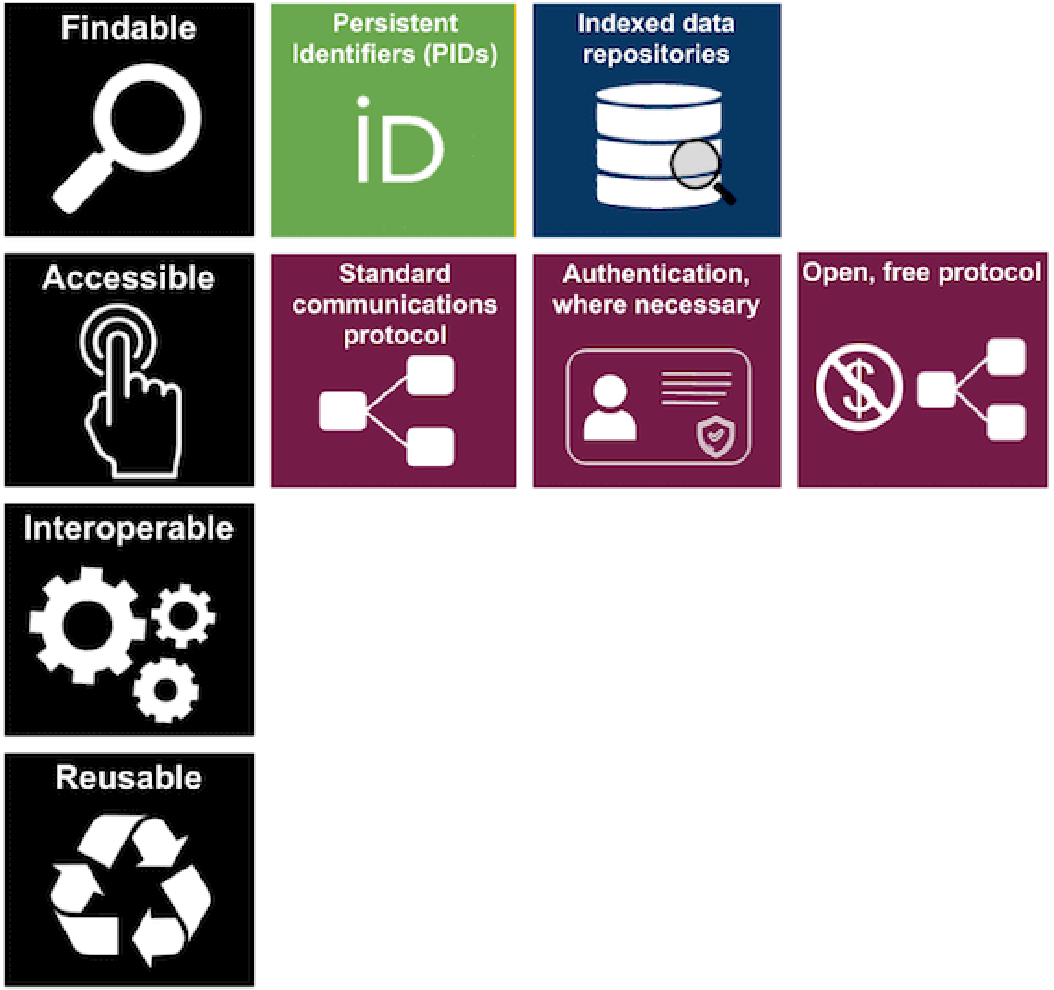
- Growing recognition of the importance of data management
- Push from institutions and funders to promote FAIR data
- Development of tools and methods for biological FAIR data



Raw sequence data

What are the specificities of making ’omic data FAIR?

FAIR framework to enhance ‘omic data reusability



Findable

Persistent Identifiers (PIDs)

Indexed data repositories

Accessible

Standard communications protocol

Authentication, where necessary

Open, free protocol

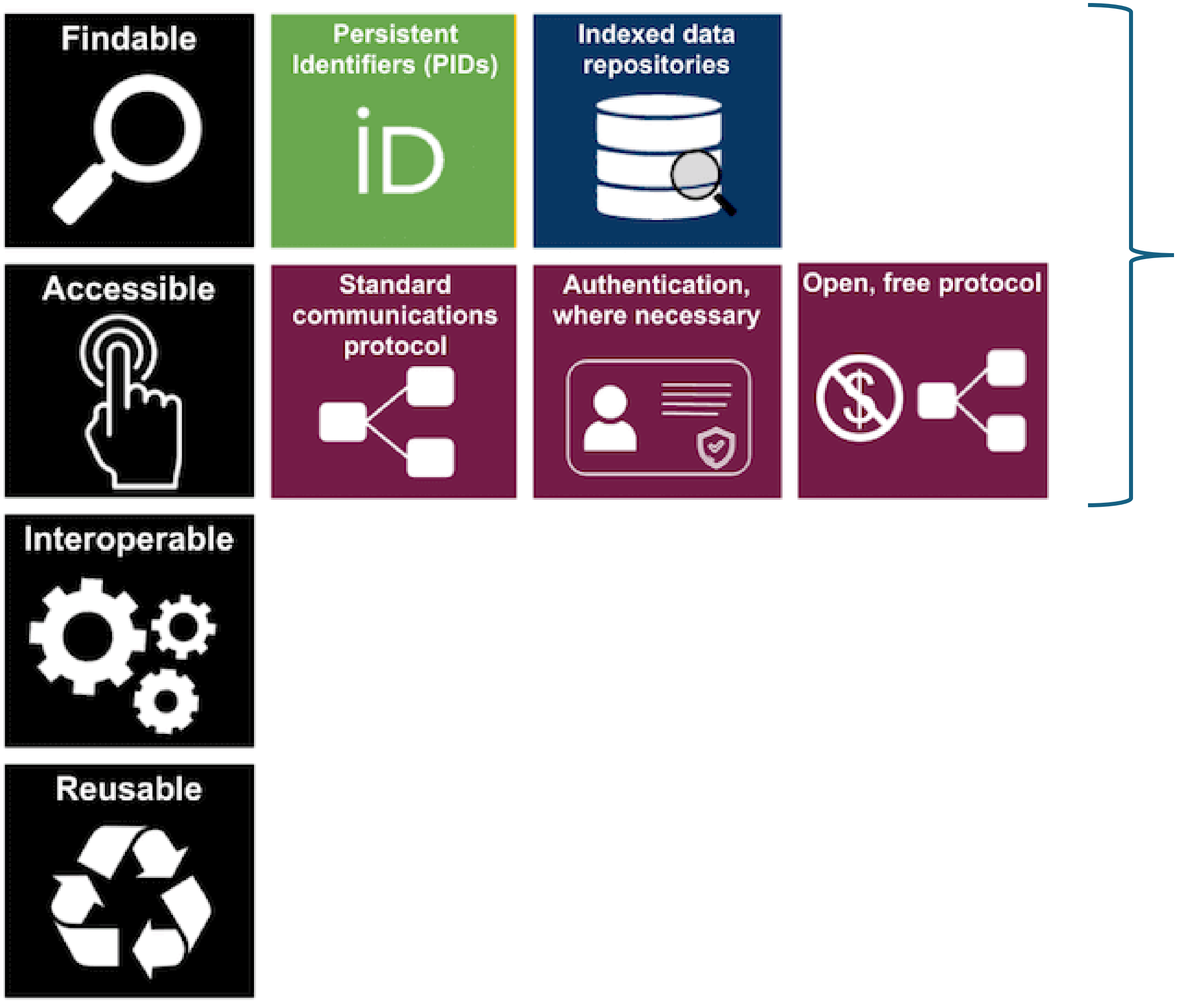
Interoperable

Reusable

Raw sequence data

Can it be found and accessed?

FAIR framework to enhance ‘omic data reusability



Findable

Persistent Identifiers (PIDs)

Indexed data repositories

Accessible

Standard communications protocol

Authentication, where necessary

Open, free protocol

Interoperable

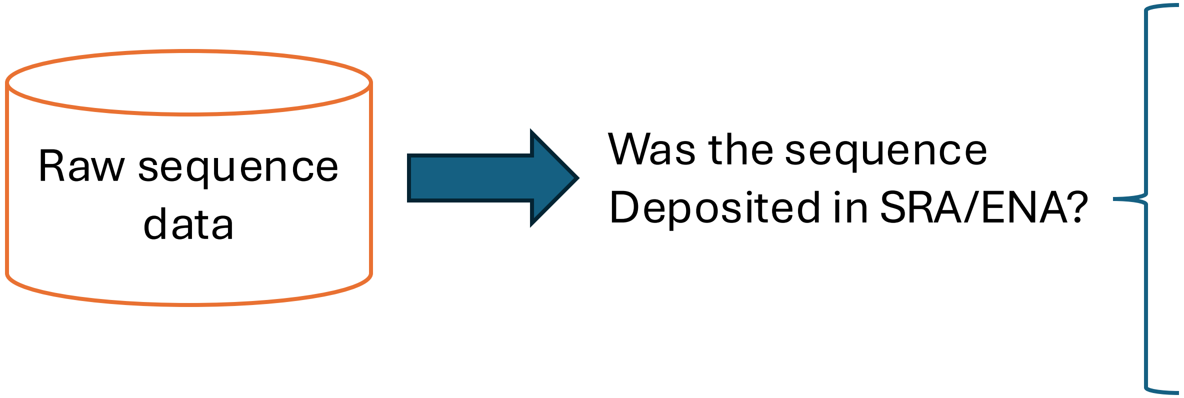
Reusable

Submission to Sequence repositories or generalist archives (Zenodo, Figshare…)

Raw sequence data

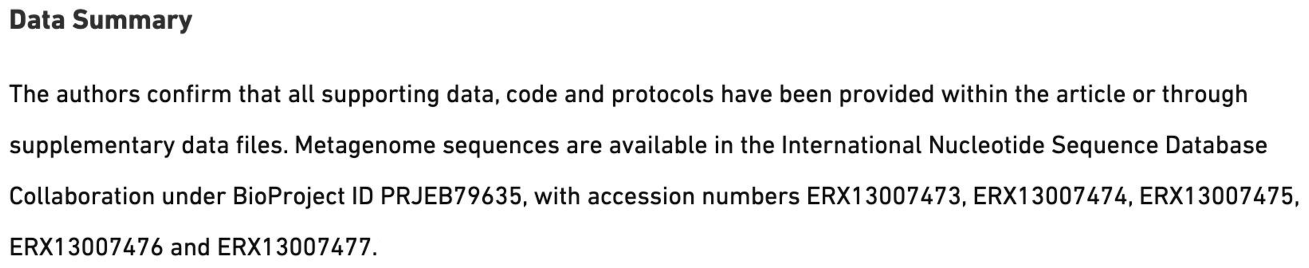
Was the sequence Deposited in SRA/ENA?

The promise vs reality of 'Omic Data



Raw sequence data

Was the sequence Deposited in SRA/ENA?

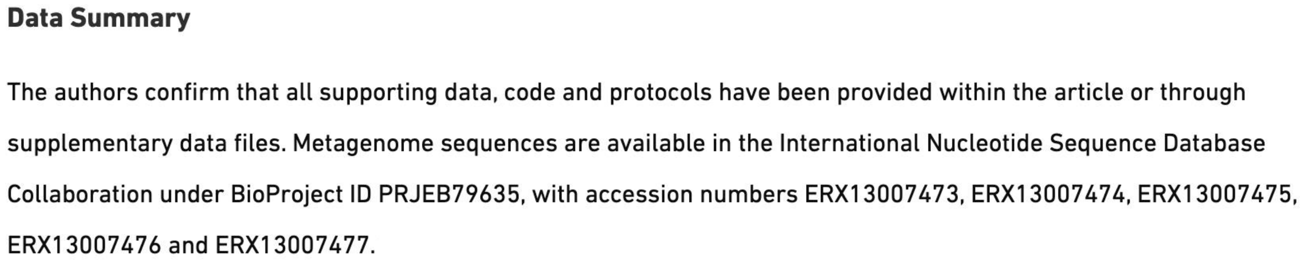


Data Summary

The authors confirm that all supporting data, code and protocols have been provided within the article or through supplementary data files. Metagenome sequences are available in the International Nucleotide Sequence Database Collaboration under BioProject ID PRJEB79635, with accession numbers ERX13007473, ERX13007474, ERX13007475, ERX13007476 and ERX13007477.

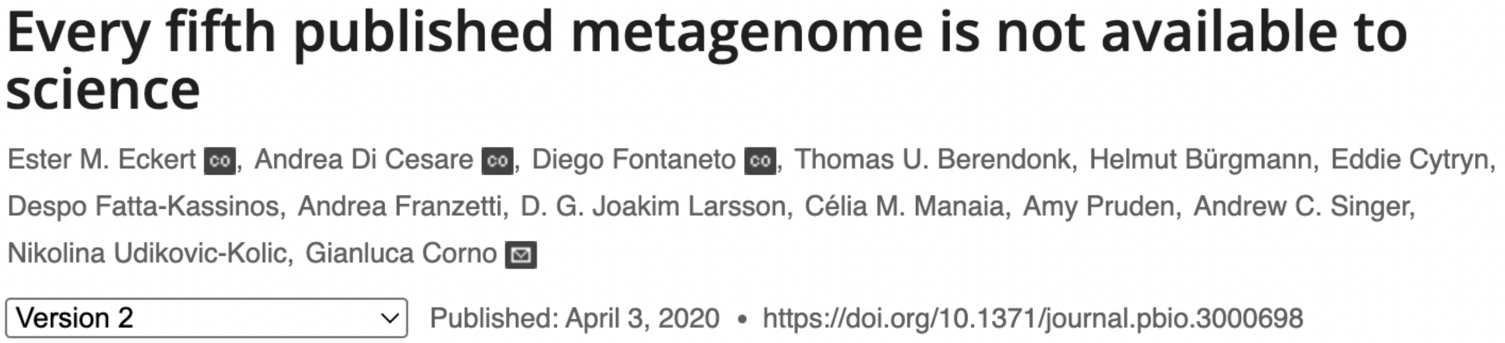
The promise vs reality of 'Omic Data

Was the sequence Deposited in SRA/ENA?



Data Summary

The authors confirm that all supporting data, code and protocols have been provided within the article or through supplementary data files. Metagenome sequences are available in the International Nucleotide Sequence Database Collaboration under BioProject ID PRJEB79635, with accession numbers ERX13007473, ERX13007474, ERX13007475, ERX13007476 and ERX13007477.



Every fifth published metagenome is not available to science

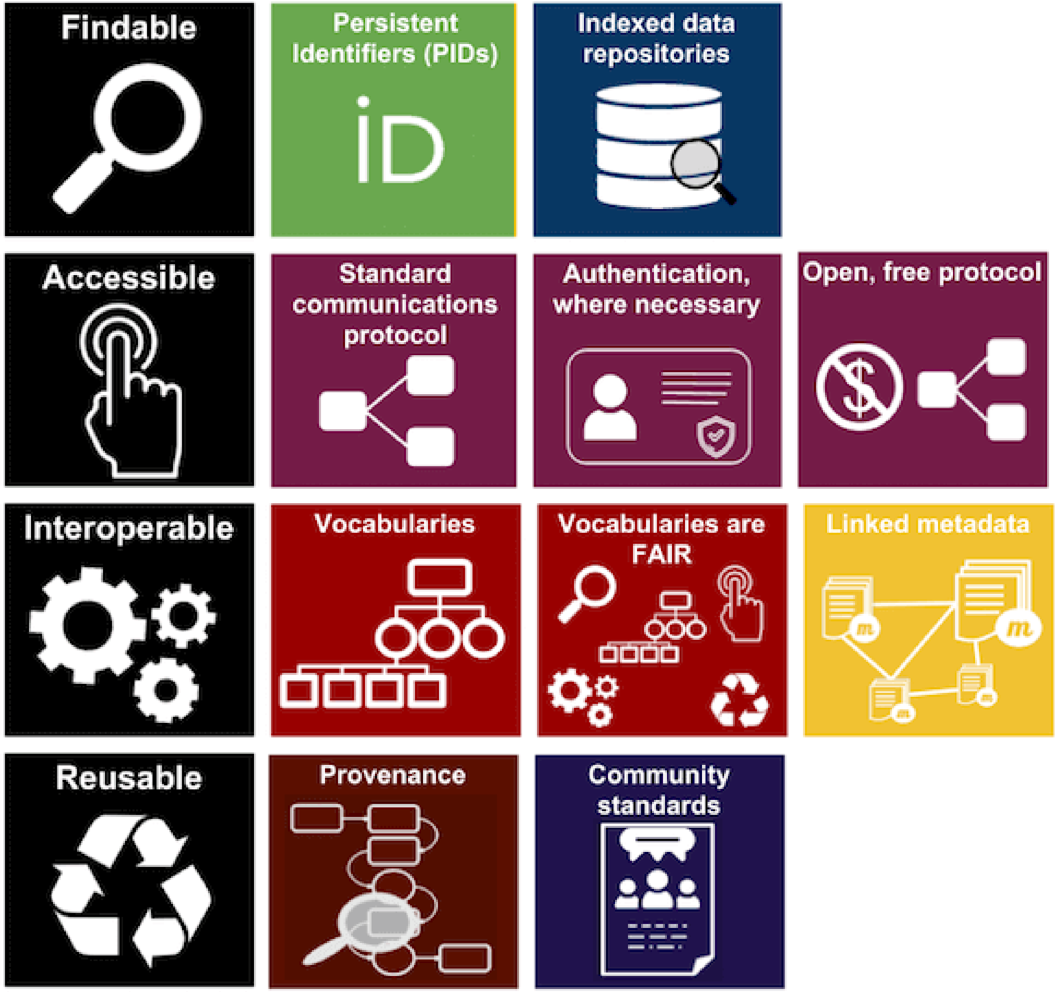
Ester M. Eckert, Andrea Di Cesare, Diego Fontaneto, Thomas U. Berendonk, Helmut Bürgmann, Eddie Cytryn, Despo Fatta-Kassinos, Andrea Franzetti, D. G. Joakim Larsson, Célia M. Manaia, Amy Pruden, Andrew C. Singer, Nikolina Udikovic-Kolic, Gianluca Corno

Version 2 · Published: April 3, 2020 • https://doi.org/10.1371/journal.pbio.3000698

13% of papers don't mention availability

8% have broken/nonexistent links

FAIR framework to enhance ‘omic data reusability



Findable | Persistent Identifiers (PIDs) | Indexed data repositories

Accessible | Standard communications protocol | Authentication, where necessary | Open, free protocol

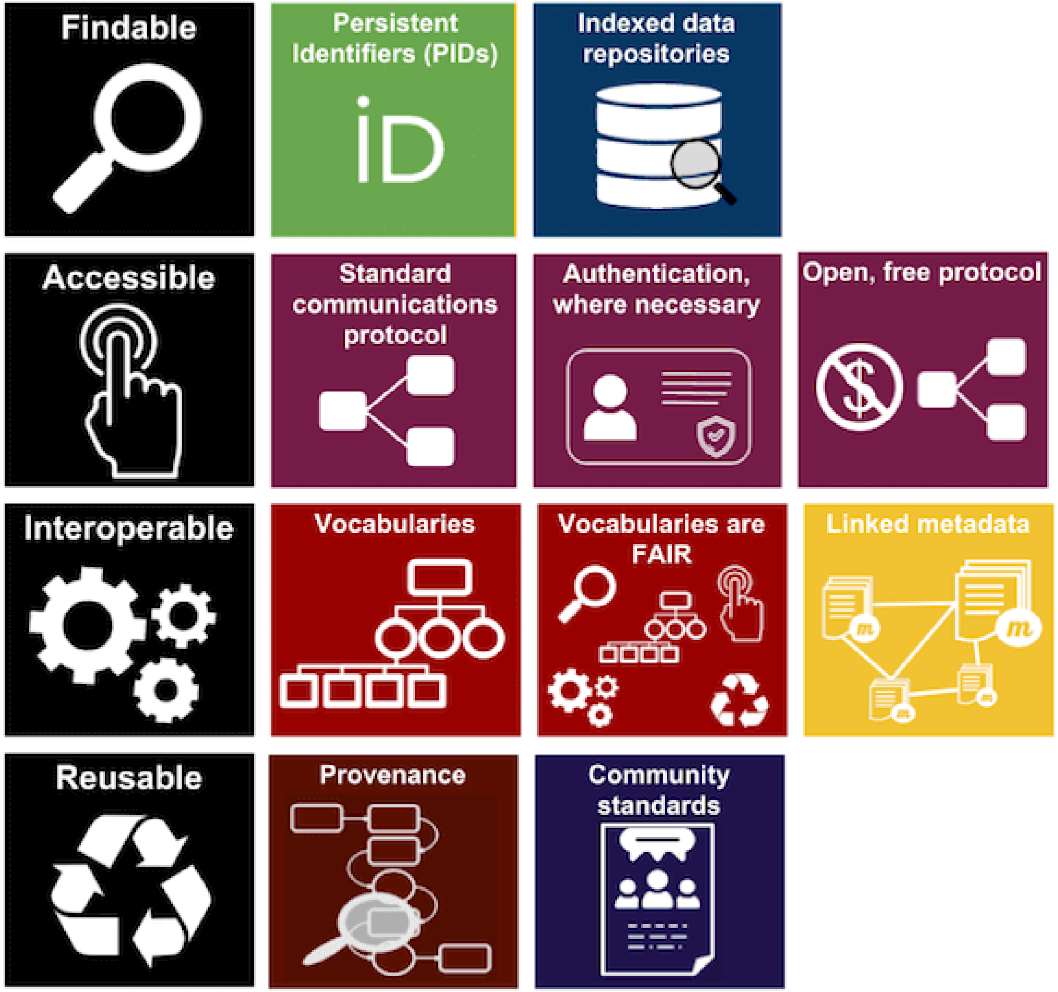
Interoperable | Vocabularies | Vocabularies are FAIR | Linked metadata

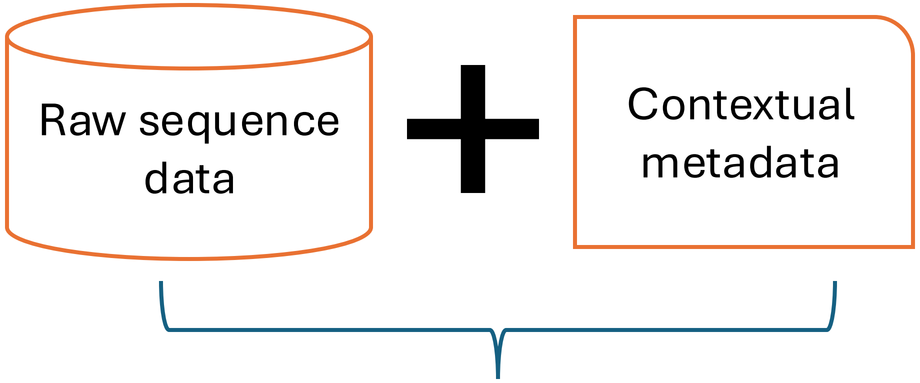
Reusable | Provenance | Community standards

Raw sequence data

Where do the sample/isolate come from?
How was it collected and when?

FAIR framework to enhance ‘omic data reusability



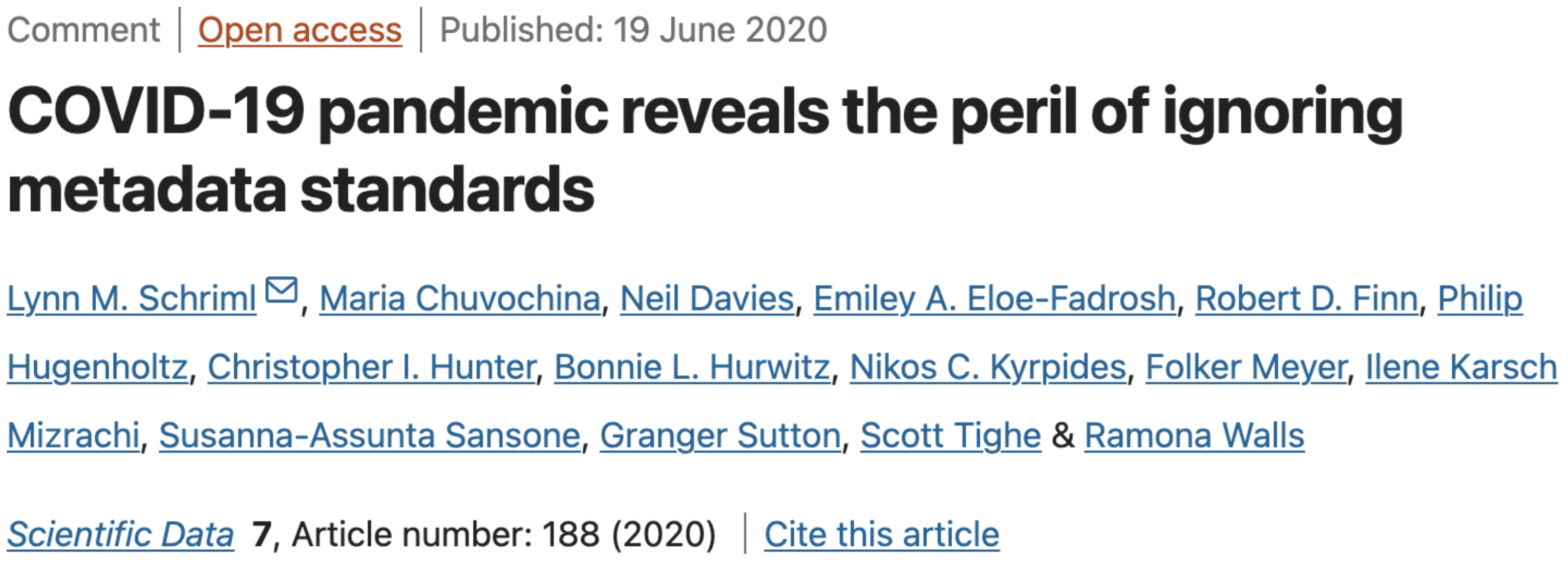

Raw sequence data

Contextual metadata

Can it be found and accessed?

Additional efforts necessary to ensure interoperability and reusability of research outputs

A metadata desert



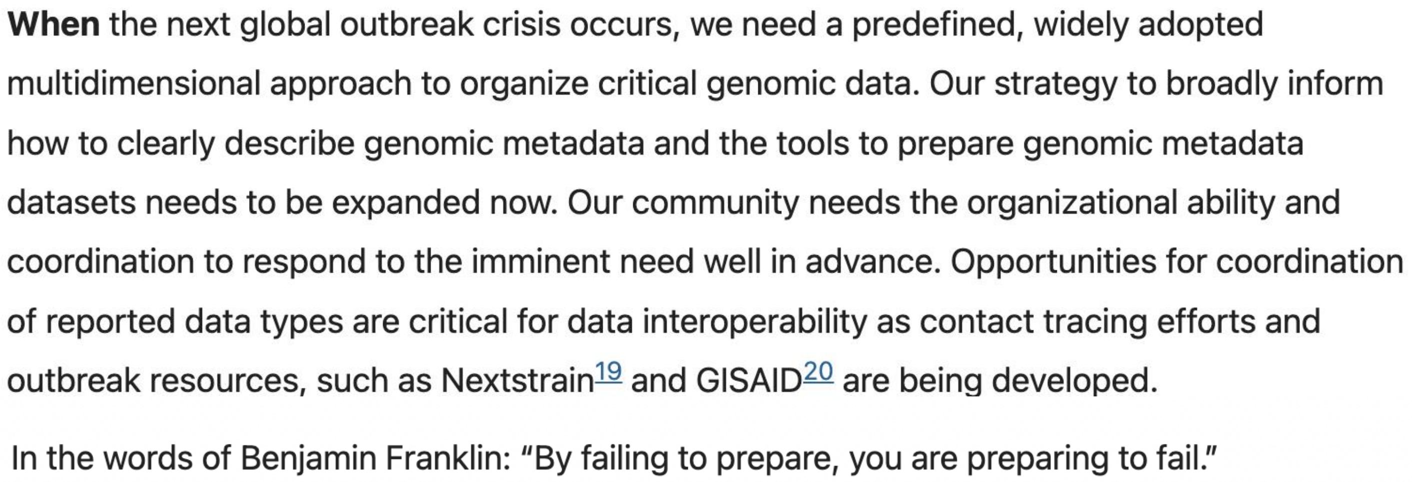
Comment | Open access | Published: 19 June 2020

COVID-19 pandemic reveals the peril of ignoring metadata standards

Lynn M. Schriml, Maria Chuvochina, Neil Davies, Emiley A. Eloe-Fadrosh, Robert D. Finn, Philip Hugenholtz, Christopher I. Hunter, Bonnie L. Hurwitz, Nikos C. Kyrpides, Folker Meyer, Ilene Karsch Mizrachi, Susanna-Assunta Sansone, Granger Sutton, Scott Tighe & Ramona Walls

Scientific Data **7**, Article number: 188 (2020) | Cite this article

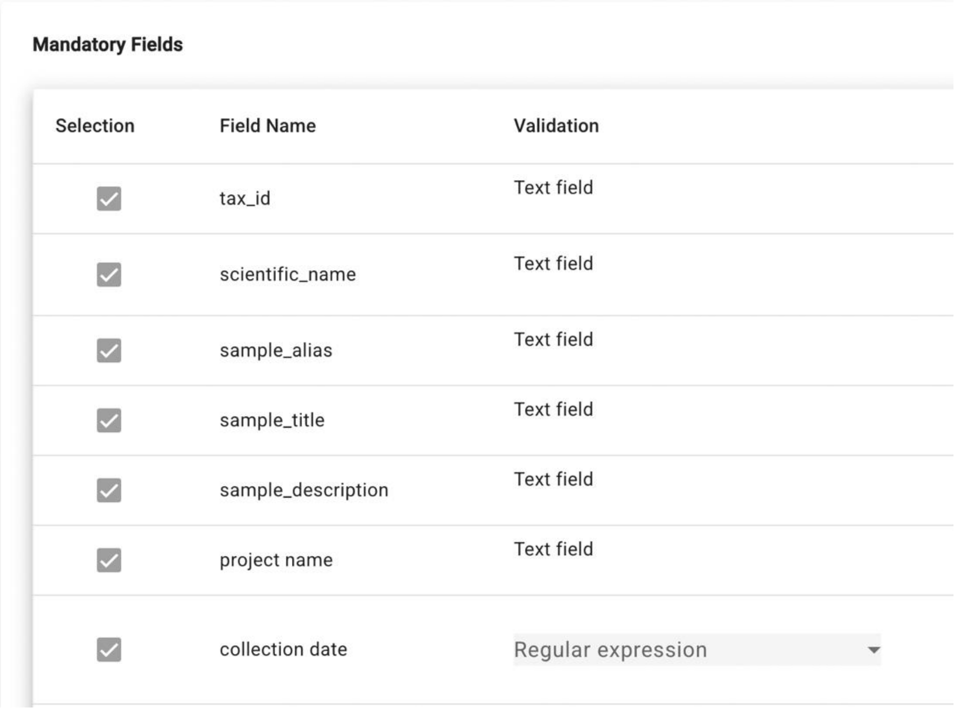
less than 33% of 2.1 million COVID genomes had basic contextual metadata



When the next global outbreak crisis occurs, we need a predefined, widely adopted multidimensional approach to organize critical genomic data. Our strategy to broadly inform how to clearly describe genomic metadata and the tools to prepare genomic metadata datasets needs to be expanded now. Our community needs the organizational ability and coordination to respond to the imminent need well in advance. Opportunities for coordination of reported data types are critical for data interoperability as contact tracing efforts and outbreak resources, such as Nextstrain[19] and GISAID[20] are being developed.

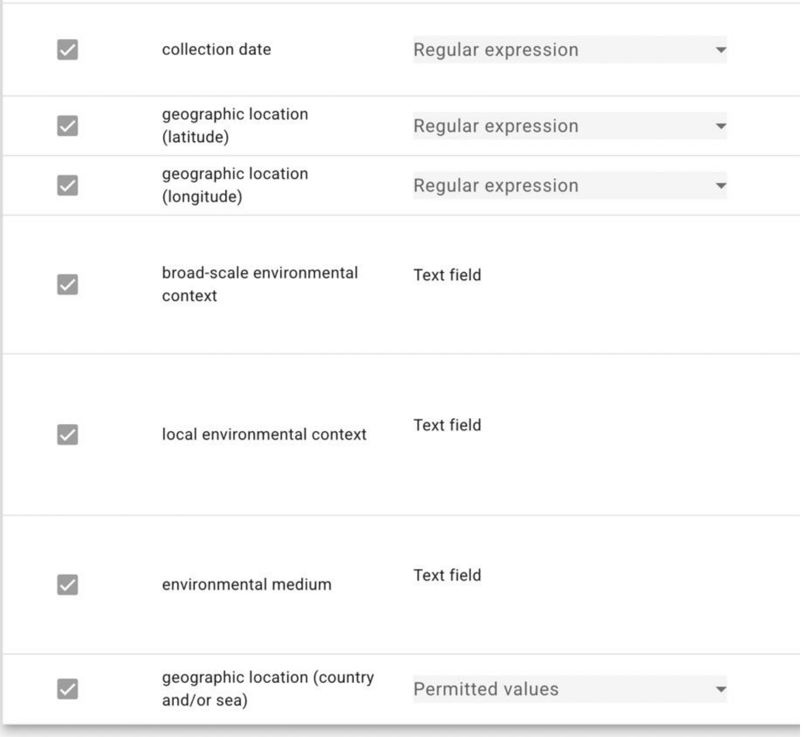
In the words of Benjamin Franklin: "By failing to prepare, you are preparing to fail."

When Standard Frameworks Don't Fit Complex Science



Mandatory Fields

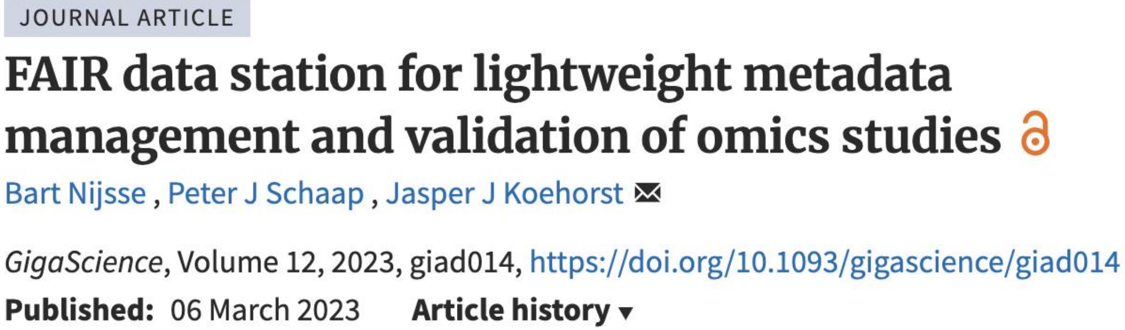
Selection	Field Name	Validation
☑	tax_id	Text field
☑	scientific_name	Text field
☑	sample_alias	Text field
☑	sample_title	Text field
☑	sample_description	Text field
☑	project name	Text field
☑	collection date	Regular expression ▾



☑	collection date	Regular expression ▾
☑	geographic location (latitude)	Regular expression ▾
☑	geographic location (longitude)	Regular expression ▾
☑	broad-scale environmental context	Text field
☑	local environmental context	Text field
☑	environmental medium	Text field
☑	geographic location (country and/or sea)	Permitted values ▾

Human gut MixS

- Limited contextual metadata data check
- Difficult to capture the complexity of study variables



JOURNAL ARTICLE

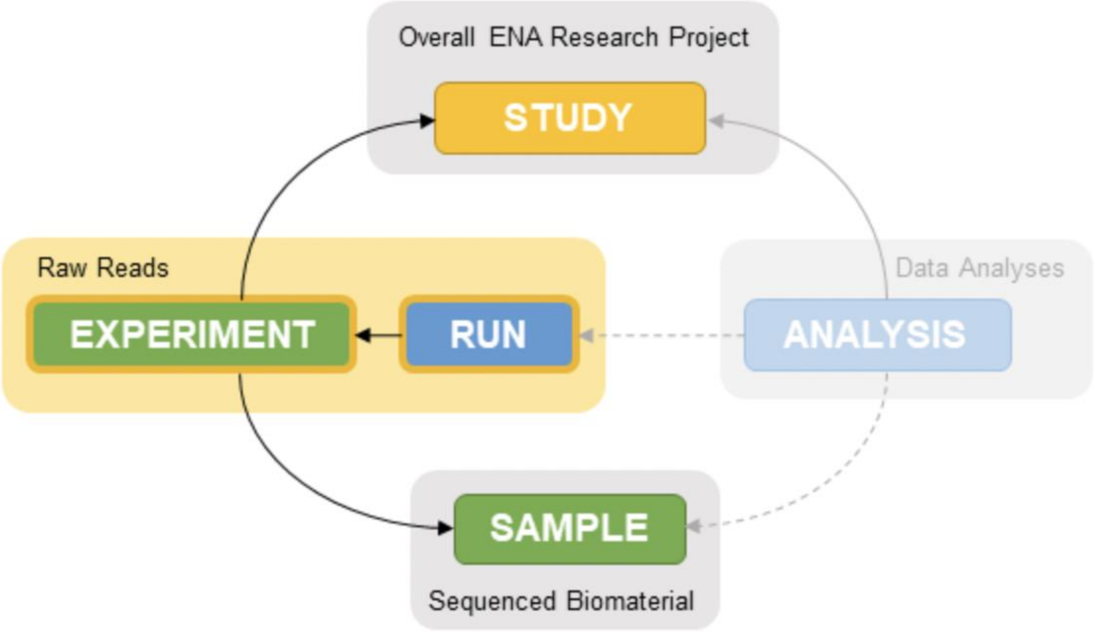
FAIR data station for lightweight metadata management and validation of omics studies

Bart Nijsse , Peter J Schaap , Jasper J Koehorst ✉

GigaScience, Volume 12, 2023, giad014, https://doi.org/10.1093/gigascience/giad014

Published: 06 March 2023 **Article history ▾**

When Standard Frameworks Don't Fit Complex Science



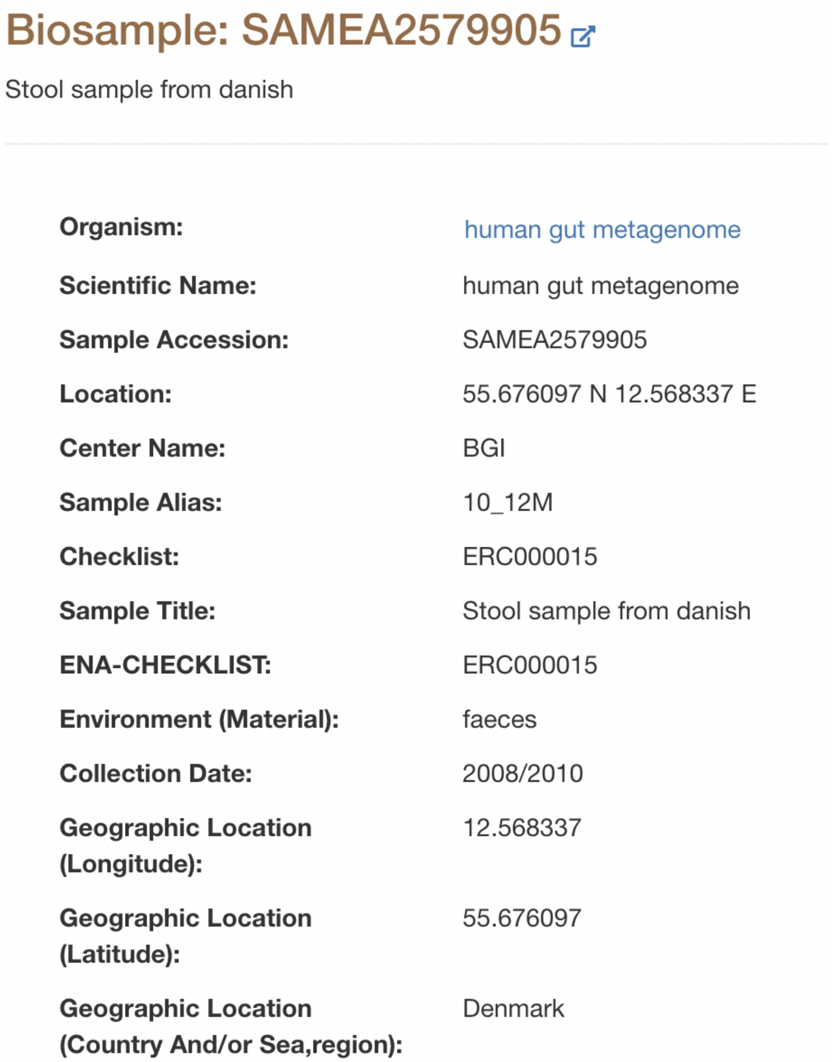
Data model design to accommodate most study design

→ Difficult for longitudinal or complex study design

→ How to link to non-sequence measurements (multi-omic)?

→ Non-sample sequences submission (blanks)?

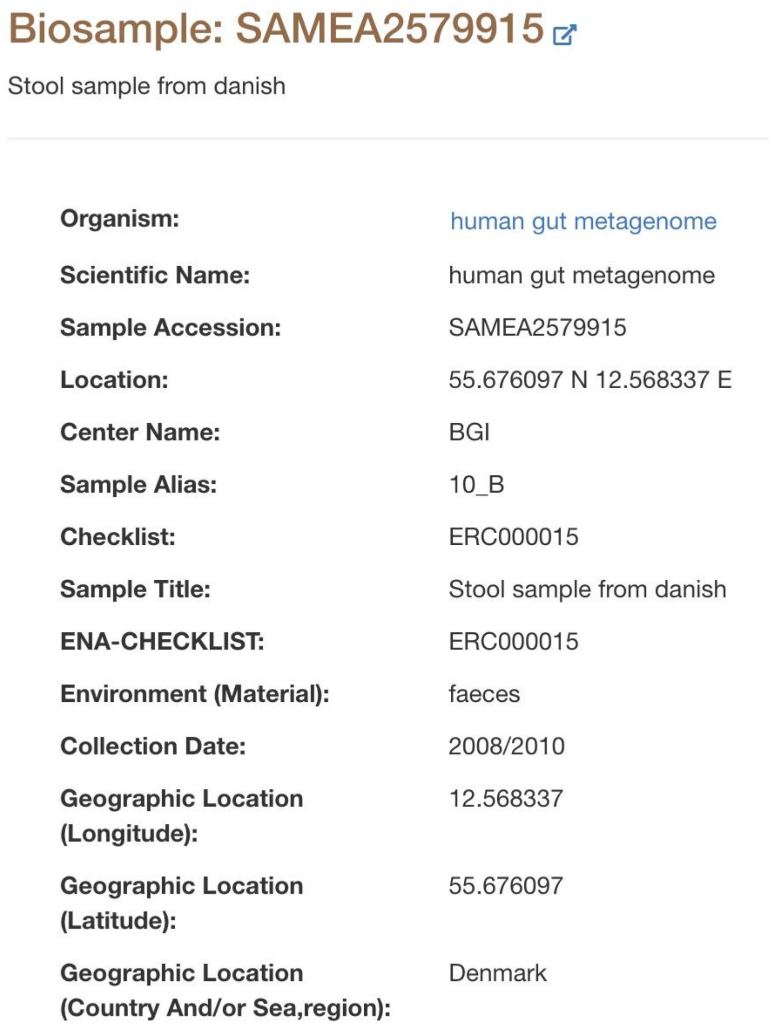
When Standard Frameworks Don't Fit Complex Science



Biosample: SAMEA2579905

Stool sample from danish

Organism:	human gut metagenome
Scientific Name:	human gut metagenome
Sample Accession:	SAMEA2579905
Location:	55.676097 N 12.568337 E
Center Name:	BGI
Sample Alias:	10_12M
Checklist:	ERC000015
Sample Title:	Stool sample from danish
ENA-CHECKLIST:	ERC000015
Environment (Material):	faeces
Collection Date:	2008/2010
Geographic Location (Longitude):	12.568337
Geographic Location (Latitude):	55.676097
Geographic Location (Country And/or Sea,region):	Denmark



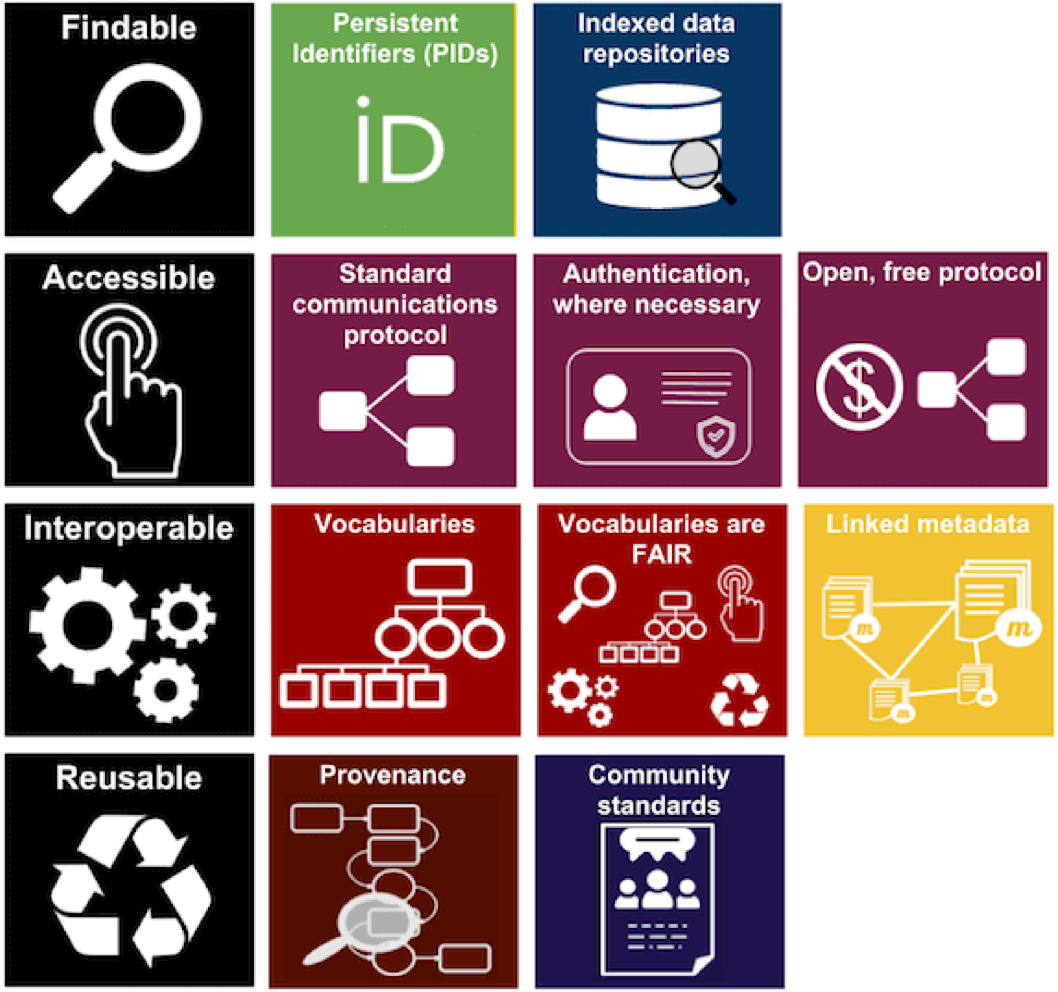
Biosample: SAMEA2579915

Stool sample from danish

Organism:	human gut metagenome
Scientific Name:	human gut metagenome
Sample Accession:	SAMEA2579915
Location:	55.676097 N 12.568337 E
Center Name:	BGI
Sample Alias:	10_B
Checklist:	ERC000015
Sample Title:	Stool sample from danish
ENA-CHECKLIST:	ERC000015
Environment (Material):	faeces
Collection Date:	2008/2010
Geographic Location (Longitude):	12.568337
Geographic Location (Latitude):	55.676097
Geographic Location (Country And/or Sea,region):	Denmark

Additional information from supplemental materials necessary to link these two samples from the same infant

FAIR framework to enhance ‘omic data reusability



Goal: not replacing standards, but enriching them to fit the institute/field needs

→ how could we improve FAIR principles implementation at QIB?

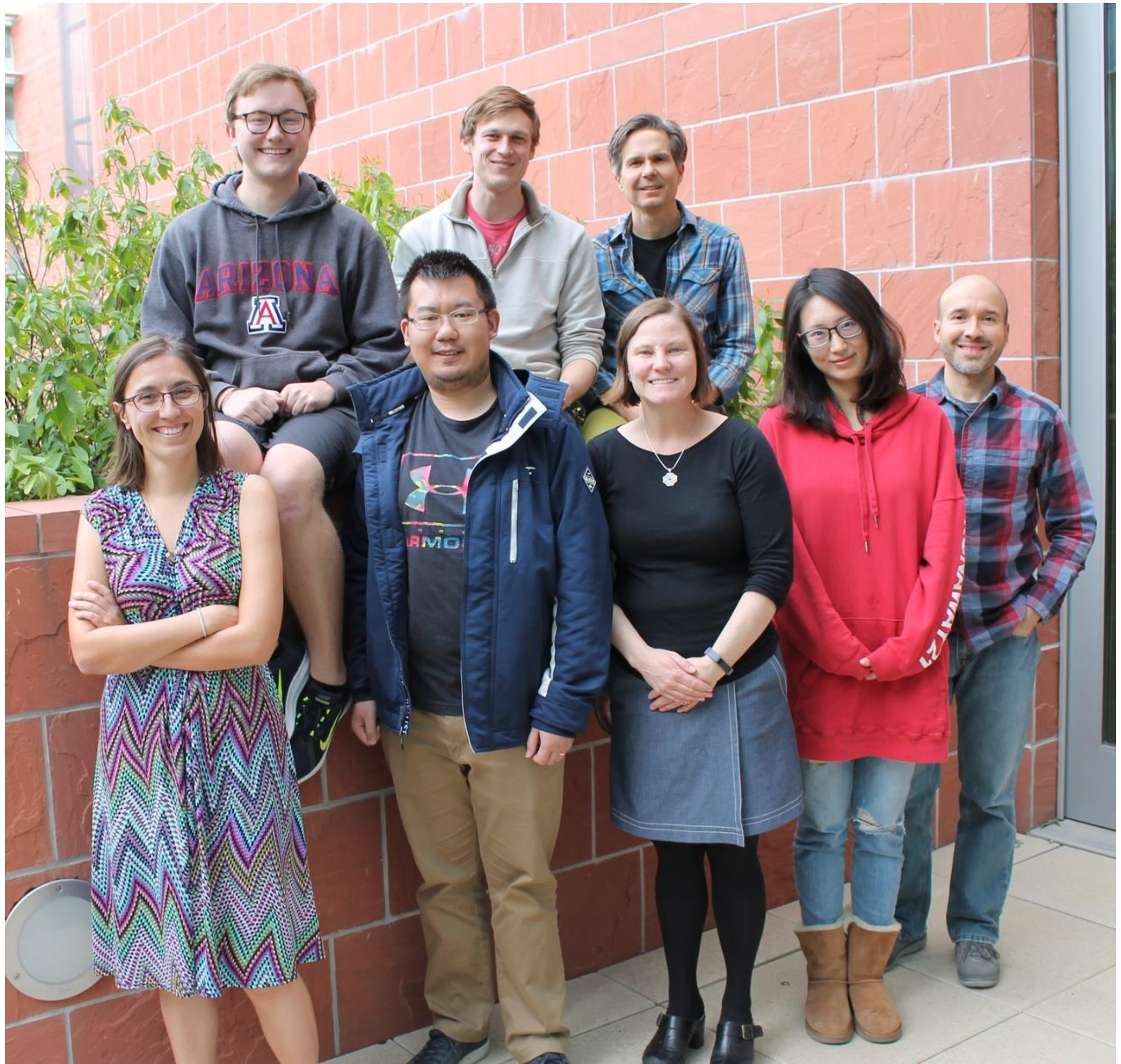
→ What are the current needs of the institute in FAIR data?

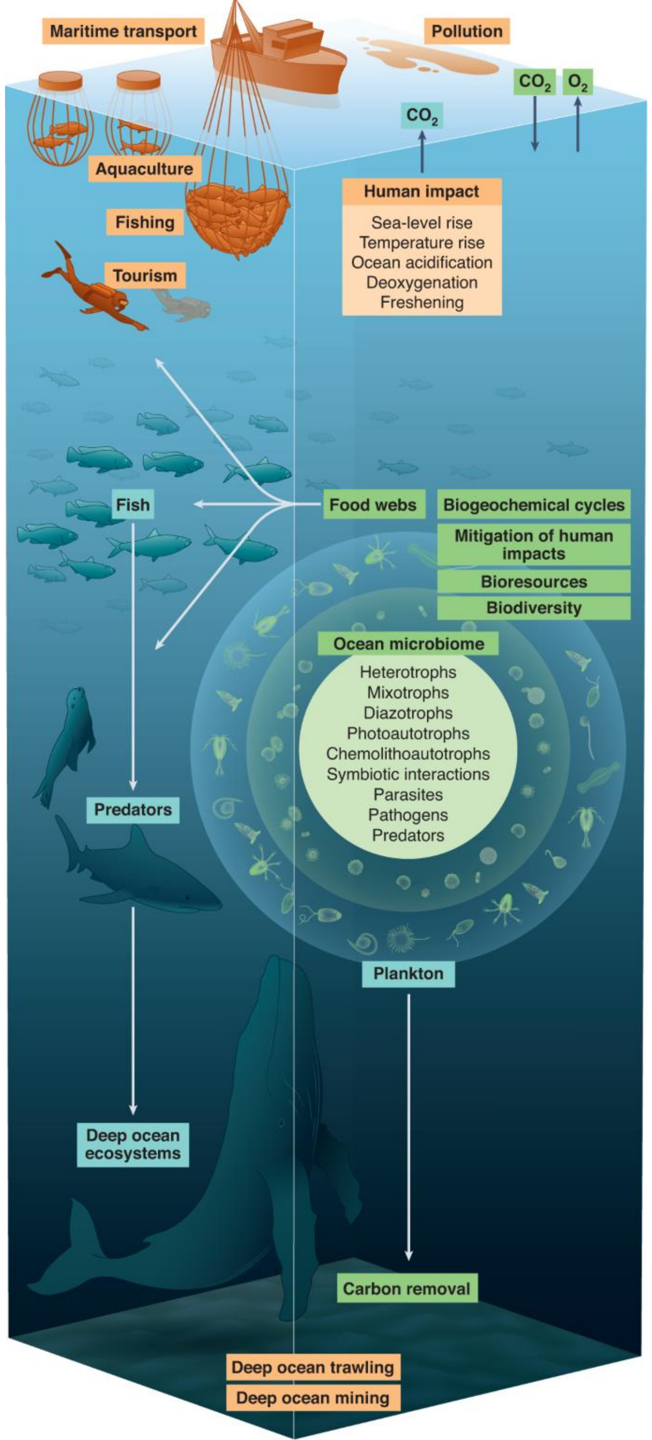
THE UNIVERSITY OF ARIZONA

Hurwitz lab

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Salonen lab





Marine environments

- Cover >70% of Earth's surface
- Largest continuous ecosystem
- Physical forces create many niches

Marine microbes

- Critical to food webs
- Drive elemental cycles
- Impact atmosphere & climate

Ecological context required

- Need well-integrated data

Great Datasets Exists

OSD

Ocean Sampling Day

C-DEBI

HAWAII OCEAN TIME-SERIES

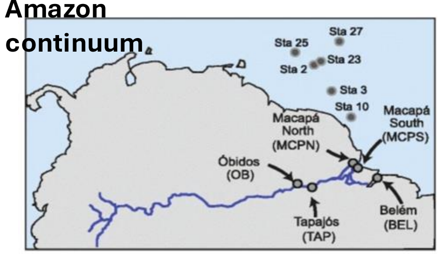
GLOBAL OCEAN FLUX STUDY

WORLD OCEAN CIRCULATION EXPERIMENT

UNIVERSITY OF HAWAII

GOS

Amazon continuum



BATS

TARA OCEANS

>30 years of sample collections

TBs of sequence data

Most oceans and sea sampled

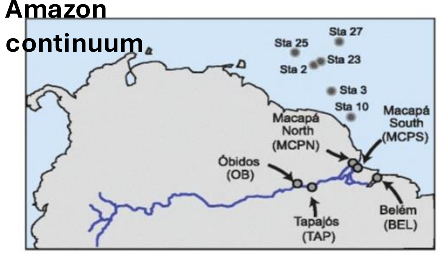
Great Datasets Exists But Can't Be Used Easily

OSD
Ocean Sampling Day

C-DEBI

HAWAII OCEAN TIME-SERIES
GLOBAL OCEAN FLUX STUDY
WORLD OCEAN CIRCULATION EXPERIMENT
UNIVERSITY OF HAWAII

Amazon continuum



GOS

BATS

TARA OCEANS

Sequence and physiochemical contextual data deposited in different repositories

Lack of common vocabularies

Different Units

Disparate data types

PlanetMicrobe

Web platform for the re-integration of -omics of historic and high-value marine datasets.

1) Harmonize

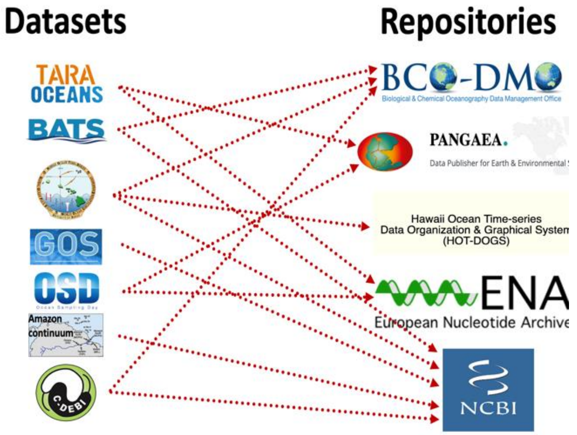
Annotate with standard vocabulary terms

2) Reconnect

Return components from separate repositories

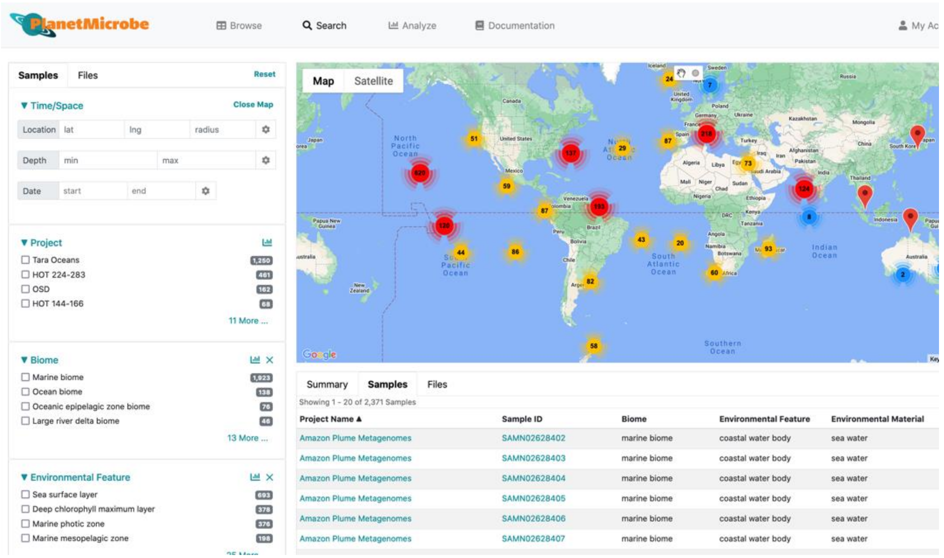
3) Load Database

Data processing, validation and unit conversion



Datasets
Repositories
TARA OCEANS
BATS
GOS
OSD
Amazon continuum
C-DEBI
BCO-DMO
Biological & Chemical Oceanography Data Management Office
PANGAEA.
Data Publisher for Earth & Environmental Science
Hawaii Ocean Time-series Data Organization & Graphical System (HOT-DOGS)
ENA
European Nucleotide Archive
NCBI

TARA OCEANS
PANGAEA.
Data Publisher for Earth & Environmental Science
ENA
European Nucleotide Archive
BCO-DMO
Biological & Chemical Oceanography Data Management Office
Hawaii Ocean Time-series Data Organization & Graphical System (HOT-DOGS)
NCBI



Samples | Files | Reset

▼ Time/Space — Close Map

Location: lat, lng, radius

Depth: min, max

Date: start, end

▼ Project

- □ Tara Oceans 1,250
- □ HOT 224-283 461
- □ OSD 162
- □ HOT 144-166 68

11 More ...

▼ Biome

- □ Marine biome 1,923
- □ Ocean biome 138
- □ Oceanic epipelagic zone biome 76
- □ Large river delta biome 46

13 More ...

▼ Environmental Feature

- □ Sea surface layer 693
- □ Deep chlorophyll maximum layer 378
- □ Marine photic zone 376
- □ Marine mesopelagic zone 198

Summary | Samples | Files

Showing 1 - 20 of 2,371 Samples

Project Name ▲	Sample ID	Biome	Environmental Feature	Environmental Material
Amazon Plume Metagenomes	SAMN02628402	marine biome	coastal water body	sea water
Amazon Plume Metagenomes	SAMN02628403	marine biome	coastal water body	sea water
Amazon Plume Metagenomes	SAMN02628404	marine biome	coastal water body	sea water
Amazon Plume Metagenomes	SAMN02628405	marine biome	coastal water body	sea water
Amazon Plume Metagenomes	SAMN02628406	marine biome	coastal water body	sea water
Amazon Plume Metagenomes	SAMN02628407	marine biome	coastal water body	sea water

Slide adapted from Blumberg 2022

Semantic harmonization

SampleID	NITRATE	Depth
TARA_1	3.082	25.6
TARA_2	0.193	125.2
TARA_3	1.967	45.7

TARA OCEANS

SampleID	NO3	Depth
BATS_1	40.02	2
BATS_2	19.2	50.2
BATS_3	36.67	75.7

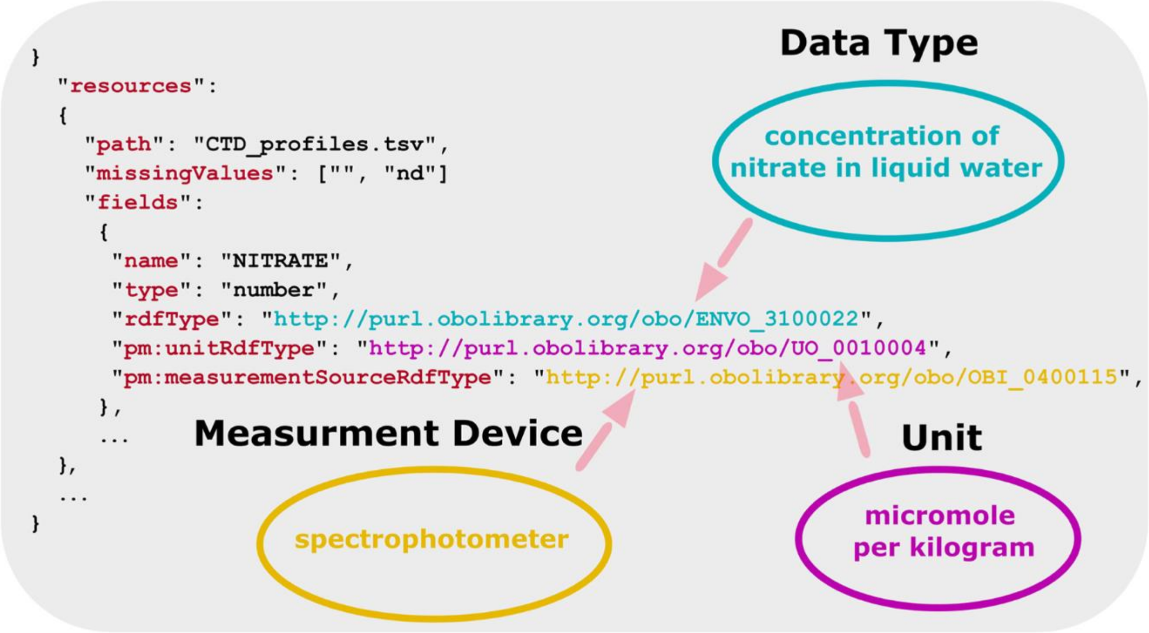
BATS

Semantic & unit harmonization

SampleID	NITRATE	Depth
TARA_1	3.082	25.6
TARA_2	0.193	125.2
TARA_3	1.967	45.7

Contextual metadata associated to the sequence (tsv)

+

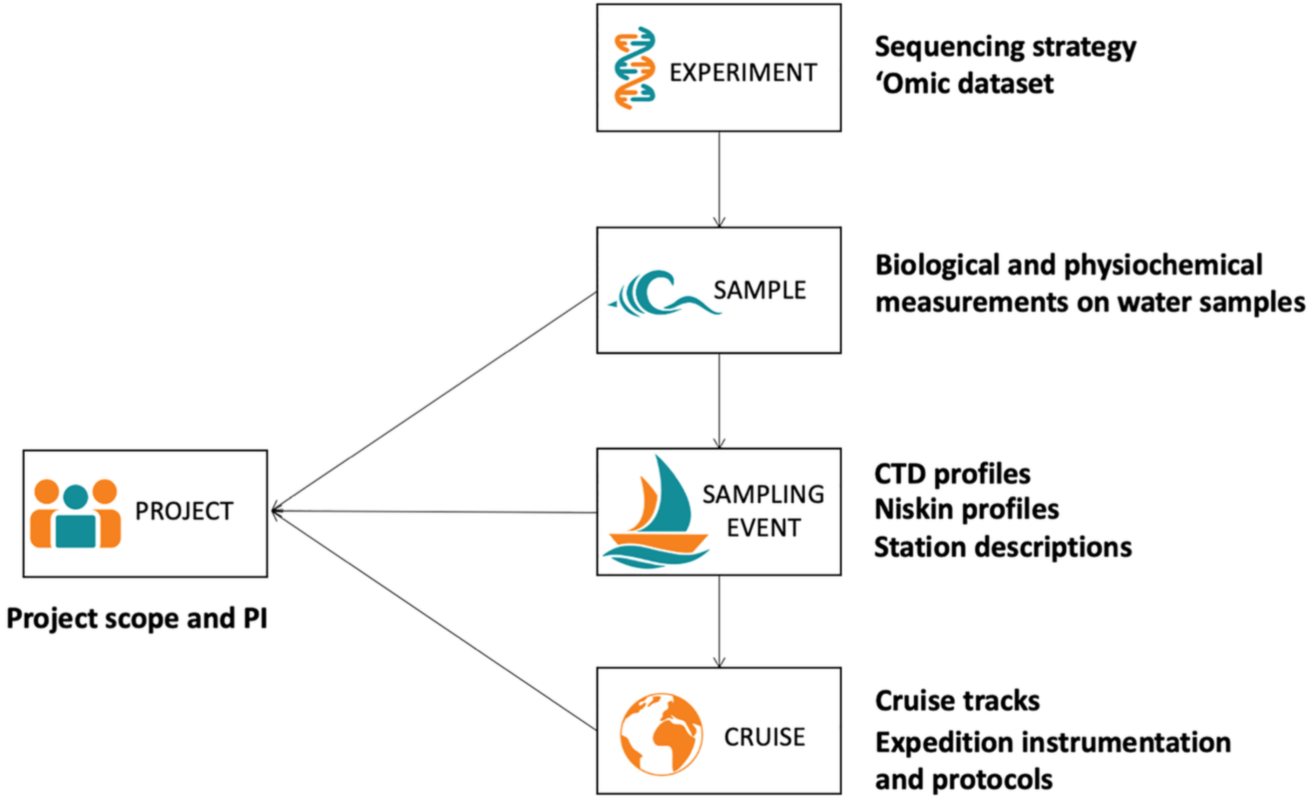


```
}
  "resources":
  {
    "path": "CTD_profiles.tsv",
    "missingValues": ["", "nd"]
    "fields":
     {
      "name": "NITRATE",
      "type": "number",
      "rdfType": "http://purl.obolibrary.org/obo/ENVO_3100022",
      "pm:unitRdfType": "http://purl.obolibrary.org/obo/UO_0010004",
      "pm:measurementSourceRdfType": "http://purl.obolibrary.org/obo/OBI_0400115",
     },
     ...
  },
  ...
}
```

Semantic layer (json)

Blumberg *et al.* 2021

Catering for a complex Data Model


EXPERIMENT

Sequencing strategy
‘Omic dataset

SAMPLE

Biological and physiochemical measurements on water samples

SAMPLING EVENT

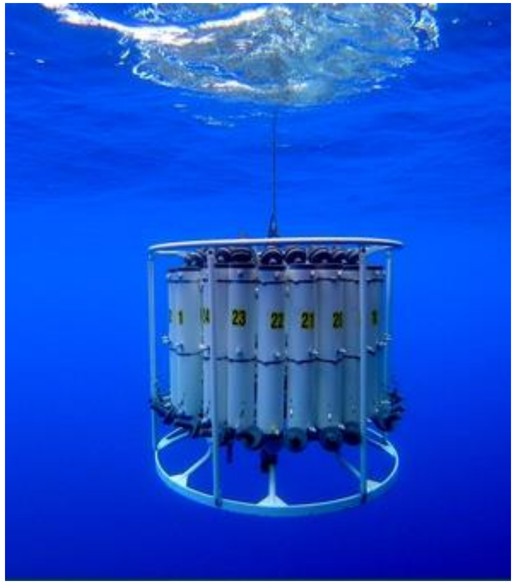
CTD profiles
Niskin profiles
Station descriptions

CRUISE

Cruise tracks
Expedition instrumentation and protocols

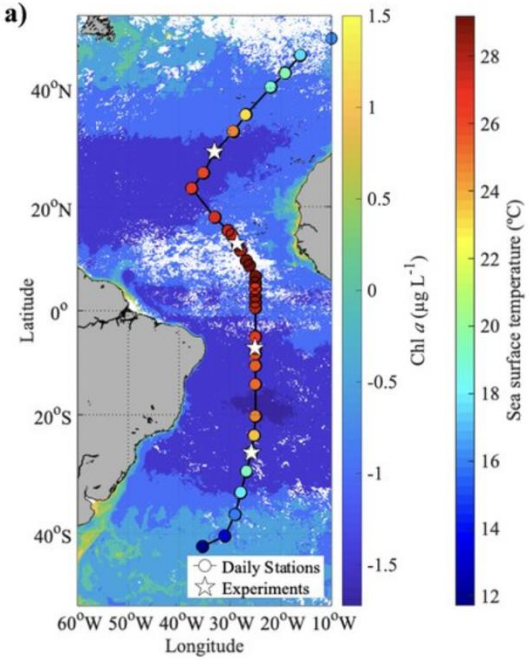
PROJECT

Project scope and PI



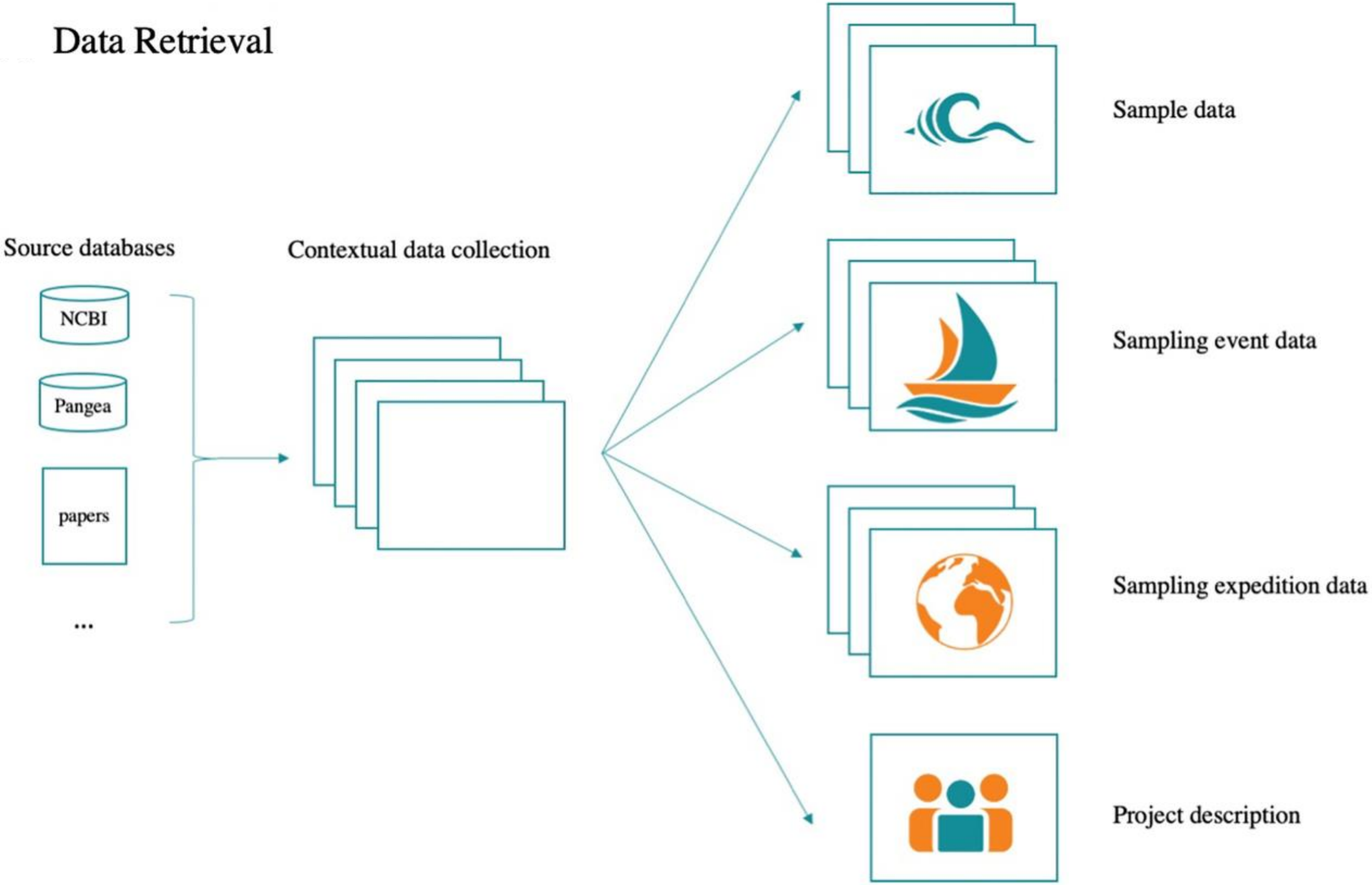
Contextual water column measure from CTD & Niskin

Large scale oceanographic information from cruise tracks



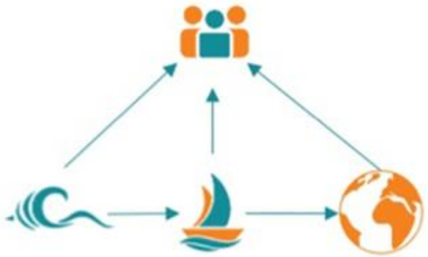
Ponsero et al. 2021

Frictionless data packages



Ponsero *et al*. 2021

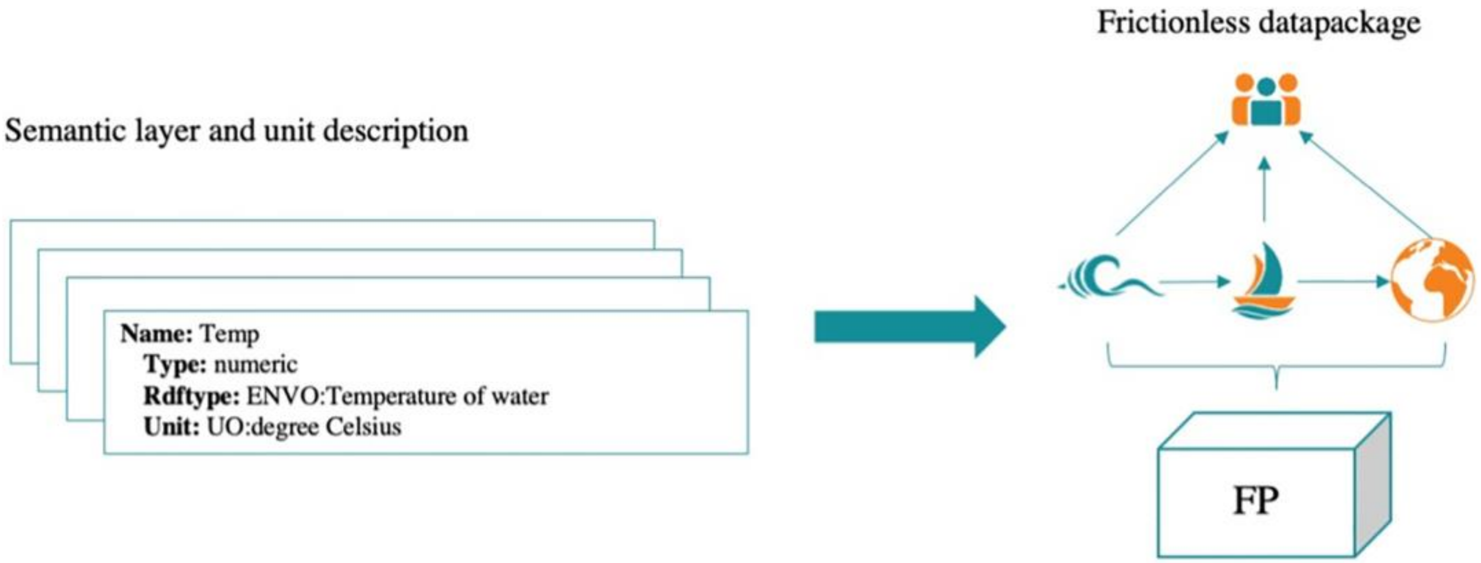
Frictionless data packages



Relational database reconnecting all contextual information to the sample

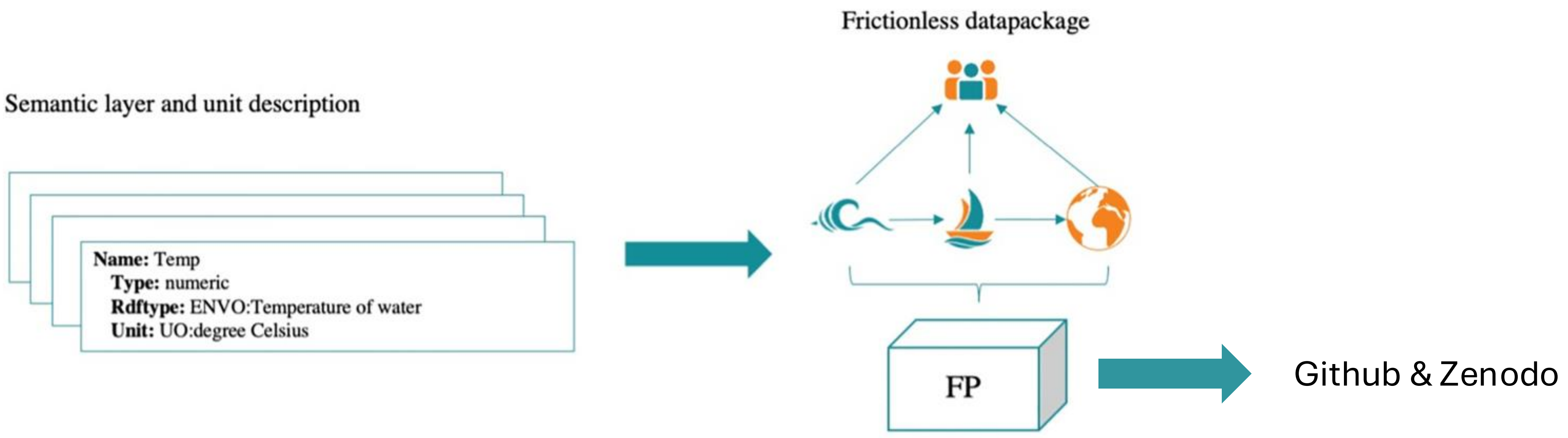
Ponsero *et al*. 2021

Frictionless data packages

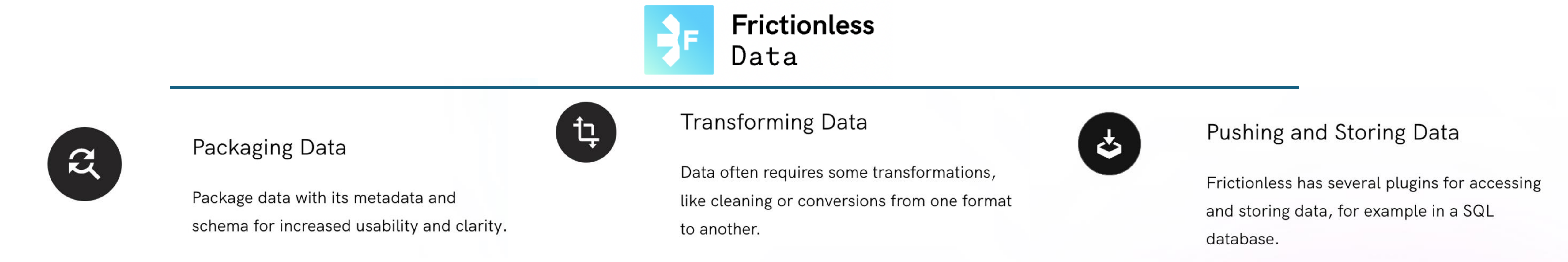


Name: Temp
Type: numeric
Rdftype: ENVO:Temperature of water
Unit: UO:degree Celsius

Frictionless data packages



Github & Zenodo



Packaging Data

Package data with its metadata and schema for increased usability and clarity.

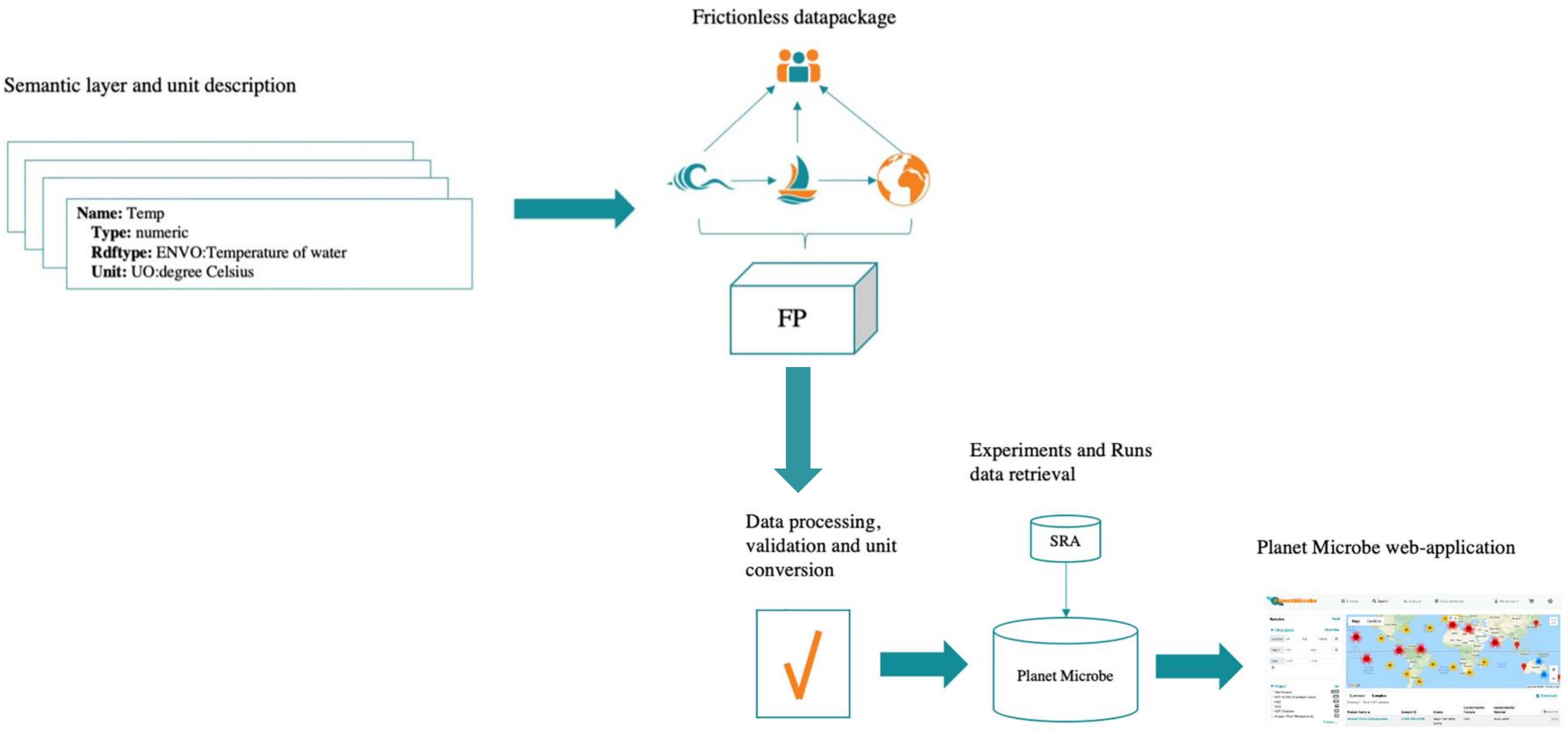
Transforming Data

Data often requires some transformations, like cleaning or conversions from one format to another.

Pushing and Storing Data

Frictionless has several plugins for accessing and storing data, for example in a SQL database.

Frictionless data packages



Semantic layer and unit description

Name: Temp
Type: numeric
Rdftype: ENVO:Temperature of water
Unit: UO:degree Celsius

Frictionless datapackage

FP

Data processing, validation and unit conversion

Experiments and Runs data retrieval

SRA

Planet Microbe

Planet Microbe web-application

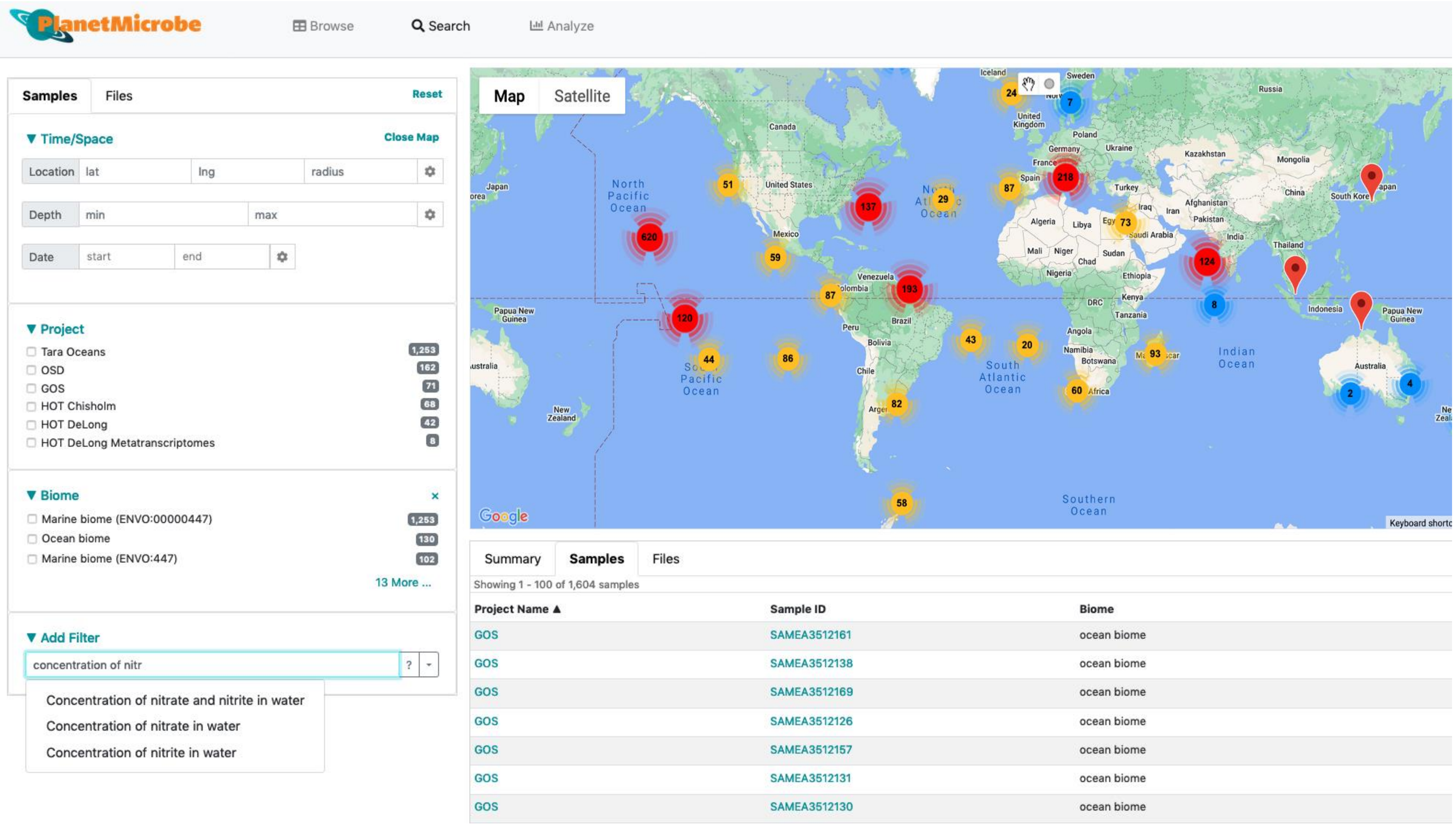
Ponsero *et al*. 2021

Increase reusability of datasets for meta-analysis

4D search

semantic search

map search



Samples | Files | Reset

▼ Time/Space — Close Map

Location: lat | lng | radius

Depth: min | max

Date: start | end

▼ Project

□ Tara Oceans	1,253
□ OSD	162
□ GOS	71
□ HOT Chisholm	68
□ HOT DeLong	42
□ HOT DeLong Metatranscriptomes	8

▼ Biome ×

□ Marine biome (ENVO:00000447)	1,253
□ Ocean biome	130
□ Marine biome (ENVO:447)	102

13 More ...

▼ Add Filter

concentration of nitr

Concentration of nitrate and nitrite in water
Concentration of nitrate in water
Concentration of nitrite in water

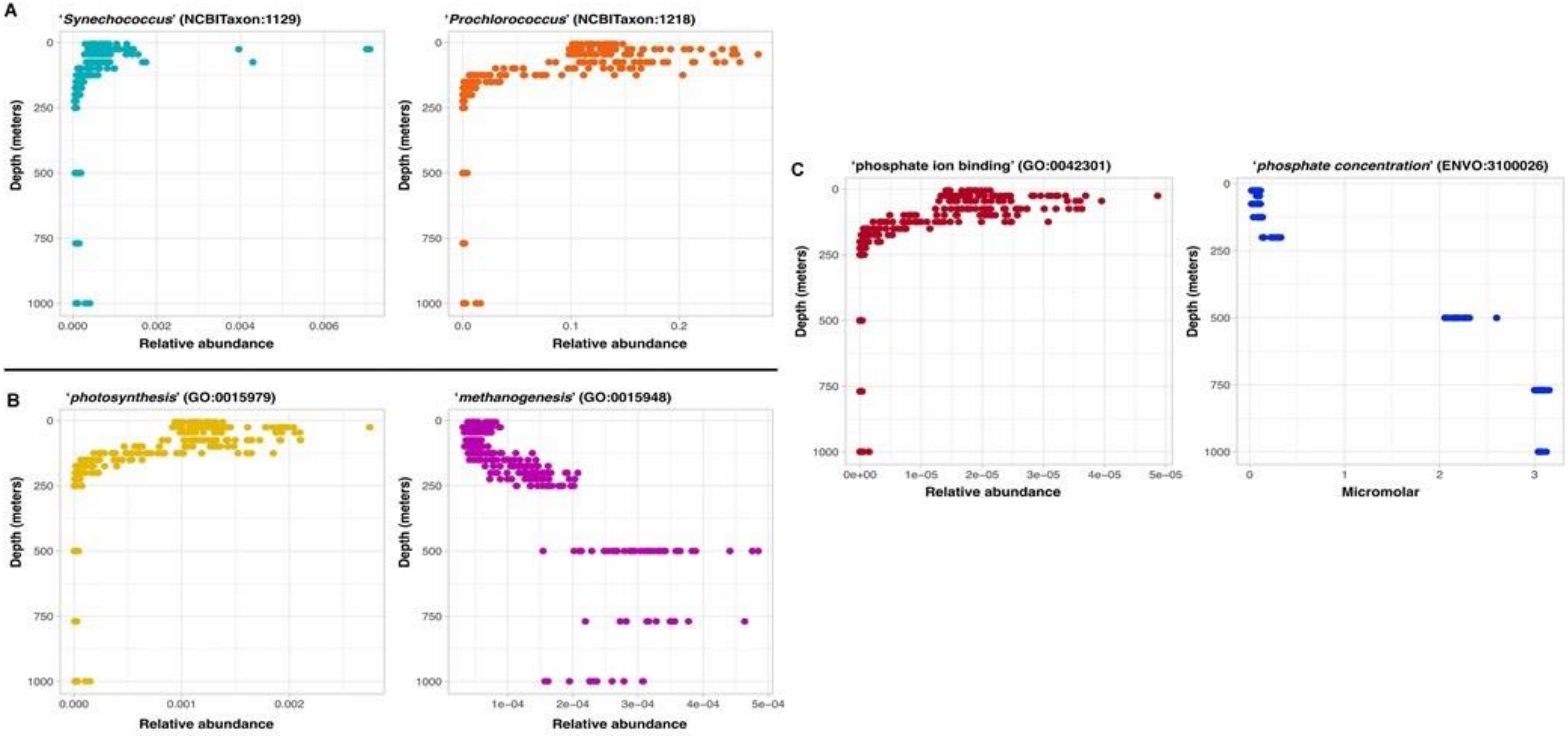
Summary | Samples | Files

Showing 1 - 100 of 1,604 samples

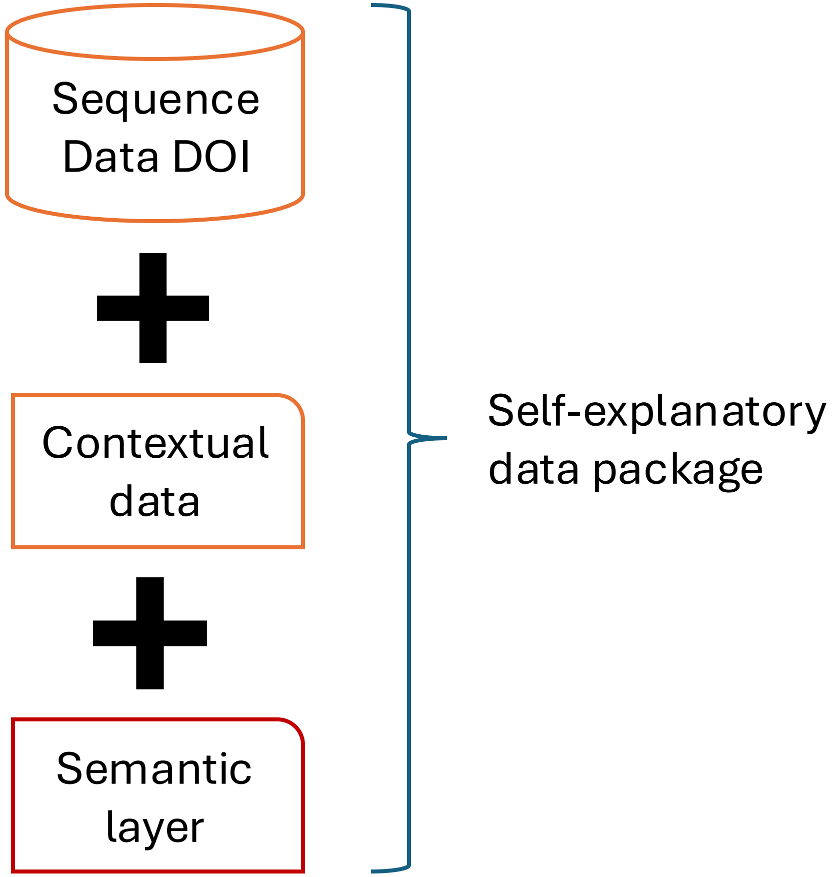
Project Name ▲	Sample ID	Biome
GOS	SAMEA3512161	ocean biome
GOS	SAMEA3512138	ocean biome
GOS	SAMEA3512169	ocean biome
GOS	SAMEA3512126	ocean biome
GOS	SAMEA3512157	ocean biome
GOS	SAMEA3512131	ocean biome
GOS	SAMEA3512130	ocean biome

Ponsero *et al.* 2021

Increased reusability of datasets for meta-analysis



Blumberg *et al.* 2023



Sequence Data DOI

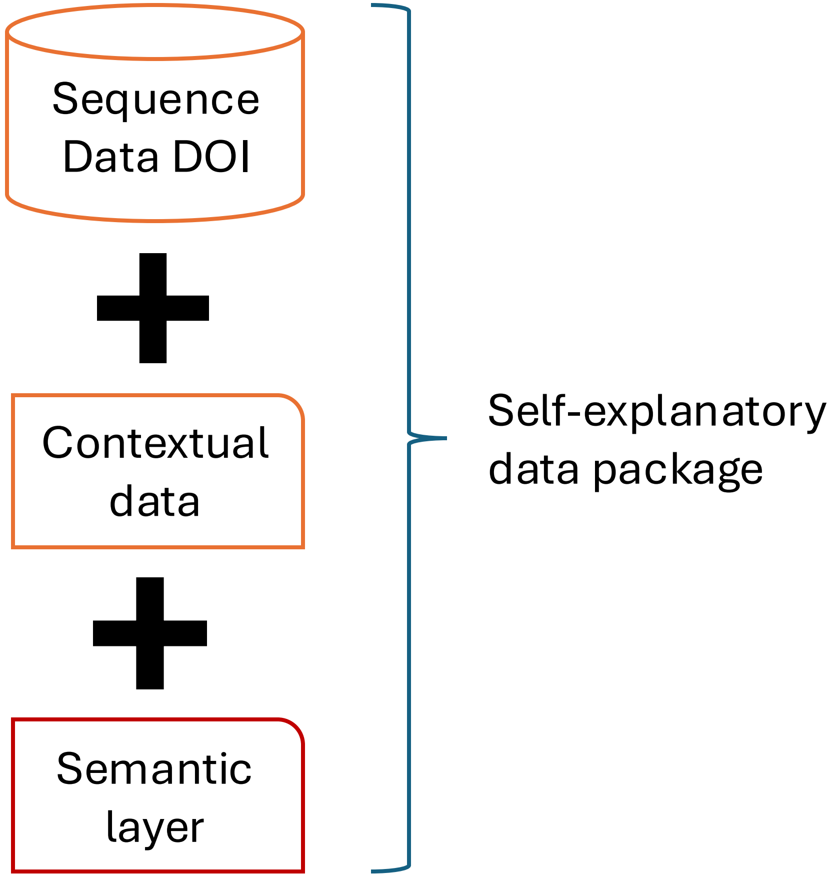
Contextual data

Semantic layer

Self-explanatory data package

In a nutshell:

- FAIR metagenomics is about capturing the complexity of the data
- Adding a semantic layer to your contextual data will help with interoperability and reusability
- Taking the effort to create curated data packages allows to make valuable ‘omics datasets interoperable and reusable



Sequence Data DOI

Contextual data

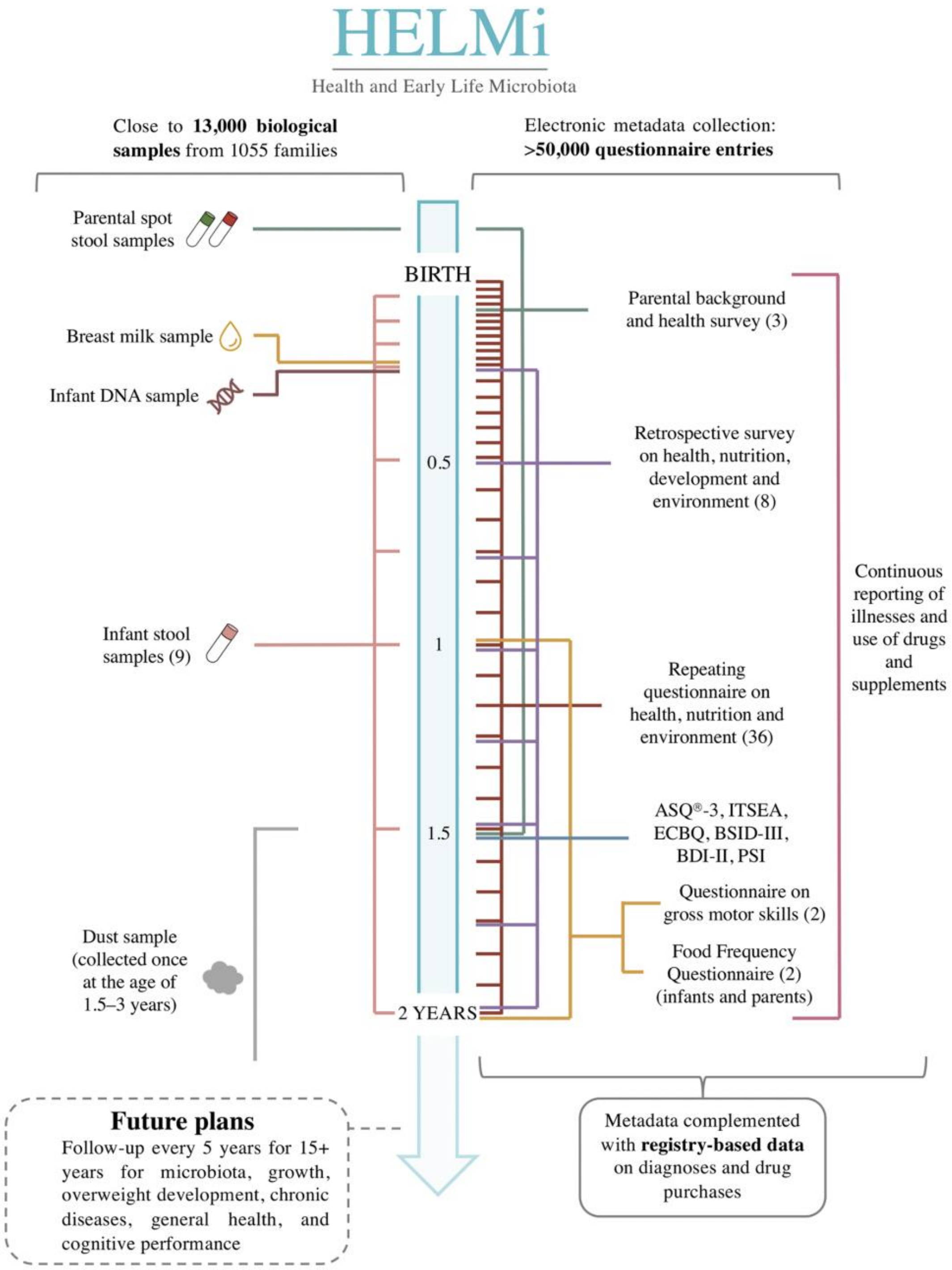
Semantic layer

Self-explanatory data package

In a nutshell:

- FAIR metagenomics is about capturing the complexity of the data
- Adding a semantic layer to your contextual data will help with interoperability and reusability
- Taking the effort to create curated data packages allows to make valuable ‘omics datasets interoperable and reusable

What if your data is sensitive or is not yet publicly available, why making your data FAIR matters?

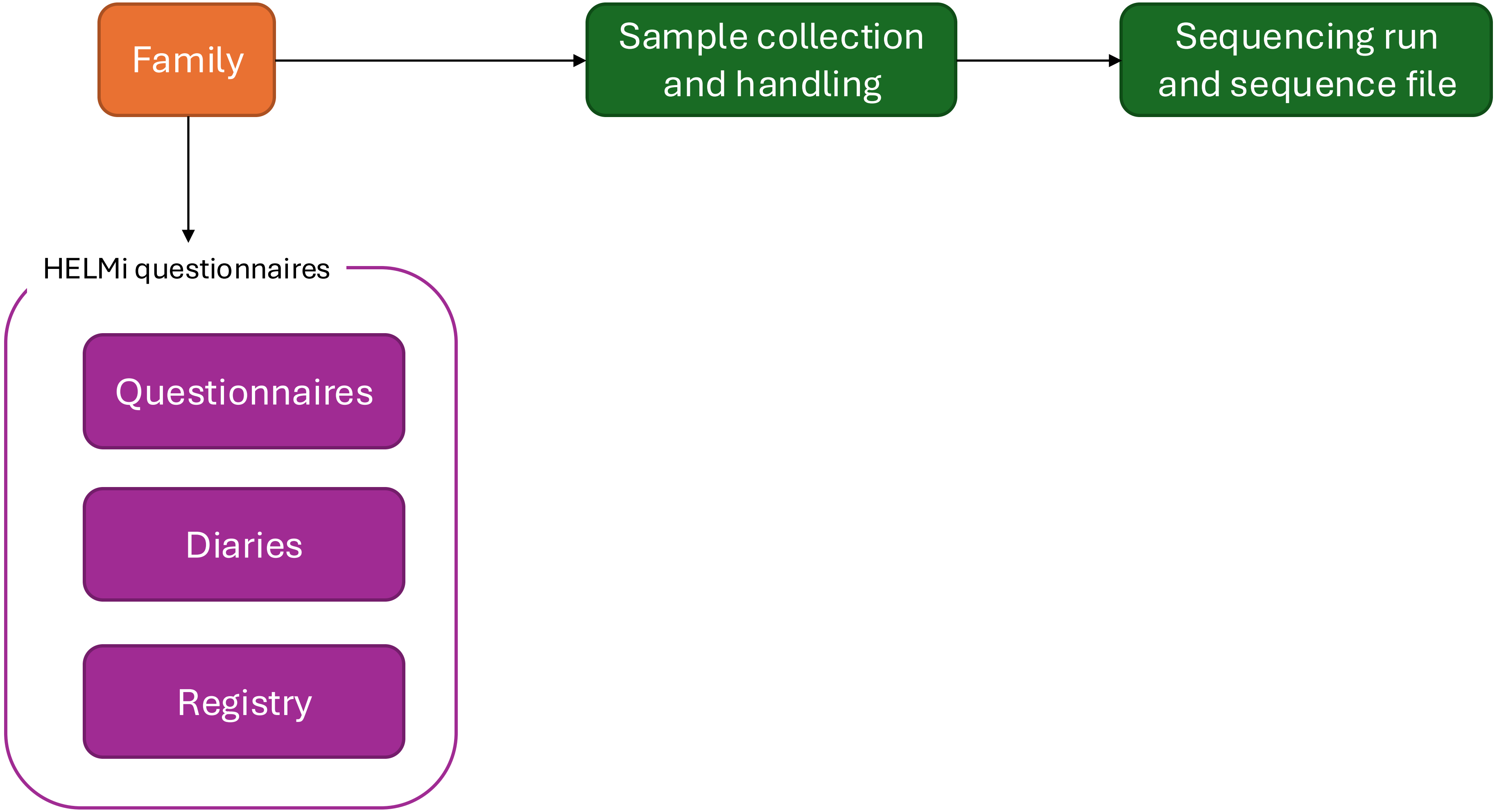

HELMi
Health and Early Life Microbiota
Close to **13,000 biological samples** from 1055 families
Electronic metadata collection: **>50,000 questionnaire entries**
Parental spot stool samples
BIRTH
Breast milk sample
Infant DNA sample
0.5
1
1.5
2 YEARS
Infant stool samples (9)
Dust sample (collected once at the age of 1.5–3 years)
Parental background and health survey (3)
Retrospective survey on health, nutrition, development and environment (8)
Repeating questionnaire on health, nutrition and environment (36)
ASQ®-3, ITSEA, ECBQ, BSID-III, BDI-II, PSI
Questionnaire on gross motor skills (2)
Food Frequency Questionnaire (2) (infants and parents)
Continuous reporting of illnesses and use of drugs and supplements
Future plans
Follow-up every 5 years for 15+ years for microbiota, growth, overweight development, chronic diseases, general health, and cognitive performance
Metadata complemented with **registry-based data** on diagnoses and drug purchases

Issues and needs for data stewardship

- Curation and harmonization (and translations) of the questionnaires data
- Tracking metagenomes, metabolomes and isolates obtained from the samples
- Coordinating multi-location collaborative projects on the dataset
- Tracking issues, contaminations and labelling errors
- Ensuring long-term reusability of the dataset even after students/postdocs have left

Korpela et al. 2021

Database to improve data management of complex datasets



Family

Sample collection and handling

Sequencing run and sequence file

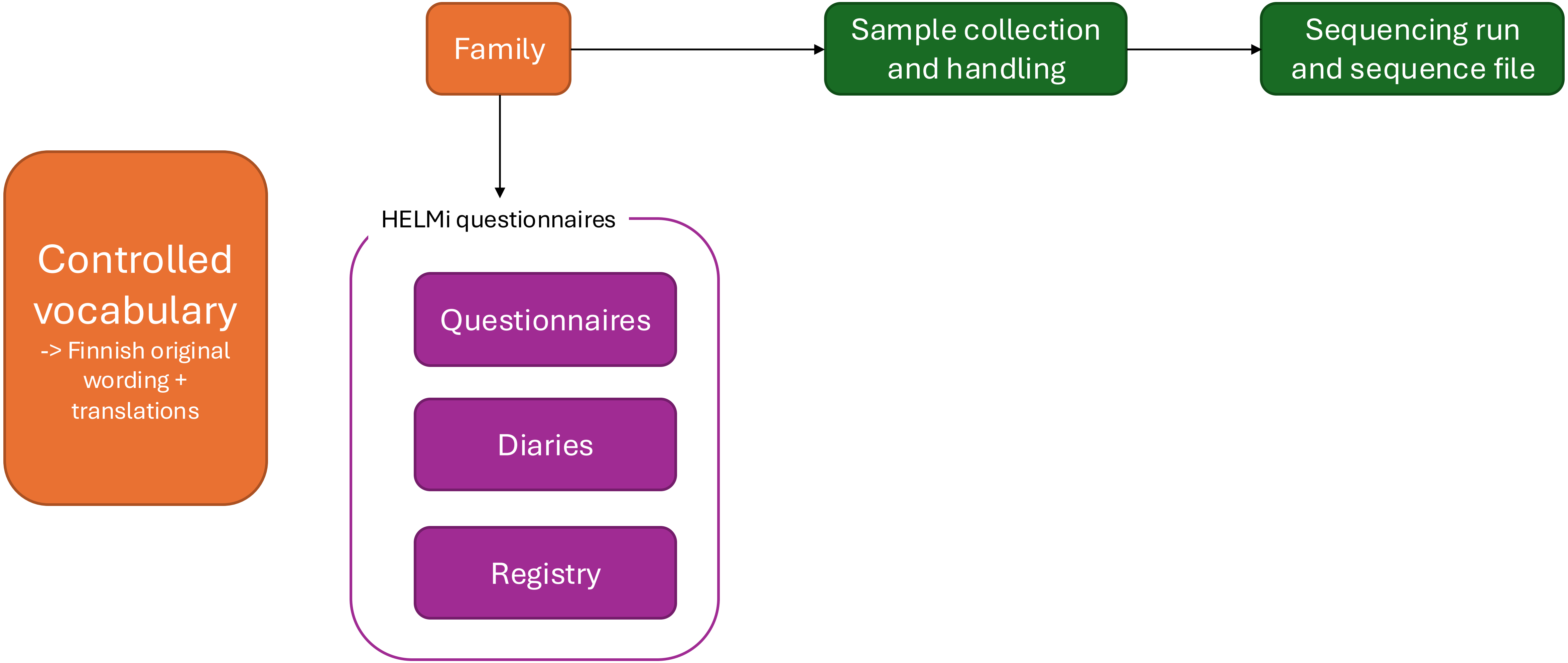
HELMi questionnaires

Questionnaires

Diaries

Registry

Database to improve data management of complex datasets



Controlled vocabulary

-> Finnish original wording + translations

Family

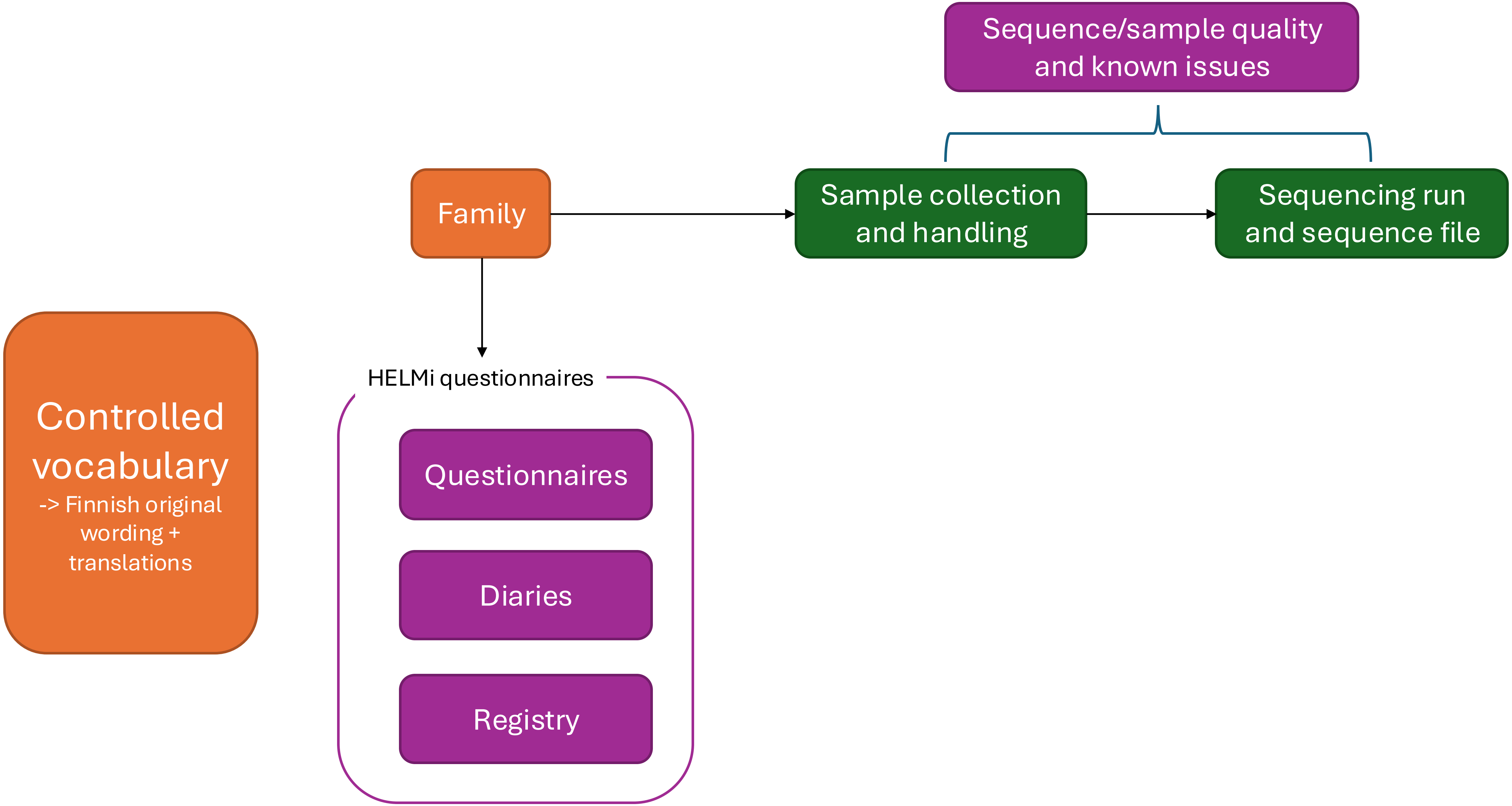
Sample collection and handling

Sequencing run and sequence file

HELMi questionnaires

- Questionnaires
- Diaries
- Registry

Database to improve data management of complex datasets



Sequence/sample quality and known issues

Family

Sample collection and handling

Sequencing run and sequence file

Controlled vocabulary

-> Finnish original wording + translations

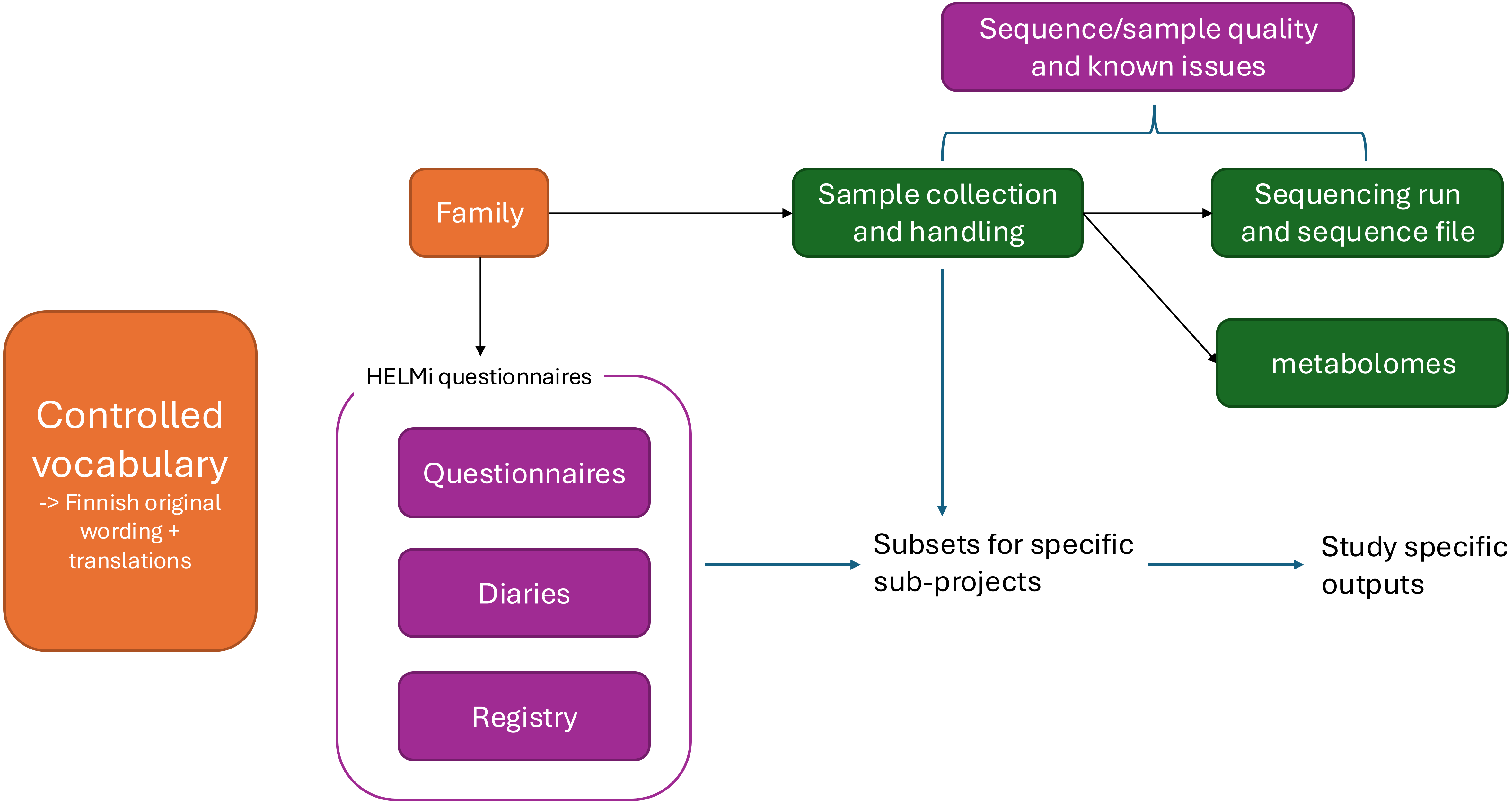
HELMi questionnaires

Questionnaires

Diaries

Registry

Database to improve data management of complex datasets



Sequence/sample quality and known issues

Family

Sample collection and handling

Sequencing run and sequence file

metabolomes

Controlled vocabulary

-> Finnish original wording + translations

HELMi questionnaires

Questionnaires

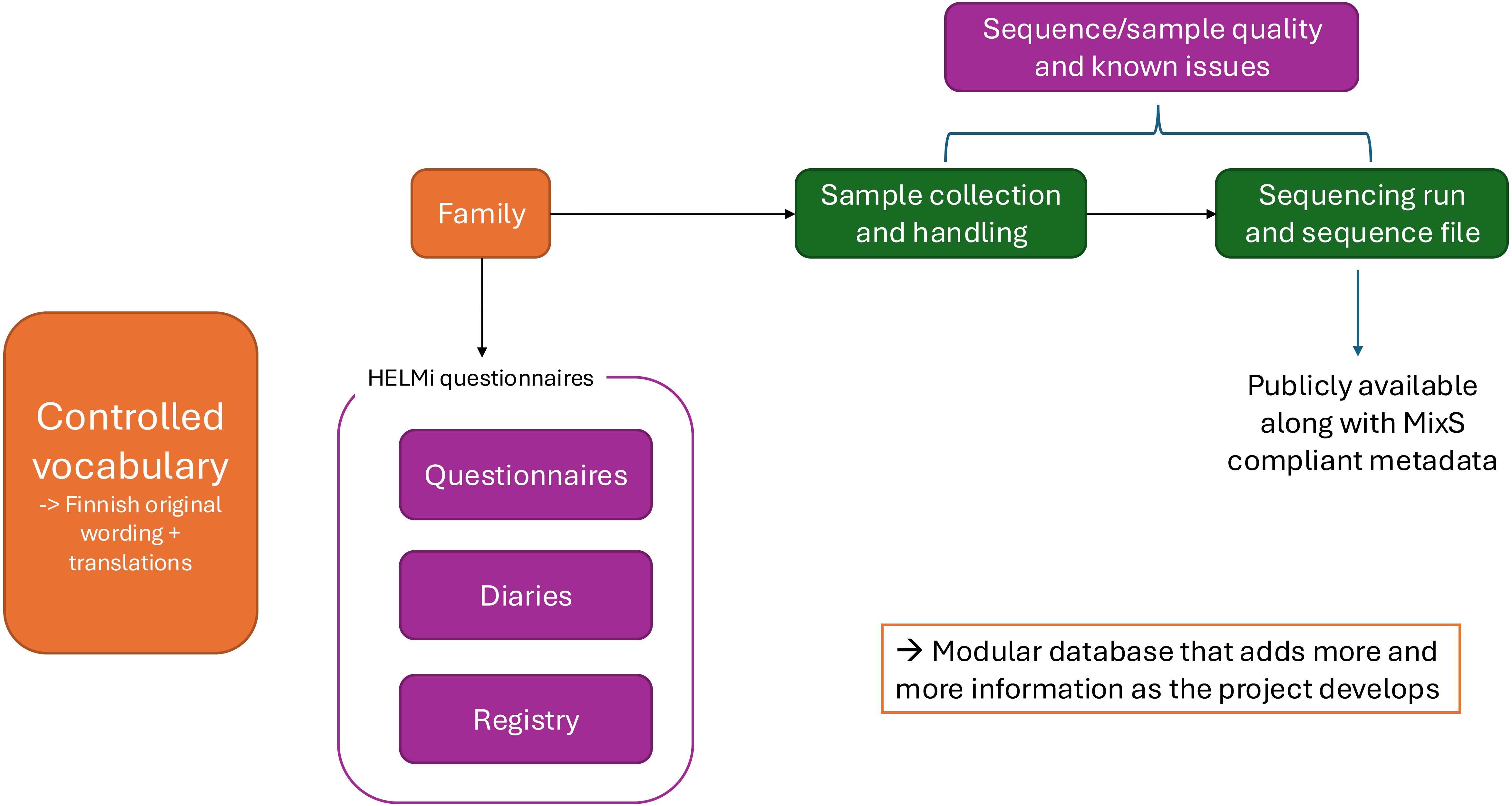
Diaries

Registry

Subsets for specific sub-projects

Study specific outputs

Database to improve data management of complex datasets



Sequence/sample quality and known issues

Family

Sample collection and handling

Sequencing run and sequence file

HELMi questionnaires

Questionnaires

Diaries

Registry

Publicly available along with MixS compliant metadata

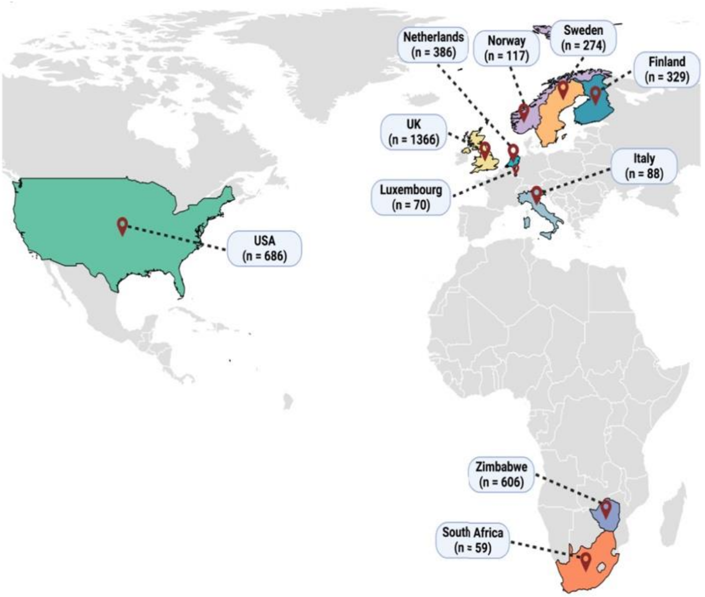
Controlled vocabulary

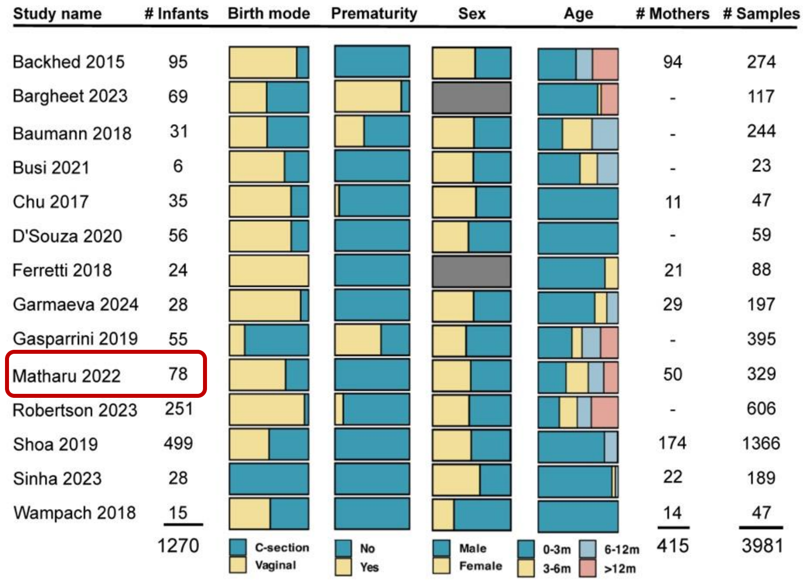
-> Finnish original wording + translations

→ Modular database that adds more and more information as the project develops

Harmonization with other datasets

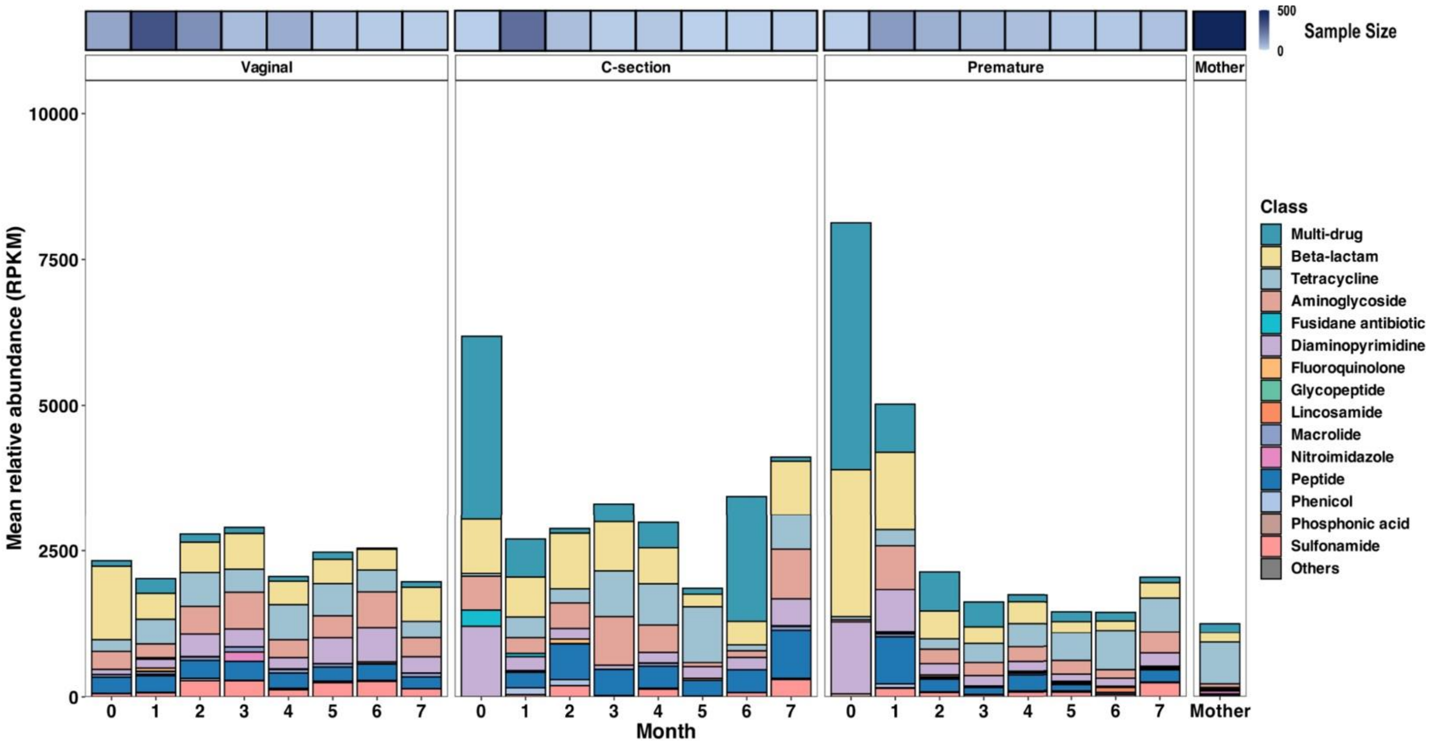
- Systematic search to identify relevant infant cohorts
- Controlled vocabulary for variables of interests
- Curation of relevant metadata across 14 cohorts



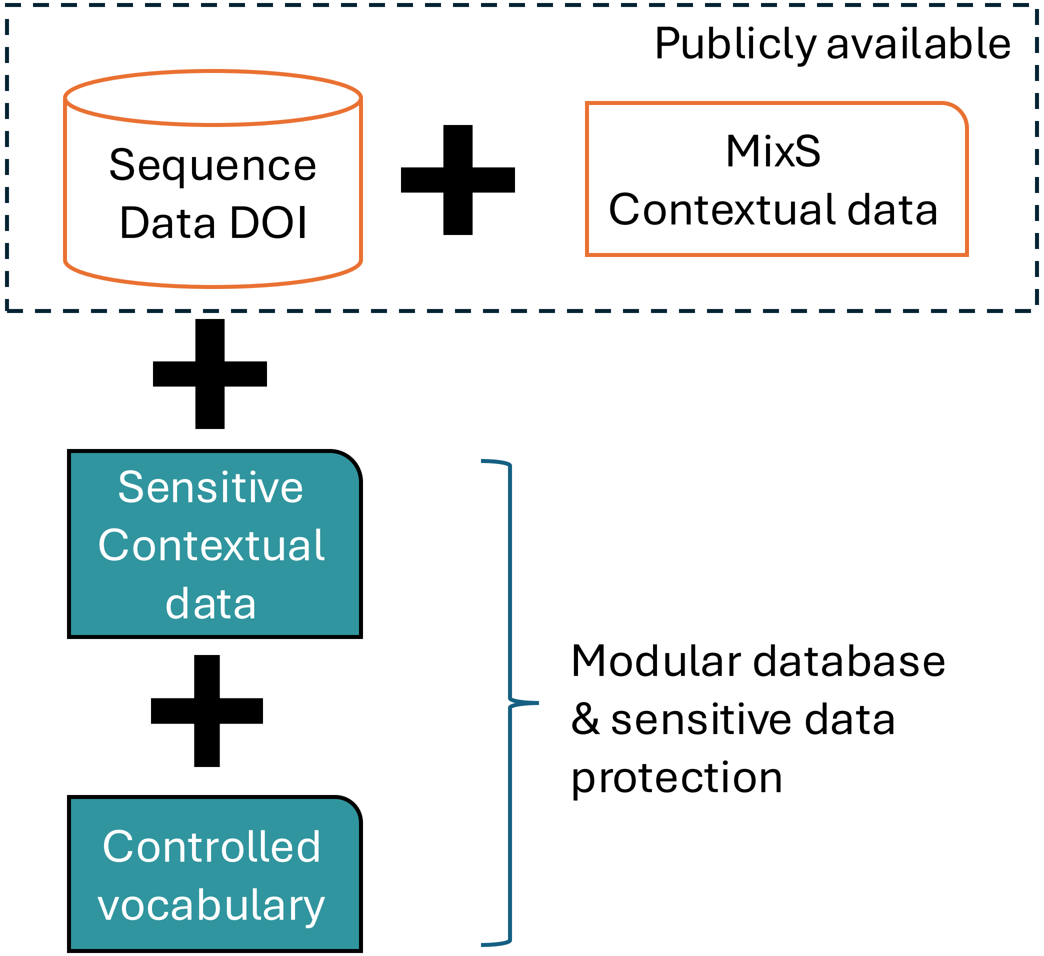


Birth mode
Prematurity
Sex
Age
C-section
Vaginal
No
Yes
Male
Female
0-3m
3-6m
6-12m
>12m

Study name	# Infants	# Mothers	# Samples
Backhed 2015	95	94	274
Bargheet 2023	69	-	117
Baumann 2018	31	-	244
Busi 2021	6	-	23
Chu 2017	35	11	47
D'Souza 2020	56	-	59
Ferretti 2018	24	21	88
Garmaeva 2024	28	29	197
Gasparrini 2019	55	-	395
Matharu 2022	78	50	329
Robertson 2023	251	-	606
Shoa 2019	499	174	1366
Sinha 2023	28	22	189
Wampach 2018	15	14	47
	1270	415	3981



Bargheet et al. 2025



In a nutshell:

- FAIR standards needs to account for sensitive data protection
- Implementing extensive data stewardship allows for long terms project management resilience
- Help supporting collaborative works and meta-analysis

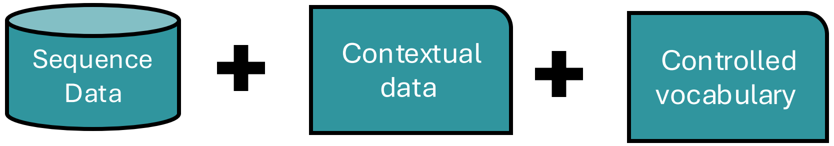
Before publication

Sequence Data + Contextual data + Controlled vocabulary

- Internal data management that document the dataset
- Internal controlled vocabulary

- Modular and comprehensive
- Limit data loss risks during the project

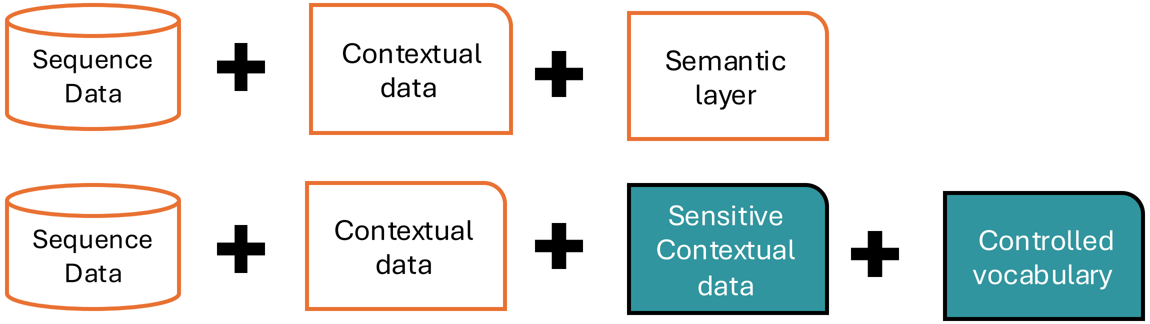
Before publication



- Internal data management that document the dataset
- Internal controlled vocabulary

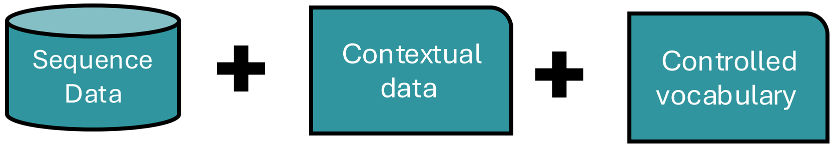
- Modular and comprehensive
- Limit data loss risks during the project

At publication



- Fixed data package that contained all relevant contextual data
- Sensitive data may be accessible upon request
- Long term storage and safety of the dataset needs to be addressed

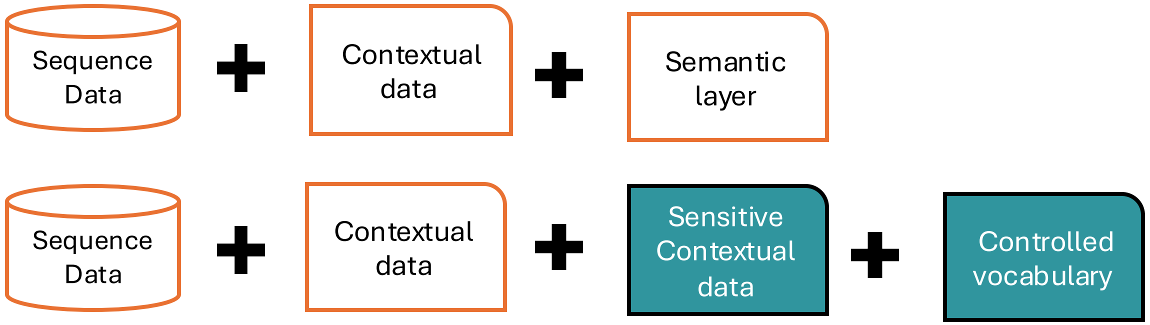
Before publication



- Internal data management that document the dataset
- Internal controlled vocabulary

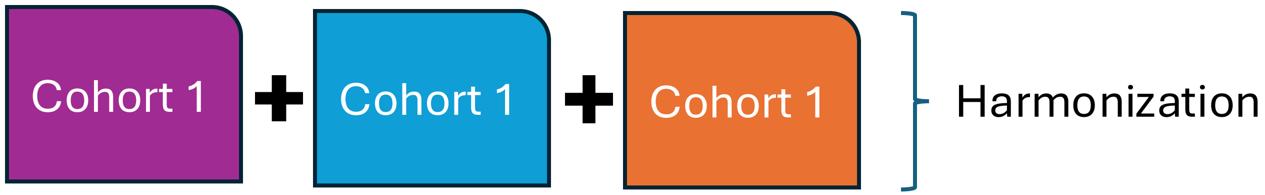
- Modular and comprehensive
- Limit data loss risks during the project

At publication



- Fixed data package that contained all relevant contextual data
- Sensitive data may be accessible upon request
- Long term storage and safety of the dataset needs to be addressed

After publication



- Curation and harmonization efforts to make high-value datasets re-usable for meta-analysis or cross-study analysis

- **What are the current pain points in data management in your lab?**
    - Drawing the data management plan → a DS meeting in October will focus on this
    - Data management during the project?
    - Making data accessible? Including secondary outputs (protein catalogues,… etc)
    - Reusing old data with current datasets?

- **What use cases would you be interested in applying FAIR principles ?**
    - What high-value datasets would benefit in being curated in a data package?
    - Highly collaborative projects that would benefit in having a data stewardship in place?