Metadata and Data Standards – What and Why

Matthew Schu, PhD

27 June 2023

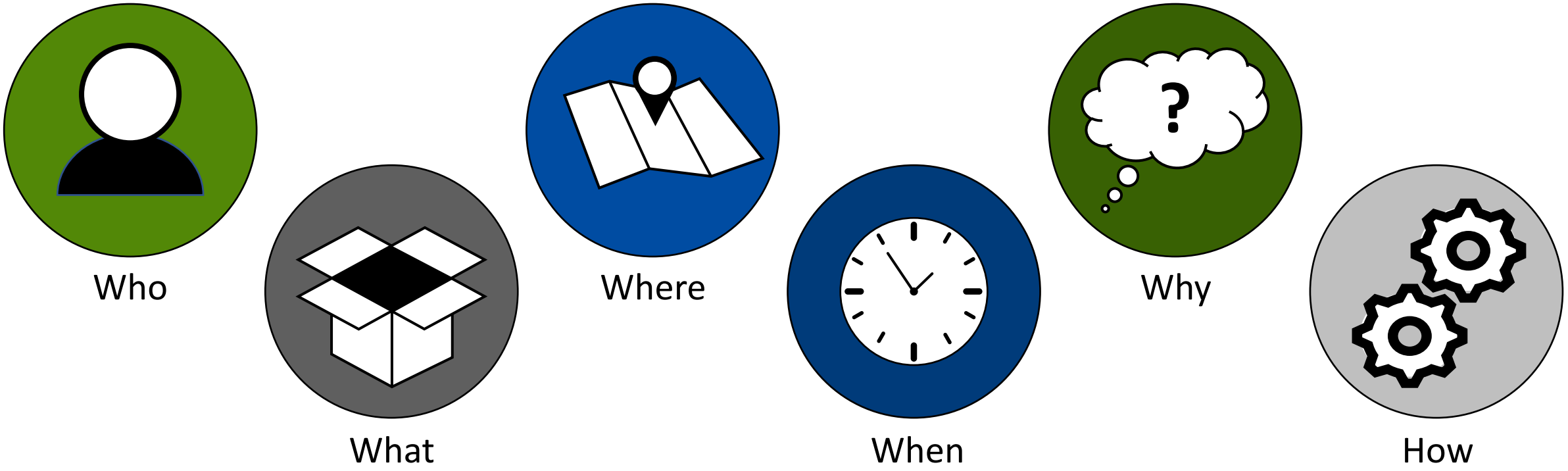
RTI INTERNATIONAL

Maximizing the value of scientific data

- Test validity of research findings
- Strengthen analysis through combined data sets
- Allow for reuse of hard to generate data
- Open new frontiers of discovery
- Foster trust in publicly funded research

Metadata - adding value to scientific data

Metadata is “data about data” and is the information needed to discover, use, and understand data and describes the



Reconstruct the context so users know what can and cannot be done with the data

What can you do with data?

What can you do with this data?

- Enjoy it (it's cute)
- Share it with a friend

When all you have is the data, there are a limitations on what you can do with it.

What can you do with data + metadata?

I'm looking for a pet and this puppy is cute. What breed of dog is this?
- Pet Lover

What were the camera settings to focus on the puppy and blur the background?
- Photographer

I want to make this into a poster. Are there any license considerations? What is the format of the image.
- Home Owner

Metadata:
Image Title: Puppy with leaves
Subject: Golden retriever puppy
Subject Age: 8 weeks
Subject Sex: Female
Photographer: Cam Jansen
Camera Manufacturer: Canon
Camera Model: EOS 5D Mark IV
Camera Lens: EF 24-105
Zoom Level: 4x
Aperture: f1.4
Focal Length: 60mm
Image Date: 9-16-2021
Image Format: PNG
Image Resolution: 300 dpi
Image Processing Software: Lightroom
Licensing: Shared under CC agreement

Metadata Levels

Study or Dataset	Variable	File
Information describing the purpose of collecting the data • Study name • Study URL or Digital Object Identifier (DOI) • Funding information • Investigator contact information • Associated publications • Release version • Study duration and dates active • Collection protocol • Sample size and population description	**Information about measured variables that belong to a study or dataset** • Variable name • Variable description • Variable data type (e.g., numeric, character, Boolean) • Data format (e.g., field length) • Dataset that contains the variable • Generation method Often stored in tabular format (data dictionary or code book)	**Information about files that have been produced during the course of the study** • File name • File URL • File description • File format (e.g., .csv, .xlsx, .png) • Access control information • License

You're probably already collecting metadata, often required for manuscript or repository submission

Metadata Standards

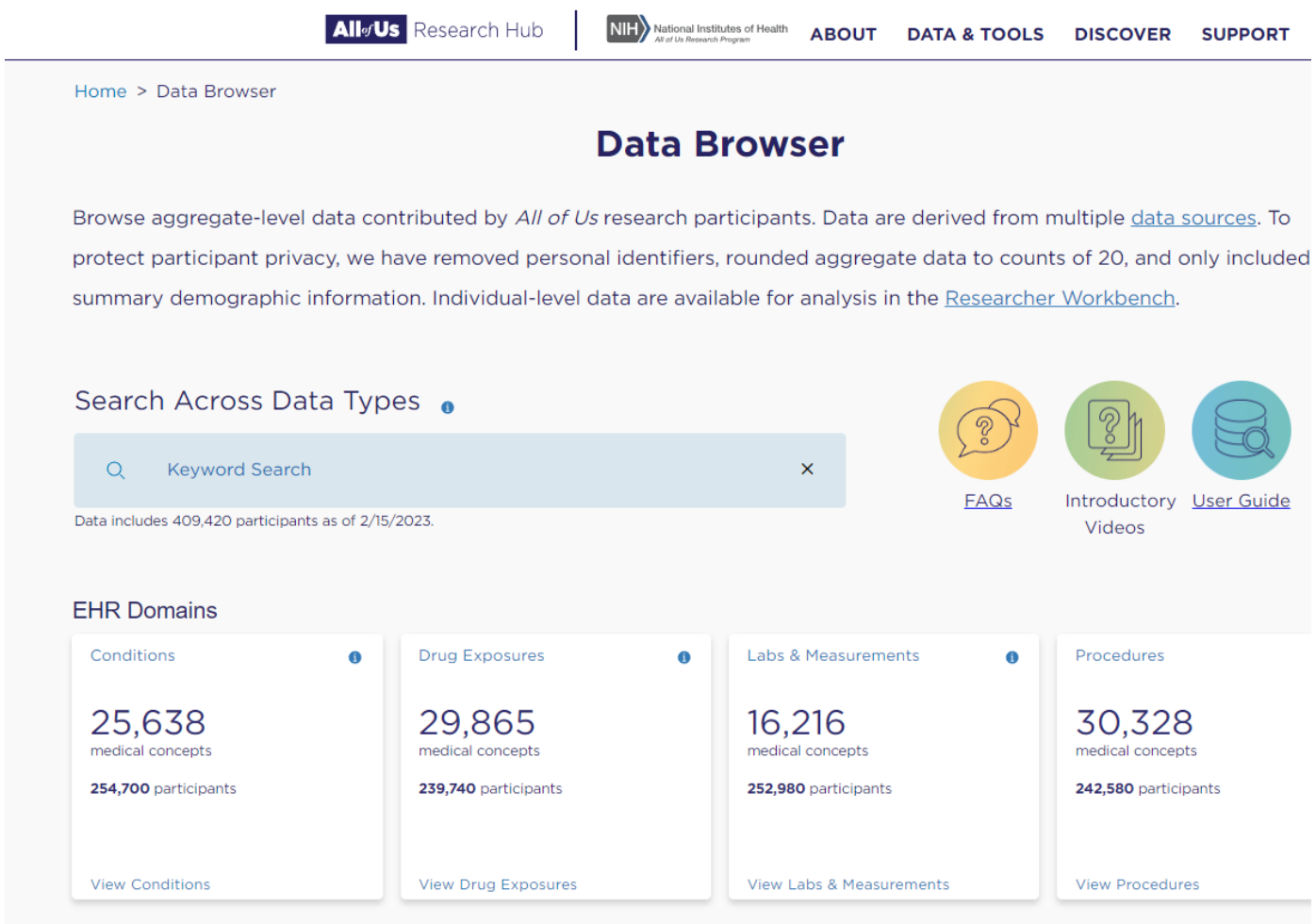
Data associated with rich metadata is more useful, but what level of metadata would allow for appropriate data reuse?

Data Type	Metadata Standards
Biochemical	Minimum Information About a Bioactive Entity (MIABE)
Clinical	Clinical Data Interchange Standards Consortium (CDISC) Analysis Data Model
Genomics	Minimum Information about a (Meta)Genome Sequence (MIGS/MIMS)
Transcriptomics	Minimal Information about a high throughput SEQuencing Experiment (MINSEQE)
Imaging	Minimum Information about Tissue Imaging (MITI)
Metabolomics	Core Information for Metabolomics Reporting (CIMR)
Proteomics	The Minimum Information About a Proteomics Experiment (MIAPE)

Metadata standards for various biological disciplines are available at fairsharing.org

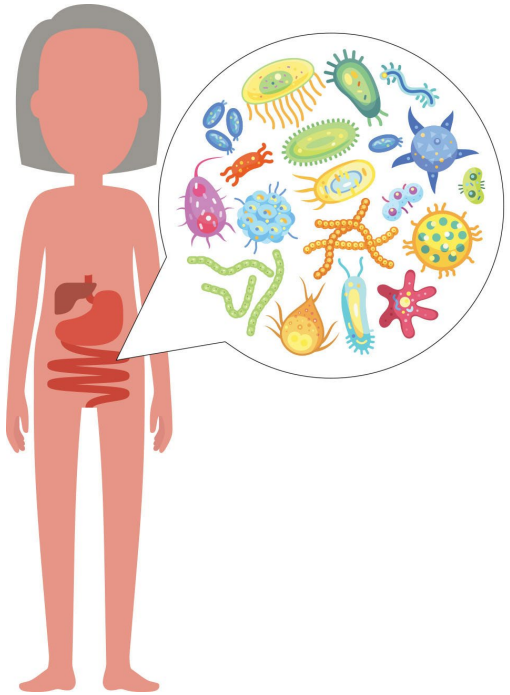
Nutrition for Precision Health Powered by *All of Us* (NPH)

All of Us Researcher Workbench



- Data from NPH will be made available via the *All of Us* Researcher workbench
  - AI analytic tools will be brought to the data to help uncover patterns within the high dimensional data captured
  - Other researchers will be able to access the data from NPH and potentially combine these results with those from other studies in the Researcher Workbench

Scientific metadata example



Microbiome metadata

1.sample_name

2.host_subject_id

3.sample_type

4.sample_quality [Bristol stool scale]

5.physical_specimen_location

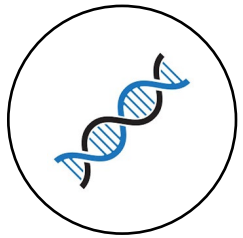
6.collection_date

7.country

- Type 1: Separate hard lumps, like nuts (difficult to pass)
- Type 2: Sausage-shaped, but lumpy
- Type 3: Like a sausage but with cracks on its surface
- Type 4: Like a sausage or snake, smooth and soft (average stool)
- Type 5: Soft blobs with clear cut edges
- Type 6: Fluffy pieces with ragged edges, a mushy stool (diarrhea)
- Type 7: Watery, no solid pieces, entirely liquid (diarrhea)

Scientific Data Reuse

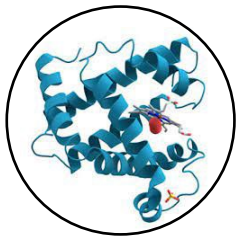
Study Data



Genomics (WGS, RNAseq)

Microscopy Images

Outcomes



Proteomics

Asking New Questions to Existing Dataset

Combining datasets

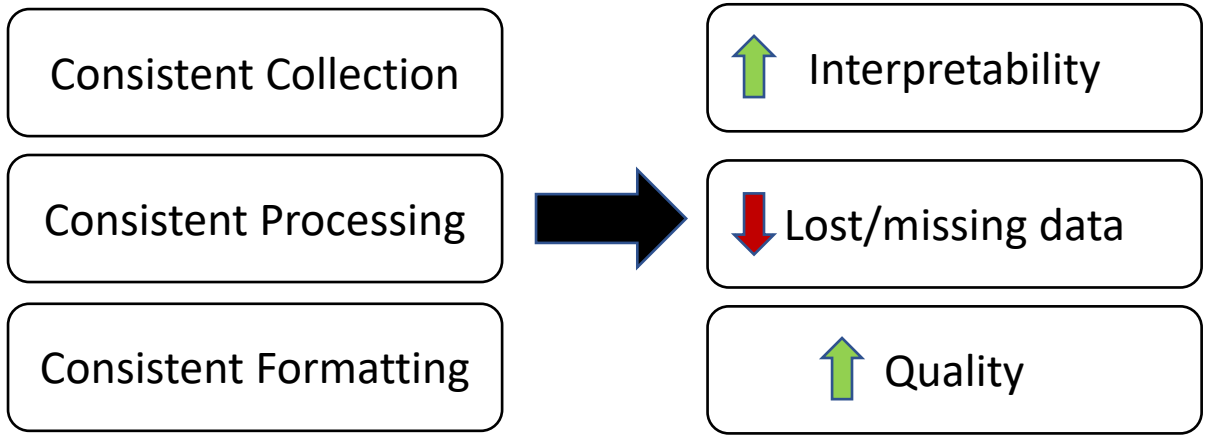
- Multiple available studies
- Investigators own data

Data Standards and Harmonization

Data Harmonization

	Study 1	Study 2
Question	How many packs a day do you <u>currently</u> smoke?	What is your smoking status?
Responses	1 – 0 2 – 1-2 packs 3 – 3-5 packs 4 – >5 packs	1 – Never smoker 2 – Past smoker 3 – Current smoker

How to combine the smoking status across the two studies?



Your own data will be readily compatible with all other studies that utilize the same or comparable standards

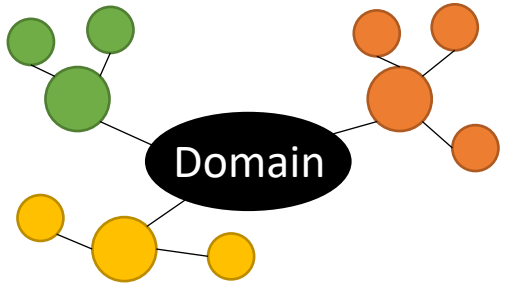
Data Standards

Vocabularies / Terminologies

Clearly communicating the definition of terms within the study context

- UMLS
- SNOMED
- LOINC
- RxNorm

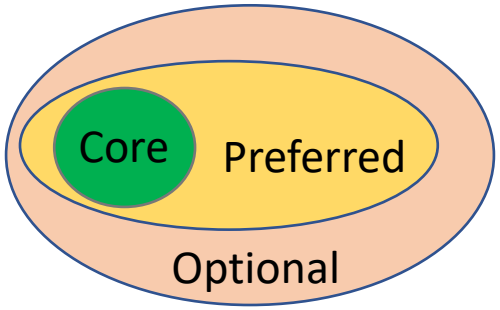
Ontologies



Describes the relationship between concepts within a scientific domain

- BioPortal
- Ontology of Precision Medicine and Investigation

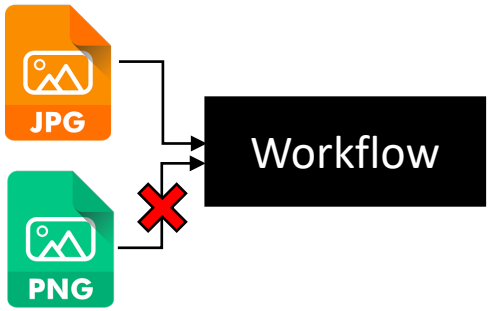
Common Data Elements



Instrument with questions and responses for consistent collecting of study data

- NIH CDE repository
- PhenX Toolkit

File Formats



Common file formats allow for consistent data processing and tool utilization

Example Data Standards for NPH (microbiome)

- Example from QIIME data standards

Sample information file

The *sample information file* will define the biological context of each sample, with categories like `sample_type`, `treatment`, etc. The `sample_name` defined in this file is used to relate each sample in the preparation file with the biological sample.

Required fields for Qiita

Note that Qiita require to have at least two columns, including `sample_name`, for a sample information file to be added to the system:

Field name	Format	Description
`sample_name`	free text with restrictions	Identifies a sample. It is the primary key and must be unique. Allowed characters are alphabetic `[A-Za-z]`, numeric `[0-9]`, and periods `.`.

Questions