What is FAIR? An outside view

28 August 2023

Mark A. Parsons
https://orcid.org/0000-0002-7723-0950
mark.parsons@uah.edu
University of Alabama in Huntsville
NASA Chief Science Data Office

This is our second webinar preparing for a workshop

FAIR for NASA Data

27-29 September 2023

Boulder CO

https://science.data.nasa.gov/news/events-fair-for-nasa-data/

Register today!

Perspectives

- FAIR assessment and FAIR qualification from GO FAIR — Erik Schultes, GO FAIR Foundation
- Improving the FAIRness of data at the US Geological Survey — Viv Hutchison, USGS
- Making biomedical data “born FAIR” — Mark A. Musen, Stanford Center for Biomedical Informatics Research

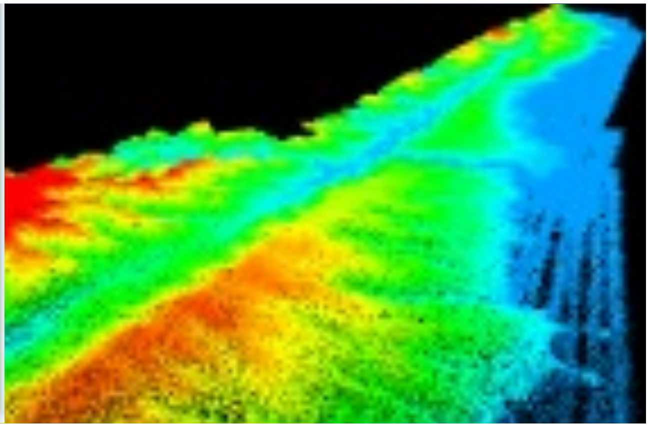
Start asking questions now

```
https://nasa.cnf.io/sessions/y7ef/
```

USGS State of the Data: Assessing the FAIRness of USGS Data Products

Viv Hutchison, Leslie Hsu, Tamar Norkin, Lisa Zolly
CSS Science Analytics and Synthesis

August 28, 2023



US Geological Survey

Science for a Changing World

The USGS serves the Nation by providing reliable scientific information to describe and understand the Earth;
minimizing loss of life and property from natural disasters;
managing water, biological, energy, and mineral resources; and enhancing and protecting our quality of life.

USGS State of the Data: Overarching Goals

- Establish a methodology using a quantitative analysis of the FAIR characteristics of USGS data and determine a baseline status for the current overall FAIRness of USGS data.

- Identify recommendations for how the USGS can improve its alignment with FAIR.

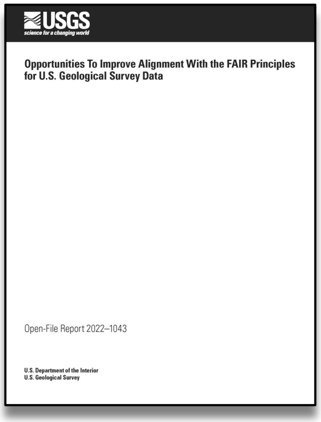
Background

USGS FAIR Roadmap Project

Project purpose: to recommend actions that USGS could take to improve alignment with the FAIR Principles.

2019 Workshop

2022 Report


USGS
science for a changing world

Opportunities To Improve Alignment With the FAIR Principles for U.S. Geological Survey Data

Open-File Report 2022–1043

U.S. Department of the Interior
U.S. Geological Survey

Project supported by the USGS Community for Data Integration (CDI)

USGS State of the Data: Methods and Status

Engaged community to develop and test a rubric based on FAIR Principles

Performed multiple analyses of rubric using a common dataset to calibrate scoring

Selected ~400 datasets randomly from Science Data Catalog for analysis

Analyzed individual datasets using rubric.

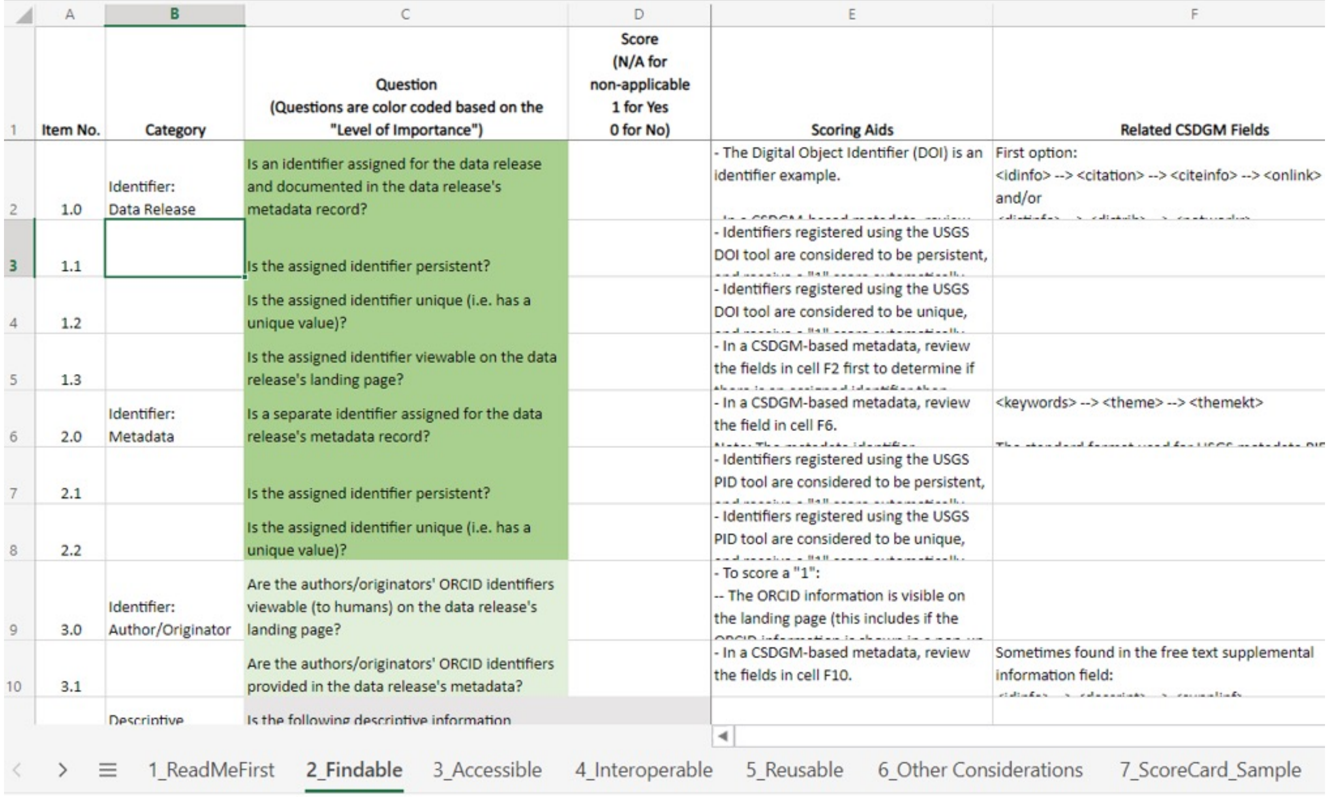
Compiled dataset to identify trends in analysis

Data Release in USGS ScienceBase (includes rubric)

Manuscript submitted to journal

USGS FAIR Rubric

Hutchison, V.B., Zolly, L.S., Norkin, T., Hsu, L., and Hou, C.-Y., 2023, USGS State of the Data Project: Rubric and Assessment Data: U.S. Geological Survey data release, https://doi.org/10.5066/P97V4XA4.



	A	B	C	D	E	F
1	Item No.	Category	Question (Questions are color coded based on the "Level of Importance")	Score (N/A for non-applicable 1 for Yes 0 for No)	Scoring Aids	Related CSDGM Fields
2	1.0	Identifier: Data Release	Is an identifier assigned for the data release and documented in the data release's metadata record?		- The Digital Object Identifier (DOI) is an identifier example. - In a CSDGM-based metadata, review	First option: <idinfo> --> <citation> --> <citeinfo> --> <onlink> and/or <distinfo> --> <distrib> --> <networkr>
3	1.1		Is the assigned identifier persistent?		- Identifiers registered using the USGS DOI tool are considered to be persistent, and receive a "1" score automatically.	
4	1.2		Is the assigned identifier unique (i.e. has a unique value)?		- Identifiers registered using the USGS DOI tool are considered to be unique, and receive a "1" score automatically.	
5	1.3		Is the assigned identifier viewable on the data release's landing page?		- In a CSDGM-based metadata, review the fields in cell F2 first to determine if there is an assigned identifier then	
6	2.0	Identifier: Metadata	Is a separate identifier assigned for the data release's metadata record?		- In a CSDGM-based metadata, review the field in cell F6. Note: The metadata identifier	<keywords> --> <theme> --> <themekt> The standard format used for USGS metadata PIF
7	2.1		Is the assigned identifier persistent?		- Identifiers registered using the USGS PID tool are considered to be persistent, and receive a "1" score automatically.	
8	2.2		Is the assigned identifier unique (i.e. has a unique value)?		- Identifiers registered using the USGS PID tool are considered to be unique, and receive a "1" score automatically.	
9	3.0	Identifier: Author/Originator	Are the authors/originators' ORCID identifiers viewable (to humans) on the data release's landing page?		- To score a "1": -- The ORCID information is visible on the landing page (this includes if the ORCID information is shown in a pop-up	
10	3.1		Are the authors/originators' ORCID identifiers provided in the data release's metadata?		- In a CSDGM-based metadata, review the fields in cell F10.	Sometimes found in the free text supplemental information field: <idinfo> --> <descript> --> <supplinf>
		Descriptive	Is the following descriptive information			

< > ≡ 1_ReadMeFirst 2_Findable 3_Accessible 4_Interoperable 5_Reusable 6_Other Considerations 7_ScoreCard_Sample

USGS FAIR Rubric

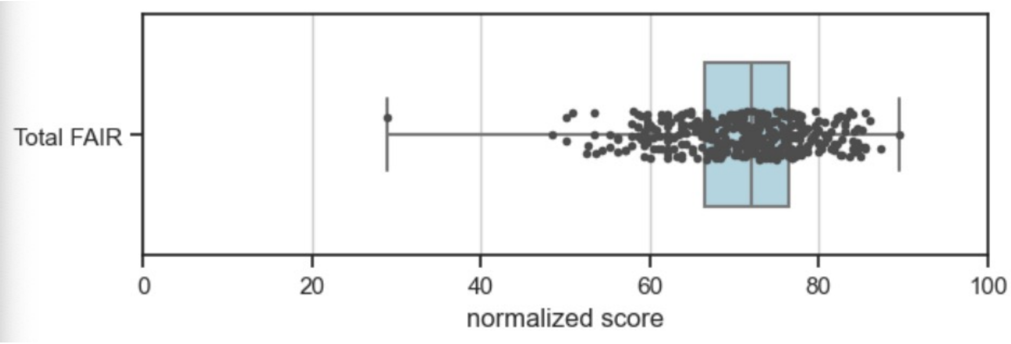
- **62 questions – y/n and n/a**
- **4 categories - F,A,I,R**
- **Essential, Intermediate, Advanced**
- **Questions based on FGDC CSDGM metadata fields**
- **Scoring guides for each question**
- **Scores are entered and totaled thru a formula**
- **Excel spreadsheet format**

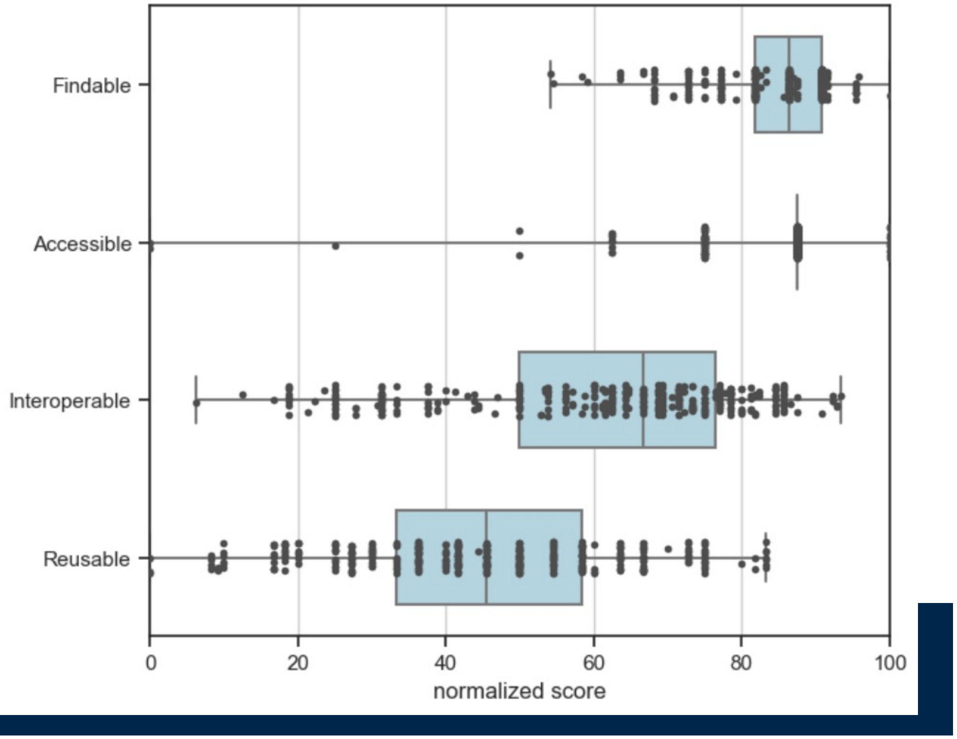
Key Findings

Total FAIR: The overall FAIR scores represent the number of relevant yeses and nos for each of the 62 rubric questions.

F, A, I, R: Scores for all 392 assessments, broken down in the four FAIR principles: Findable, Accessible, Interoperable, and Reusable.

Each score is normalized to a maximum of 100 and does not take into account questions that are Not Applicable.

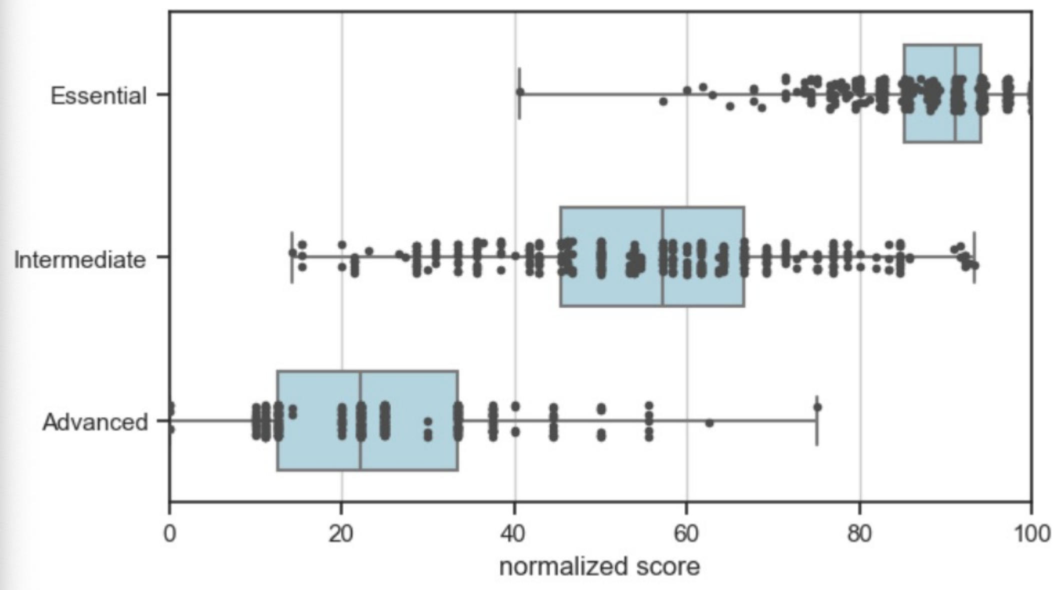




Key Findings

Each overall FAIR score can be broken down into the three designated levels of importance: Essential, Intermediate and Advanced.

Intermediate and Advanced category questions may be not be relevant to all datasets, but their lower scores indicate that there are areas for improvement.

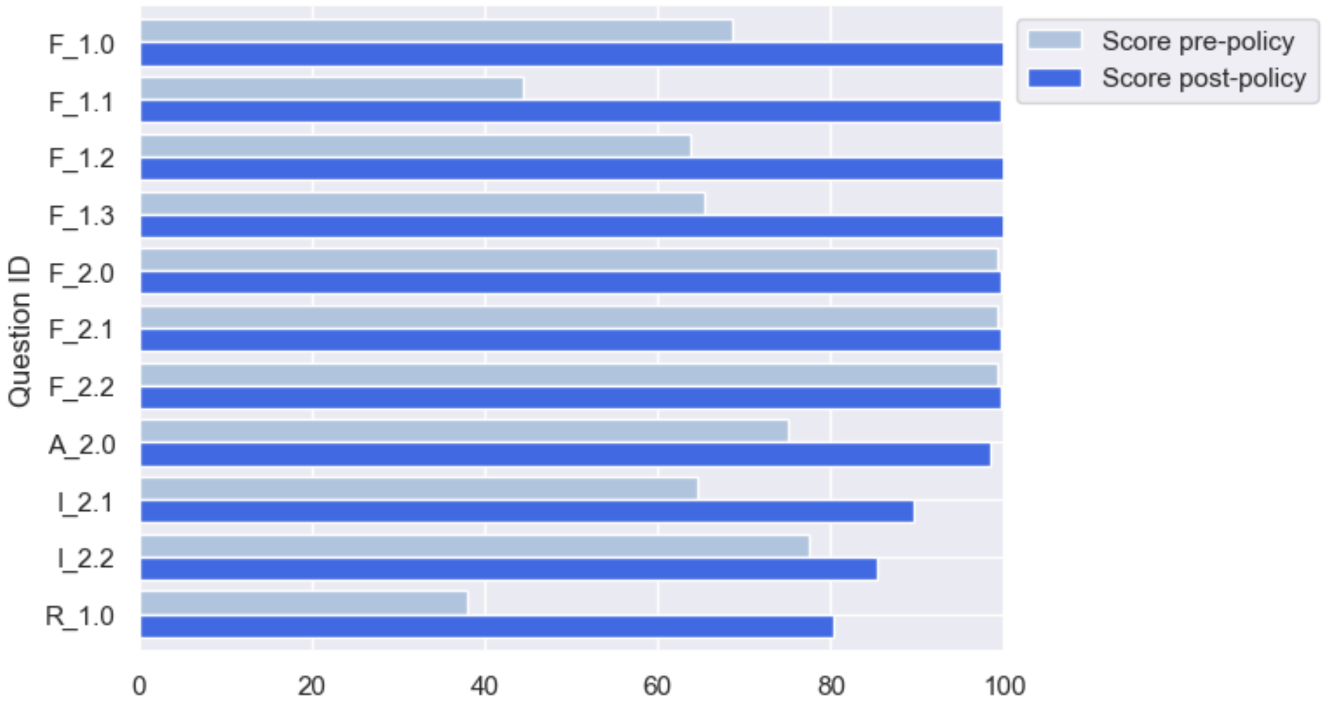


Key Findings

Pre and Post Policy:

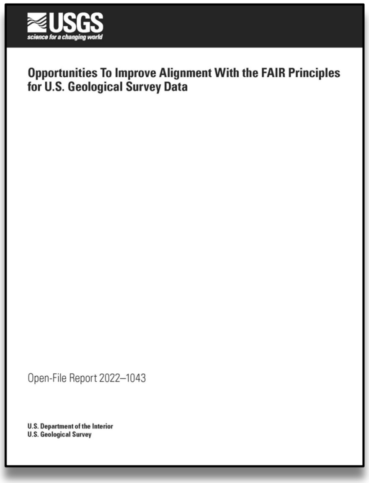
USGS introduced data management policies in 2016

11 questions in rubric address elements affected by the USGS data policy implementation, showing an increase in “Yes responses” for all questions.



Recommendations

Findings and recommendations resulting from the State of the Data analysis, align nicely with the recommendations in the CDI FAIR roadmap publication

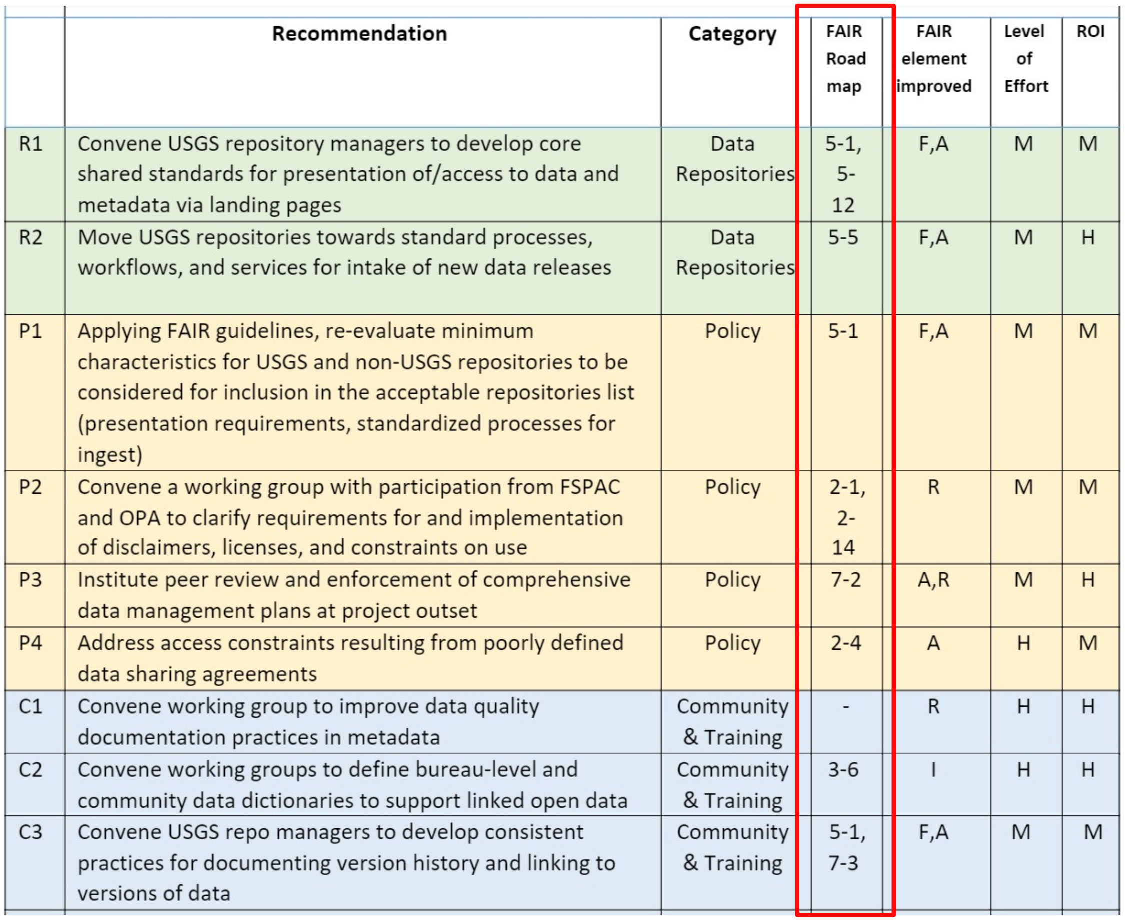


USGS
science for a changing world

Opportunities To Improve Alignment With the FAIR Principles for U.S. Geological Survey Data

Open-File Report 2022–1043

U.S. Department of the Interior
U.S. Geological Survey



	Recommendation	Category	FAIR Road map	FAIR element improved	Level of Effort	ROI
R1	Convene USGS repository managers to develop core shared standards for presentation of/access to data and metadata via landing pages	Data Repositories	5-1, 5-12	F,A	M	M
R2	Move USGS repositories towards standard processes, workflows, and services for intake of new data releases	Data Repositories	5-5	F,A	M	H
P1	Applying FAIR guidelines, re-evaluate minimum characteristics for USGS and non-USGS repositories to be considered for inclusion in the acceptable repositories list (presentation requirements, standardized processes for ingest)	Policy	5-1	F,A	M	M
P2	Convene a working group with participation from FSPAC and OPA to clarify requirements for and implementation of disclaimers, licenses, and constraints on use	Policy	2-1, 2-14	R	M	M
P3	Institute peer review and enforcement of comprehensive data management plans at project outset	Policy	7-2	A,R	M	H
P4	Address access constraints resulting from poorly defined data sharing agreements	Policy	2-4	A	H	M
C1	Convene working group to improve data quality documentation practices in metadata	Community & Training	-	R	H	H
C2	Convene working groups to define bureau-level and community data dictionaries to support linked open data	Community & Training	3-6	I	H	H
C3	Convene USGS repo managers to develop consistent practices for documenting version history and linking to versions of data	Community & Training	5-1, 7-3	F,A	M	M

Next Phases

- Develop a method for automated analysis of datasets for FAIRness
- Test the use of Artificial Intelligence to conduct the next analysis and compare to baseline
- Ensure training, and action on other recommendations, occurs based on results
- Use the State of the Data report to continue to increase community engagement in expanding a USGS culture of FAIR

Thank you!

Viv Hutchison

US Geological Survey
vhutchison@usgs.gov

Making Biomedical Data “Born FAIR”

Mark A. Musen, M.D., Ph.D
Stanford University

musen@stanford.edu

CEDAR

CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL

The FAIR Guiding Principles

F1: (Meta) data are assigned globally unique and persistent identifiers

F2: Data are described with rich metadata

F3: Metadata clearly and explicitly include the identifier of the data they describe

F4: (Meta)data are registered or indexed in a searchable resource

A1: (Meta)data are retrievable by their identifier using a standardised communication protocol

A1.1: The protocol is open, free and universally implementable

A1.2: The protocol allows for an authentication and authorisation where necessary

A2: Metadata should be accessible even when the data is no longer available

I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

I2: (Meta)data use vocabularies that follow the FAIR principles

I3: (Meta)data include qualified references to other (meta)data

R1: (Meta)data are richly described with a plurality of accurate and relevant attributes

R1.1: (Meta)data are released with a clear and accessible data usage license

R1.2: (Meta)data are associated with detailed provenance

R1.3: (Meta)data meet domain-relevant community standards

Most FAIR principles are about *metadata*

F1: (Meta) data are assigned globally unique and persistent identifiers

F2: Data are described with rich metadata

F3: Metadata clearly and explicitly include the identifier of the data they describe

F4: (Meta)data are registered or indexed in a searchable resource

A1: (Meta)data are retrievable by their identifier using a standardised communication protocol

A1.1: The protocol is open, free and universally implementable

A1.2: The protocol allows for an authentication and authorisation where necessary

A2: Metadata should be accessible even when the data is no longer available

I1: (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

I2: (Meta)data use vocabularies that follow the FAIR principles

I3: (Meta)data include qualified references to other (meta)data

R1: (Meta)data are richly described with a plurality of accurate and relevant attributes

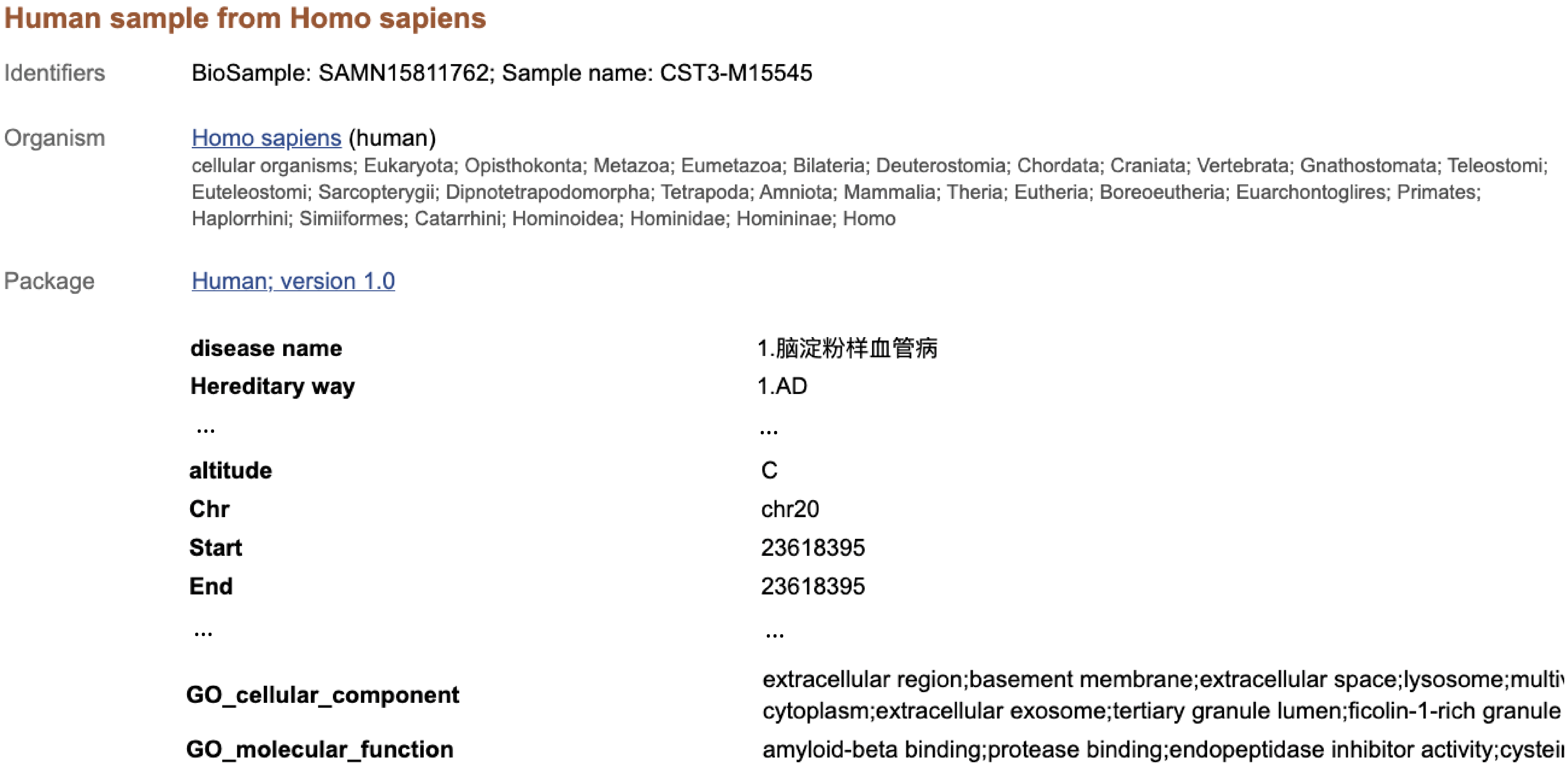
R1.1: (Meta)data are released with a clear and accessible data usage license

R1.2: (Meta)data are associated with detailed provenance

R1.3: (Meta)data meet domain-relevant community standards

Metadata in public repositories are a mess!

- Investigators view their work as publishing papers, not leaving a legacy of reusable data
- Sponsors may require data sharing, but they do not encourage the use of grant funds to pay for it
- Creating the metadata to describe data sets is unbearably hard



Human sample from Homo sapiens

Identifiers BioSample: SAMN15811762; Sample name: CST3-M15545

Organism Homo sapiens (human)
cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homininae; Homo

Package Human; version 1.0

disease name	1.脑淀粉样血管病
Hereditary way	1.AD
...	...
altitude	C
Chr	chr20
Start	23618395
End	23618395
...	...
GO_cellular_component	extracellular region;basement membrane;extracellular space;lysosome;multi cytoplasm;extracellular exosome;tertiary granule lumen;ficolin-1-rich granule
GO_molecular_function	amyloid-beta binding;protease binding;endopeptidase inhibitor activity;cystei

Full metadata record available at: https://www.ncbi.nlm.nih.gov/biosample/15811762

Metadata need to adhere to standards!

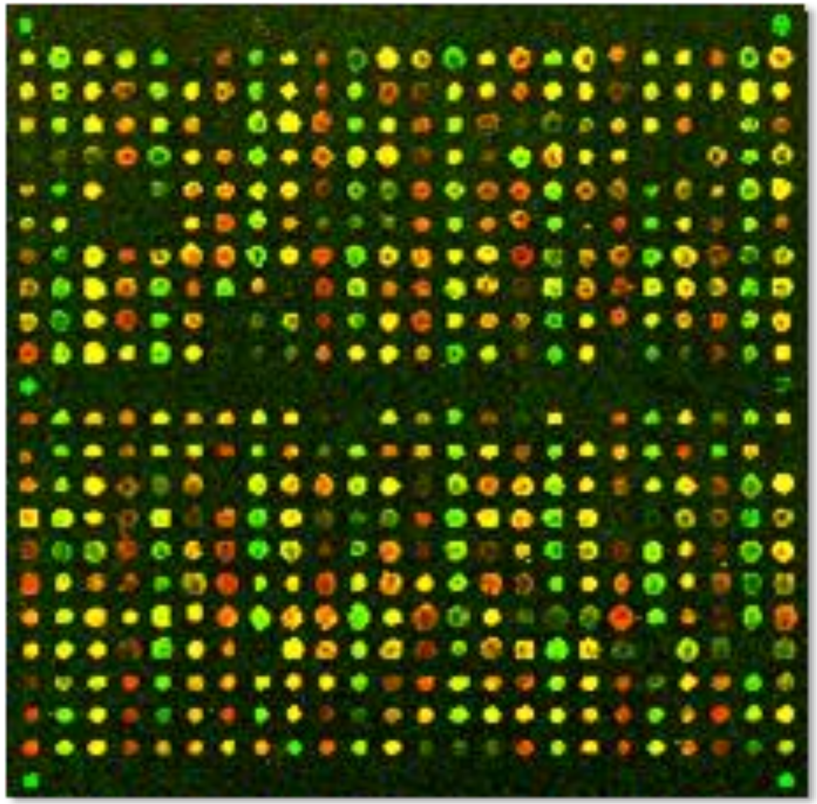
age
Age
AGE
`Age
age (after birth)
age (in years)
age (y)
age (year)
age (years)
Age (years)
Age (Years)
age (yr)
age (yr-old)
age (yrs)
Age (yrs)

age [y]
age [year]
age [years]
age in years
age of patient
Age of patient
age of subjects
age(years)
Age(years)
Age(yrs.)
Age, year
age, years
age, yrs
age.year
age_years

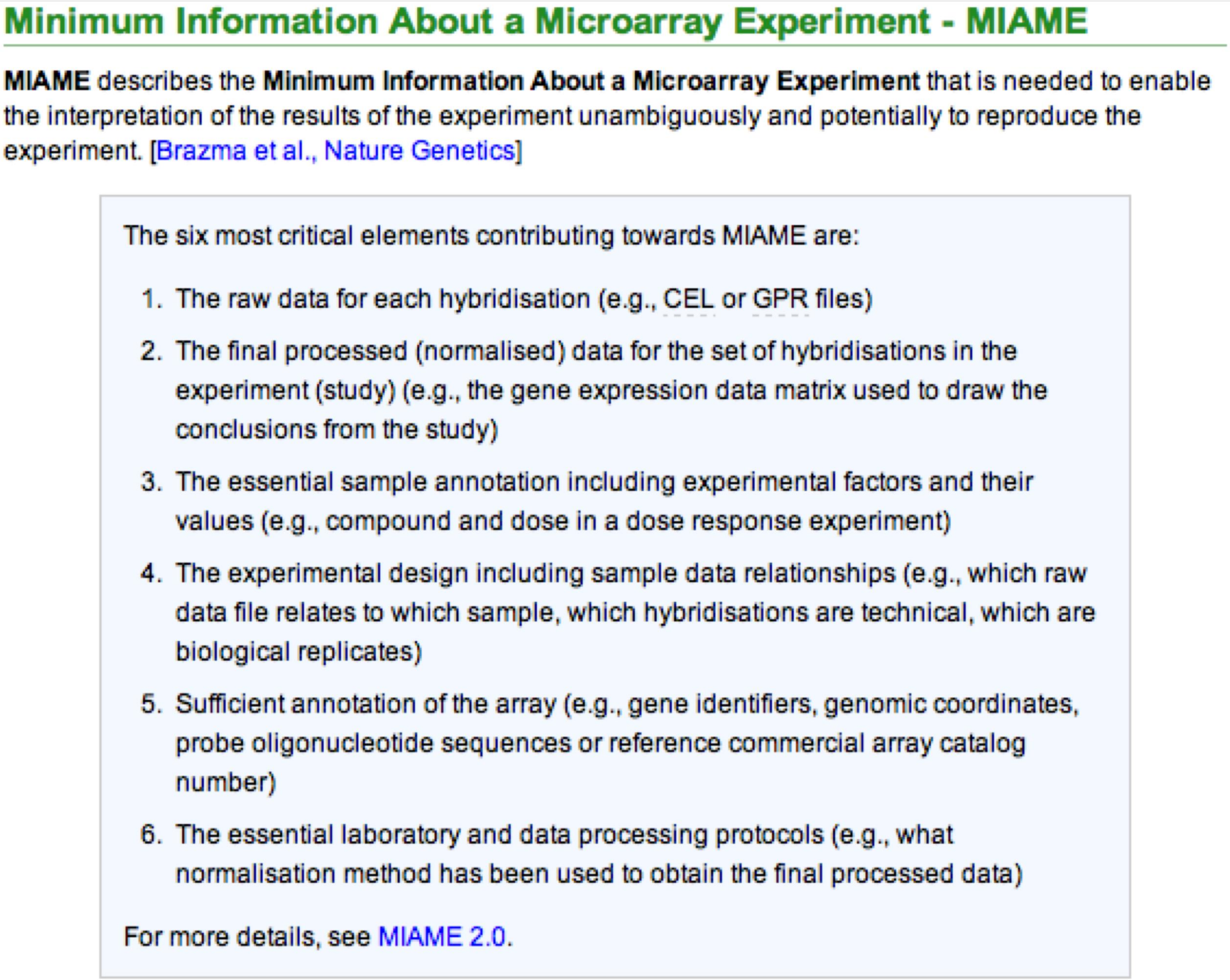
Gene Expression Omnibus

The microarray community took the lead in standardizing metadata **reporting guidelines**

- What was the substrate of the experiment?
- What array platform was used?
- What were the experimental conditions?



DNA Microarray



Minimum Information About a Microarray Experiment - MIAME

MIAME describes the **Minimum Information About a Microarray Experiment** that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment. [Brazma et al., Nature Genetics]

The six most critical elements contributing towards MIAME are:

1. The raw data for each hybridisation (e.g., CEL or GPR files)
2. The final processed (normalised) data for the set of hybridisations in the experiment (study) (e.g., the gene expression data matrix used to draw the conclusions from the study)
3. The essential sample annotation including experimental factors and their values (e.g., compound and dose in a dose response experiment)
4. The experimental design including sample data relationships (e.g., which raw data file relates to which sample, which hybridisations are technical, which are biological replicates)
5. Sufficient annotation of the array (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalog number)
6. The essential laboratory and data processing protocols (e.g., what normalisation method has been used to obtain the final processed data)

For more details, see MIAME 2.0.

But it didn’t stop with MIAME!

- Minimal Information About T Cell Assays (MIATA)
- Minimal Information Required in the Annotation of biochemical Models (MIRIAM)
- MINImal MEtagemome Sequence analysis Standard (MINIMESS)
- Minimal Information Specification For In Situ Hybridization and Immunohistochemistry Experiments (MISFISHIE)

These are exactly the kinds of community standards
that we need to structure metadata!

If we want to have FAIR data, we need good metadata. Good metadata need:

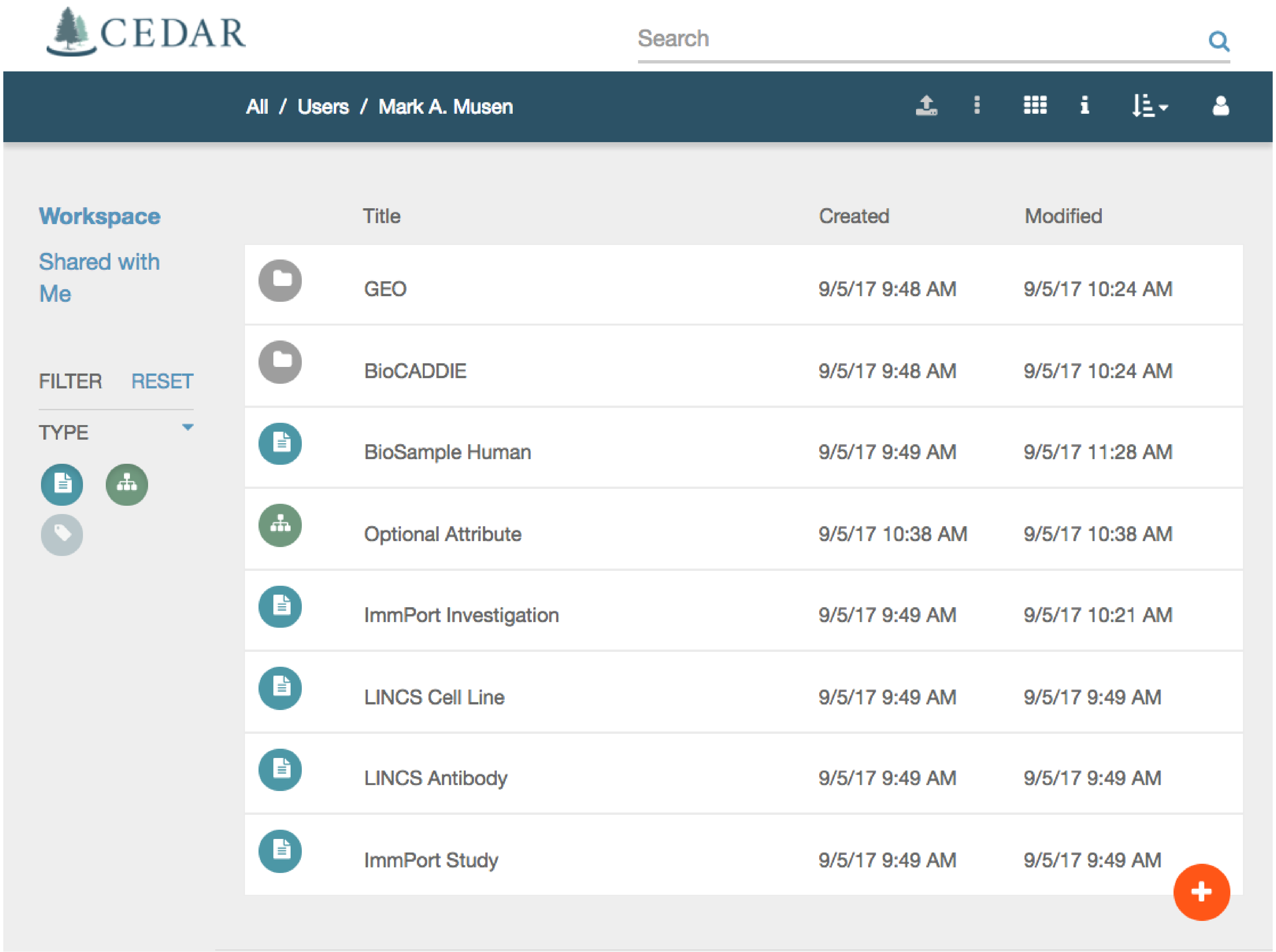
- **Ontologies** to provide controlled terms
- **Reporting guidelines**—like MIAME—to provide a standardized structure for the metadata components
- **Technology** to make it easy to author good metadata in the first place
- **Procedures** to create community-based standards in the first place

Our approach in CEDAR

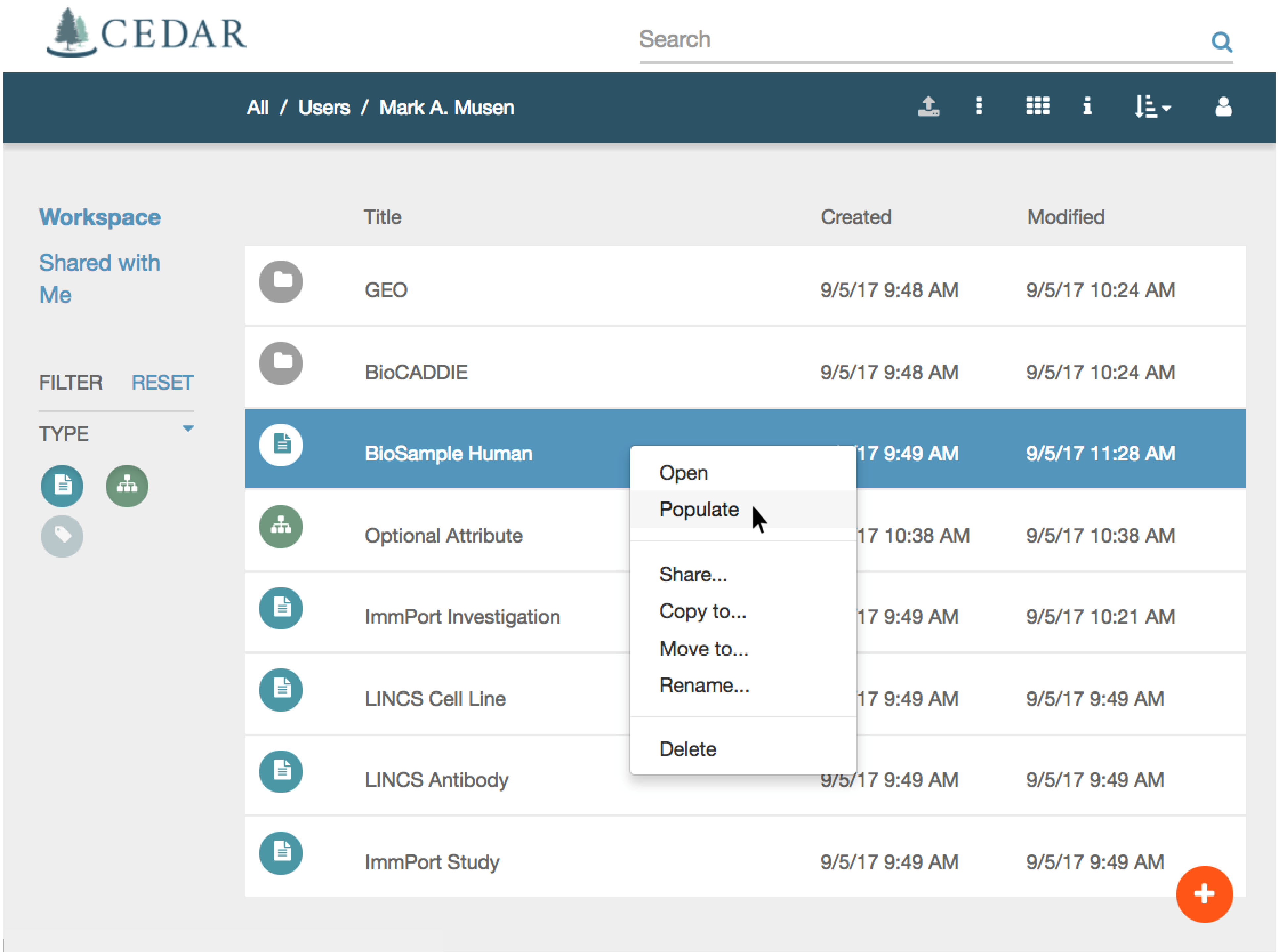
- Encode standard, community-endorsed *reporting guidelines* as **templates** that offer fill-in-the-blank authoring opportunities
- Use selections from *ontologies* whenever possible to provide **standardized values** for the template fields

CEDAR

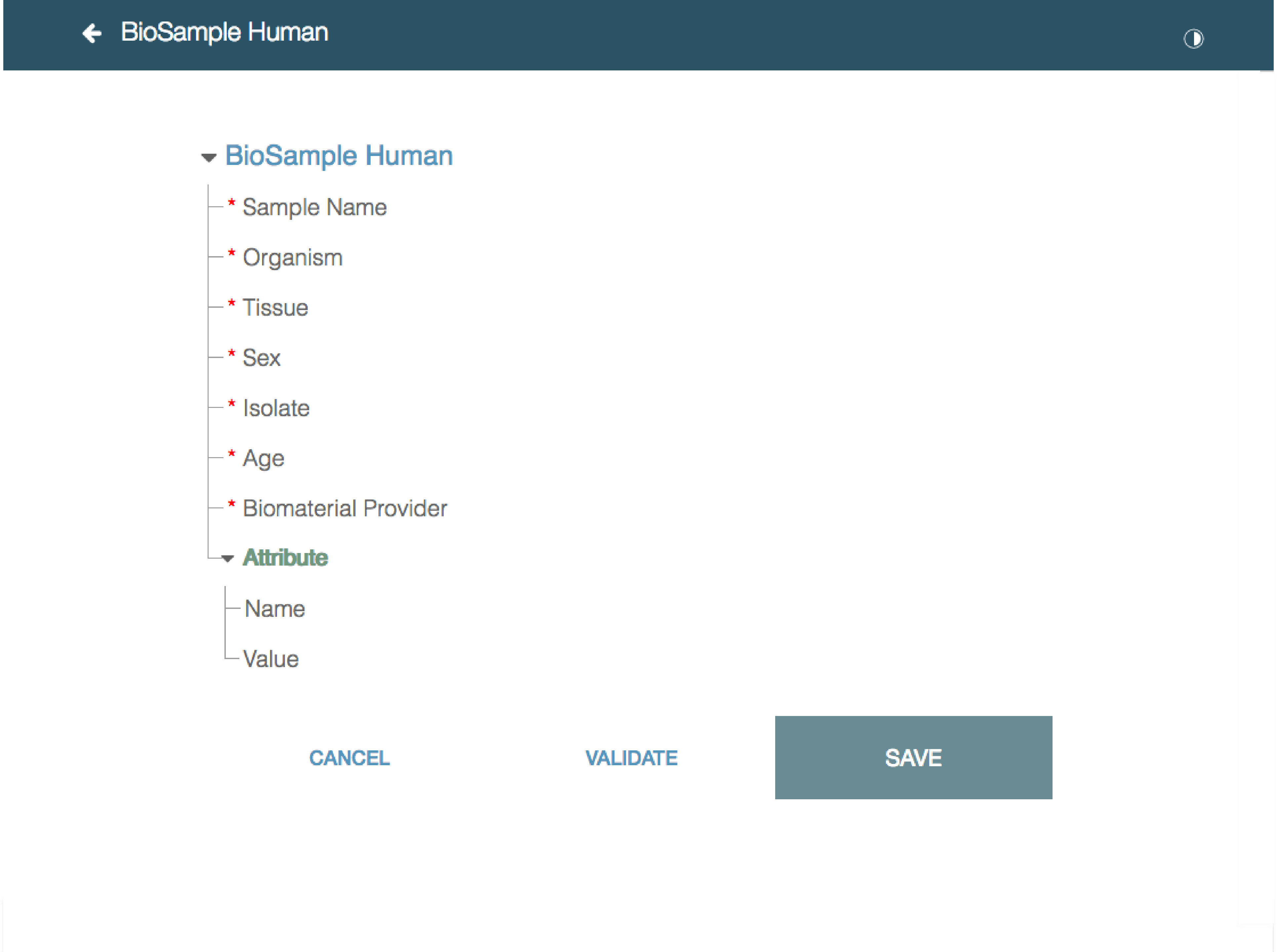
CENTER FOR EXPANDED DATA
ANNOTATION AND RETRIEVAL



Title	Created	Modified
GEO	9/5/17 9:48 AM	9/5/17 10:24 AM
BioCADDIE	9/5/17 9:48 AM	9/5/17 10:24 AM
BioSample Human	9/5/17 9:49 AM	9/5/17 11:28 AM
Optional Attribute	9/5/17 10:38 AM	9/5/17 10:38 AM
ImmPort Investigation	9/5/17 9:49 AM	9/5/17 10:21 AM
LINCS Cell Line	9/5/17 9:49 AM	9/5/17 9:49 AM
LINCS Antibody	9/5/17 9:49 AM	9/5/17 9:49 AM
ImmPort Study	9/5/17 9:49 AM	9/5/17 9:49 AM



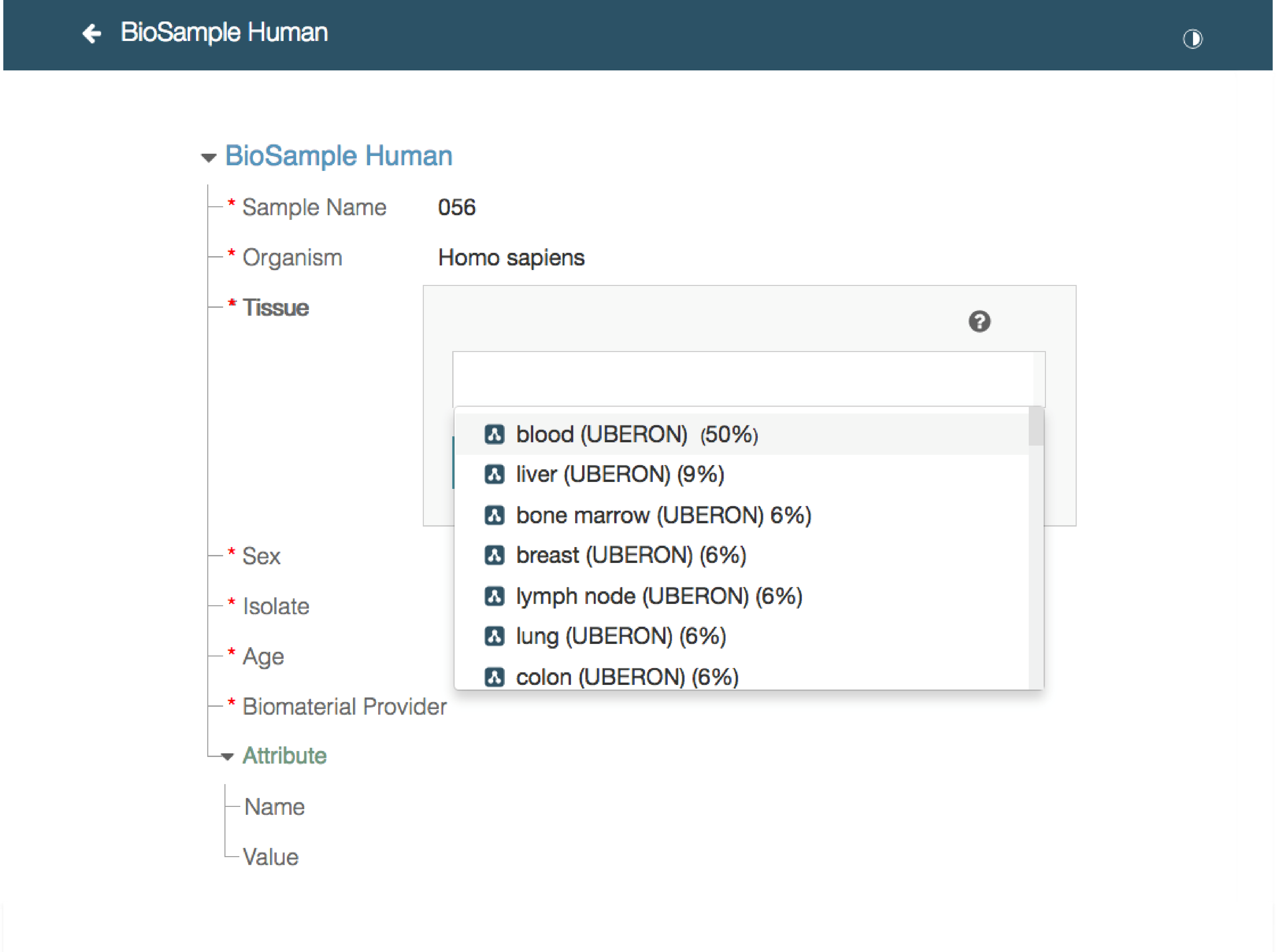
Title	Created	Modified
GEO	9/5/17 9:48 AM	9/5/17 10:24 AM
BioCADDIE	9/5/17 9:48 AM	9/5/17 10:24 AM
BioSample Human	17 9:49 AM	9/5/17 11:28 AM
Optional Attribute	17 10:38 AM	9/5/17 10:38 AM
ImmPort Investigation	17 9:49 AM	9/5/17 10:21 AM
LINCS Cell Line	17 9:49 AM	9/5/17 9:49 AM
LINCS Antibody	9/5/17 9:49 AM	9/5/17 9:49 AM
ImmPort Study	9/5/17 9:49 AM	9/5/17 9:49 AM



BioSample Human

BioSample Human

- * Sample Name
- * Organism
- * Tissue
- * Sex
- * Isolate
- * Age
- * Biomaterial Provider
- Attribute
  - Name
  - Value



BioSample Human

BioSample Human

* Sample Name 056

* Organism Homo sapiens

* Tissue

- blood (UBERON) (50%)
- liver (UBERON) (9%)
- bone marrow (UBERON) 6%)
- breast (UBERON) (6%)
- lymph node (UBERON) (6%)
- lung (UBERON) (6%)
- colon (UBERON) (6%)

* Sex

* Isolate

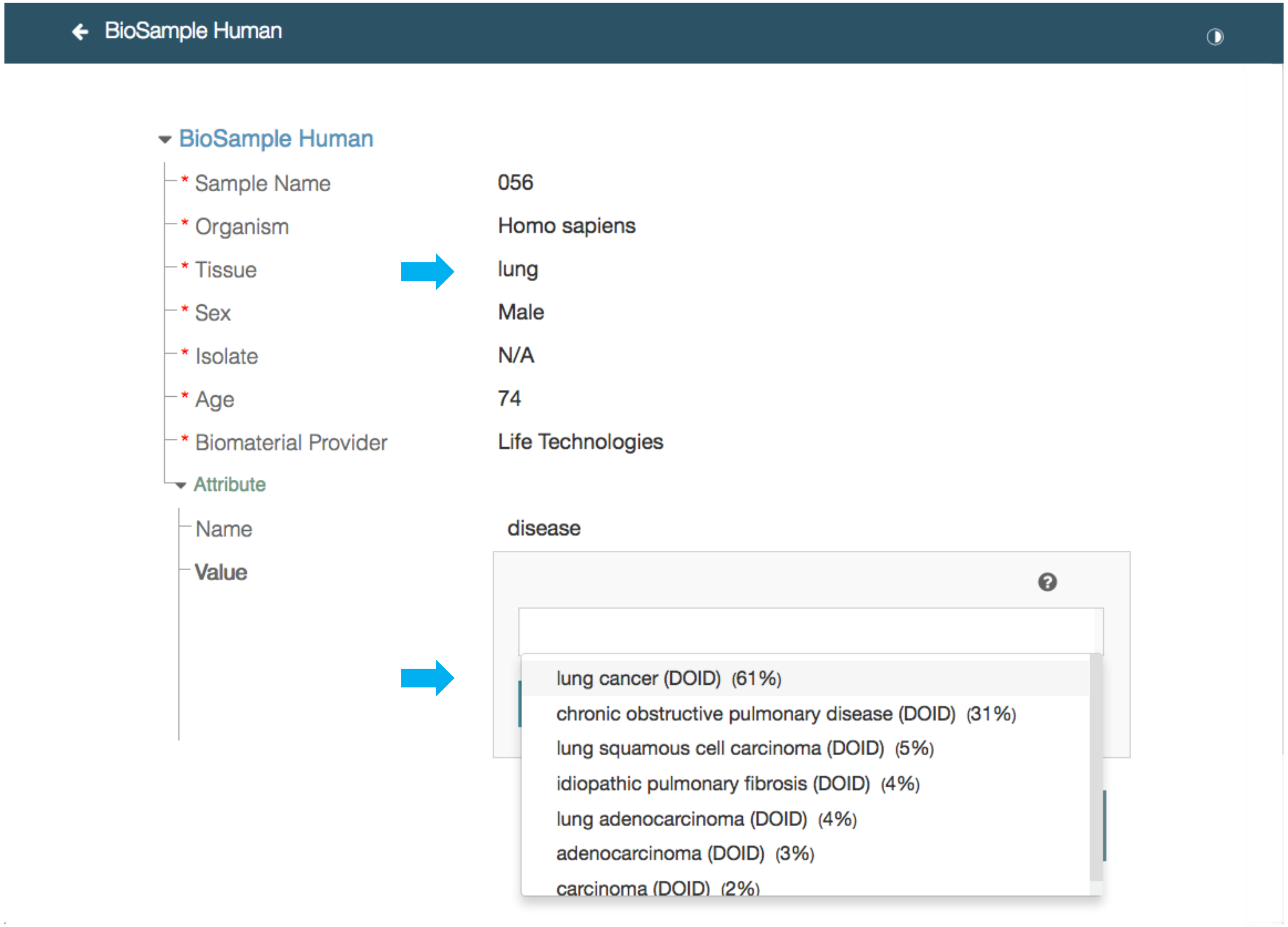
* Age

* Biomaterial Provider

Attribute

Name

Value



BioSample Human

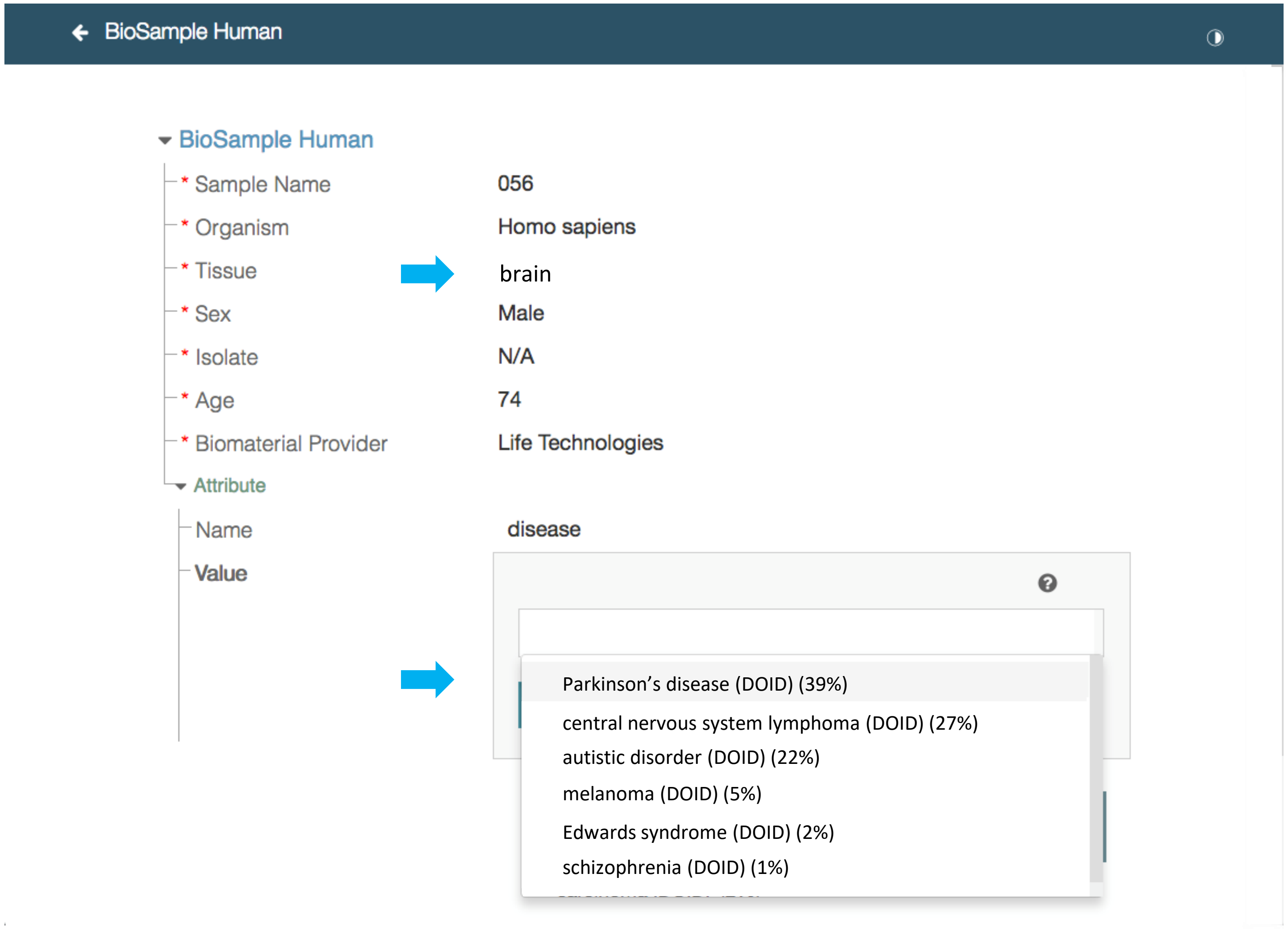
BioSample Human

* Sample Name	056
* Organism	Homo sapiens
* Tissue	lung
* Sex	Male
* Isolate	N/A
* Age	74
* Biomaterial Provider	Life Technologies

Attribute

Name	disease
Value	

- lung cancer (DOID) (61%)
- chronic obstructive pulmonary disease (DOID) (31%)
- lung squamous cell carcinoma (DOID) (5%)
- idiopathic pulmonary fibrosis (DOID) (4%)
- lung adenocarcinoma (DOID) (4%)
- adenocarcinoma (DOID) (3%)
- carcinoma (DOID) (2%)



BioSample Human

* Sample Name	056
* Organism	Homo sapiens
* Tissue	brain
* Sex	Male
* Isolate	N/A
* Age	74
* Biomaterial Provider	Life Technologies

Attribute

Name	disease
Value	

- Parkinson's disease (DOID) (39%)
- central nervous system lymphoma (DOID) (27%)
- autistic disorder (DOID) (22%)
- melanoma (DOID) (5%)
- Edwards syndrome (DOID) (2%)
- schizophrenia (DOID) (1%)

VODAN AFRICA & ASIA

NIH HEAL INITIATIVE

HuBMAP

RADX DATA HUB

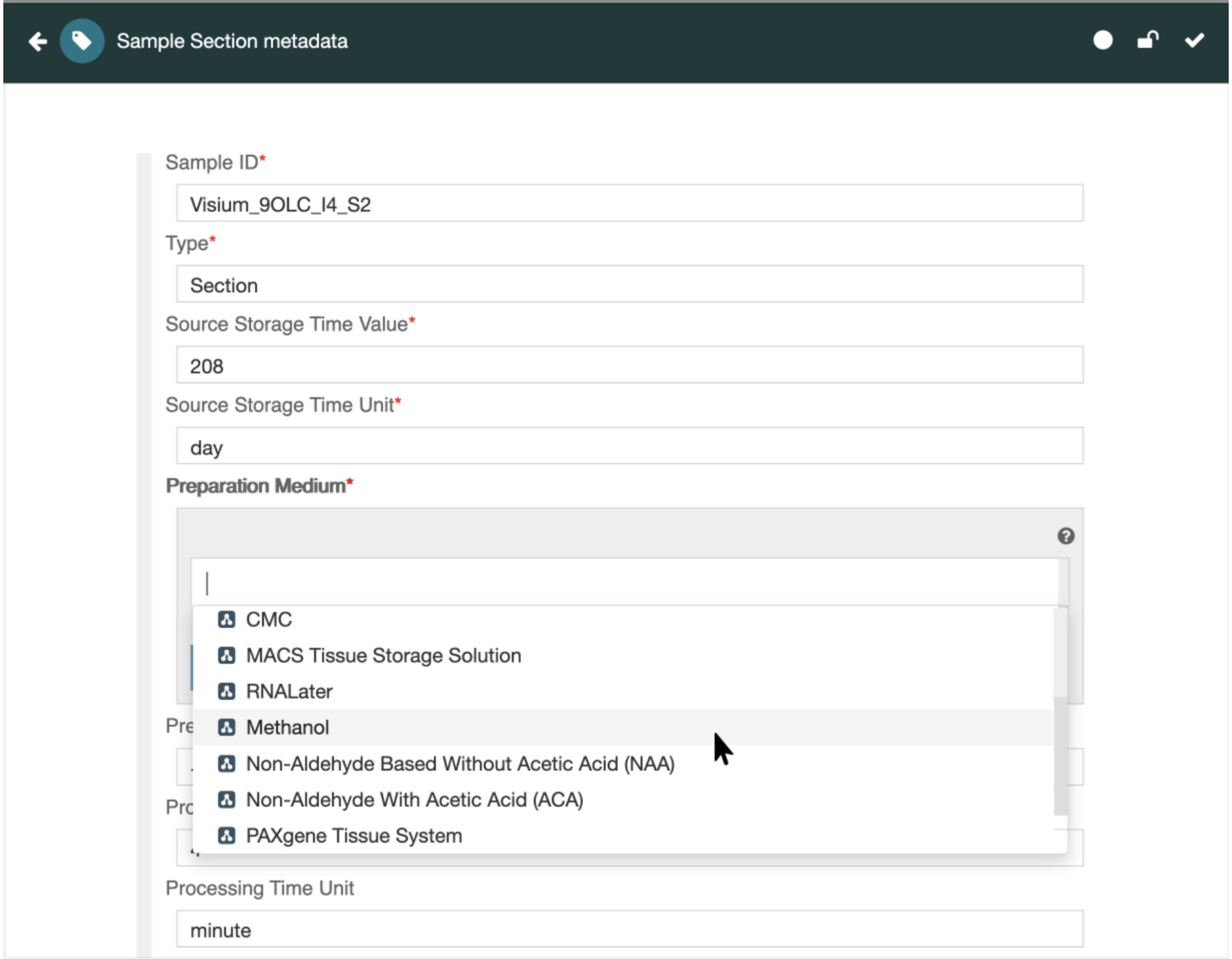
MCBK

Mobilizing Computable Biomedical Knowledge

IDG

Illuminating the Druggable Genome

HuBMAP Metadata Spreadsheet Validator



Sample Section metadata

Sample ID*

Visium_9OLC_I4_S2

Type*

Section

Source Storage Time Value*

208

Source Storage Time Unit*

day

Preparation Medium*

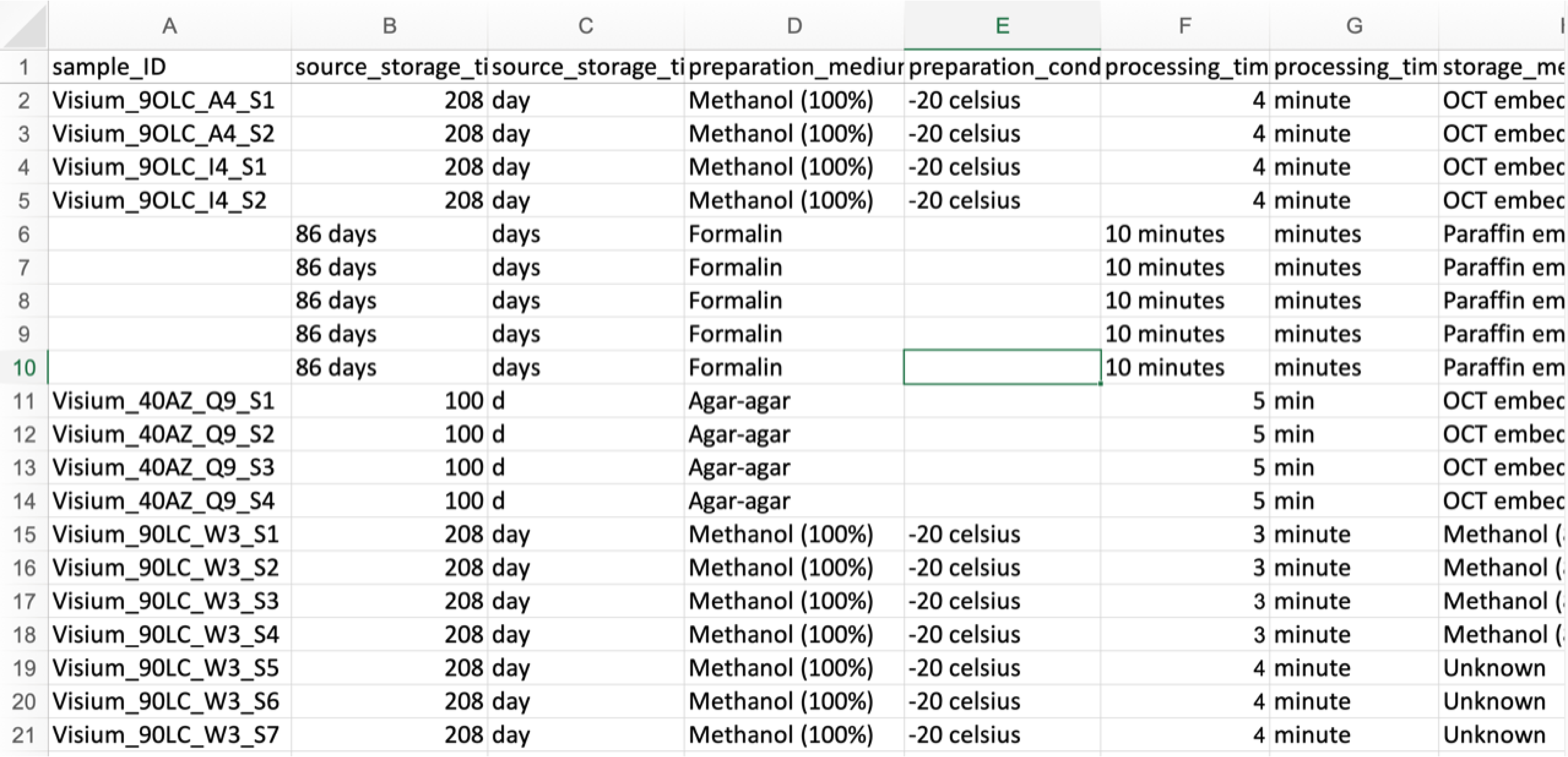
- CMC
- MACS Tissue Storage Solution
- RNALater
- Methanol
- Non-Aldehyde Based Without Acetic Acid (NAA)
- Non-Aldehyde With Acetic Acid (ACA)
- PAXgene Tissue System

Pre

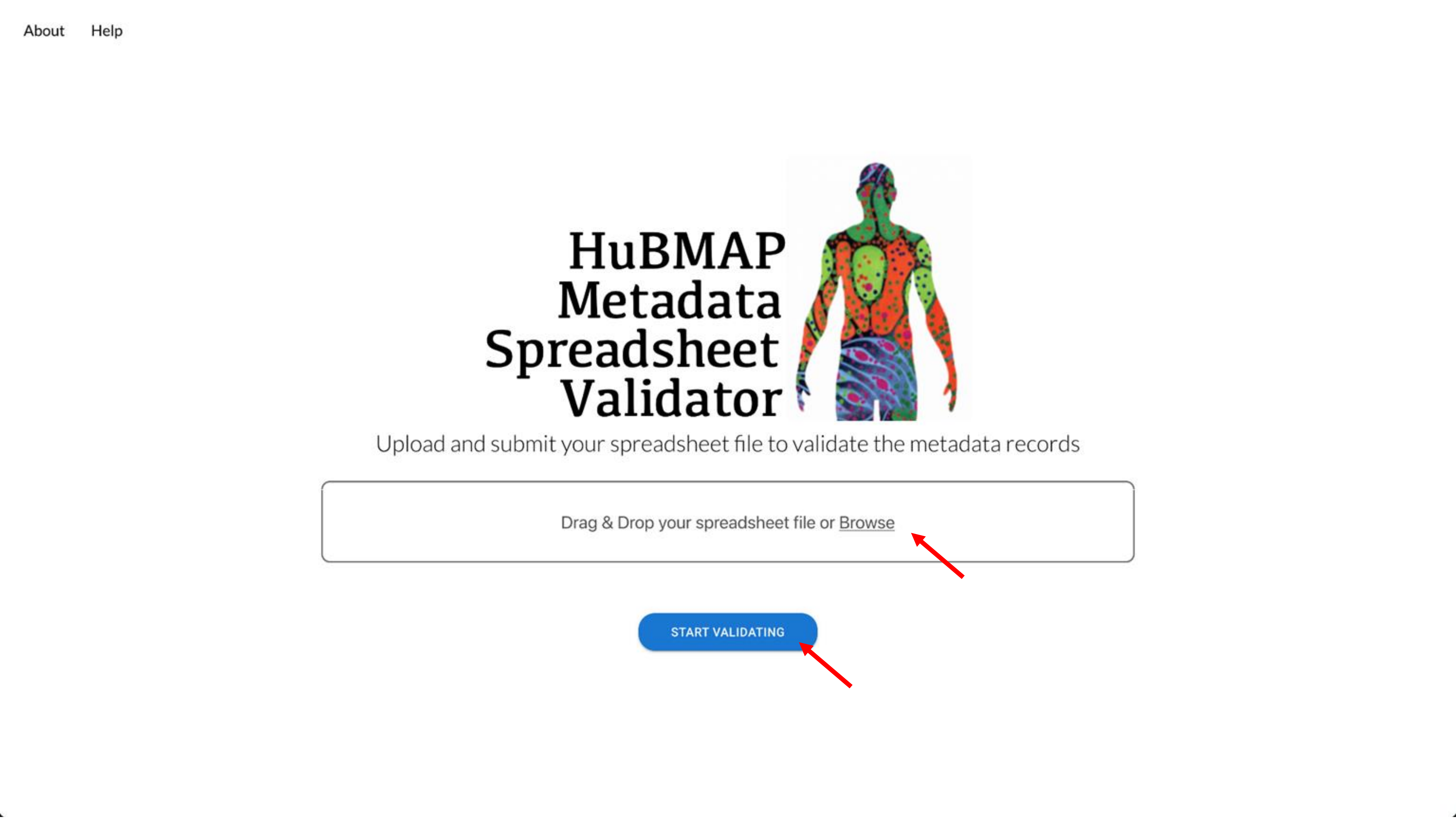
Pro

Processing Time Unit

minute



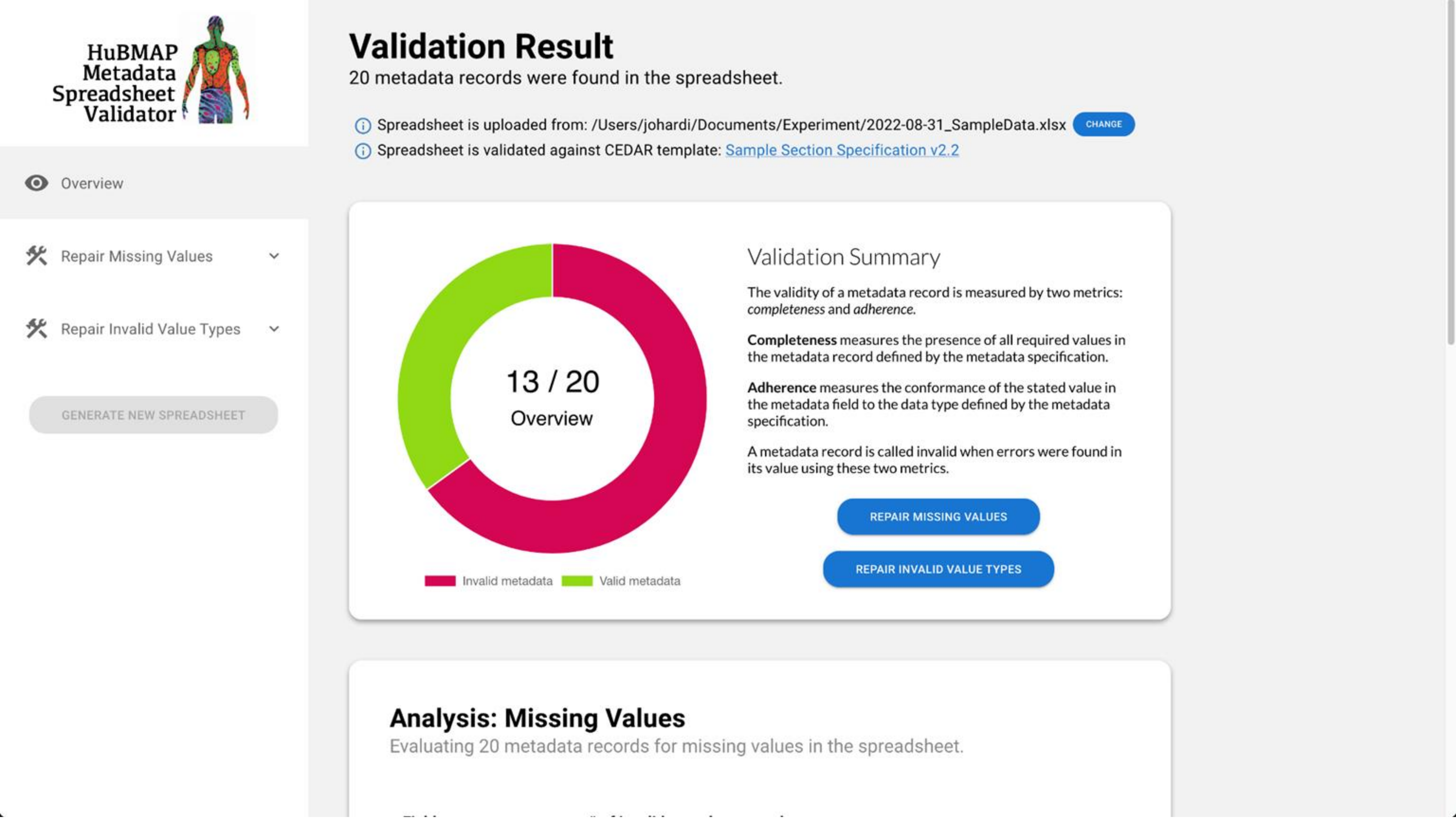
	A	B	C	D	E	F	G	
1	sample_ID	source_storage_ti	source_storage_ti	preparation_mediur	preparation_cond	processing_tim	processing_tim	storage_me
2	Visium_9OLC_A4_S1	208	day	Methanol (100%)	-20 celsius	4	minute	OCT embec
3	Visium_9OLC_A4_S2	208	day	Methanol (100%)	-20 celsius	4	minute	OCT embec
4	Visium_9OLC_I4_S1	208	day	Methanol (100%)	-20 celsius	4	minute	OCT embec
5	Visium_9OLC_I4_S2	208	day	Methanol (100%)	-20 celsius	4	minute	OCT embec
6		86 days	days	Formalin		10 minutes	minutes	Paraffin em
7		86 days	days	Formalin		10 minutes	minutes	Paraffin em
8		86 days	days	Formalin		10 minutes	minutes	Paraffin em
9		86 days	days	Formalin		10 minutes	minutes	Paraffin em
10		86 days	days	Formalin		10 minutes	minutes	Paraffin em
11	Visium_40AZ_Q9_S1	100	d	Agar-agar		5	min	OCT embec
12	Visium_40AZ_Q9_S2	100	d	Agar-agar		5	min	OCT embec
13	Visium_40AZ_Q9_S3	100	d	Agar-agar		5	min	OCT embec
14	Visium_40AZ_Q9_S4	100	d	Agar-agar		5	min	OCT embec
15	Visium_90LC_W3_S1	208	day	Methanol (100%)	-20 celsius	3	minute	Methanol (
16	Visium_90LC_W3_S2	208	day	Methanol (100%)	-20 celsius	3	minute	Methanol (
17	Visium_90LC_W3_S3	208	day	Methanol (100%)	-20 celsius	3	minute	Methanol (
18	Visium_90LC_W3_S4	208	day	Methanol (100%)	-20 celsius	3	minute	Methanol (
19	Visium_90LC_W3_S5	208	day	Methanol (100%)	-20 celsius	4	minute	Unknown
20	Visium_90LC_W3_S6	208	day	Methanol (100%)	-20 celsius	4	minute	Unknown
21	Visium_90LC_W3_S7	208	day	Methanol (100%)	-20 celsius	4	minute	Unknown



HuBMAP Metadata Spreadsheet Validator

Upload and submit your spreadsheet file to validate the metadata records

Drag & Drop your spreadsheet file or Browse



HuBMAP Metadata Spreadsheet Validator

Overview

Repair Missing Values

Repair Invalid Value Types

GENERATE NEW SPREADSHEET

Validation Result

20 metadata records were found in the spreadsheet.

ⓘ Spreadsheet is uploaded from: /Users/johardi/Documents/Experiment/2022-08-31_SampleData.xlsx CHANGE

ⓘ Spreadsheet is validated against CEDAR template: Sample Section Specification v2.2

Validation Summary

The validity of a metadata record is measured by two metrics: *completeness* and *adherence.*

Completeness measures the presence of all required values in the metadata record defined by the metadata specification.

Adherence measures the conformance of the stated value in the metadata field to the data type defined by the metadata specification.

A metadata record is called invalid when errors were found in its value using these two metrics.

REPAIR MISSING VALUES

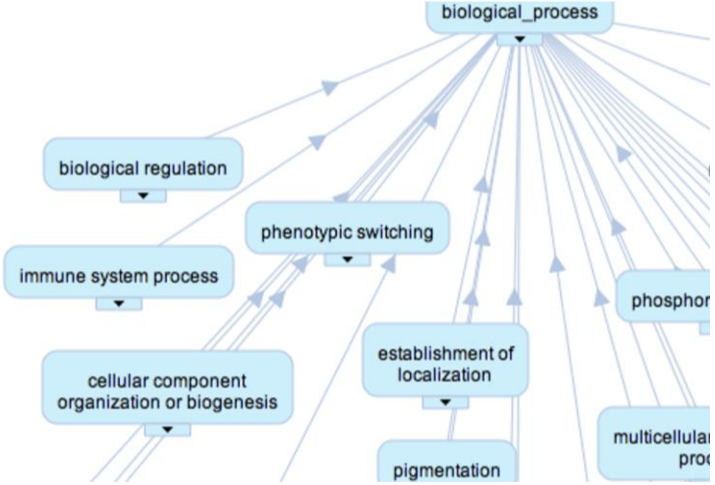
REPAIR INVALID VALUE TYPES

Analysis: Missing Values

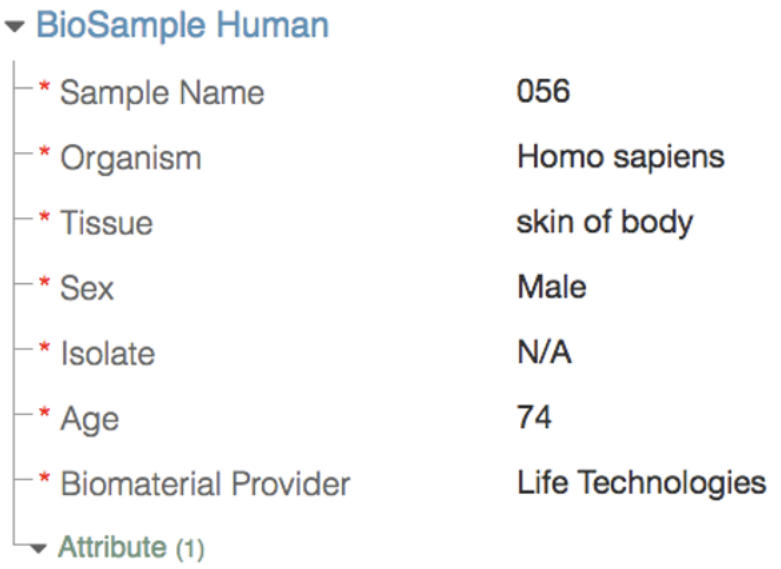
Evaluating 20 metadata records for missing values in the spreadsheet.

There are two kinds of community standards that guide the authoring of scientific metadata

1. **Ontologies**: Collections of standard terms for salient entities in a discipline (e.g., Gene Ontology, International Classification of Diseases)



2. **Reporting Guidelines**: Enumerations of those aspects of a class of experiment that useful metadata need to mention (e.g., Minimum Information About a Microrray Experiment; MIAME)



BioSample Human

* Sample Name	056
* Organism	Homo sapiens
* Tissue	skin of body
* Sex	Male
* Isolate	N/A
* Age	74
* Biomaterial Provider	Life Technologies

Attribute (1)

Online data will never be FAIR

- Until we standardize metadata structure using common **templates** to capture reporting guidelines
- Until we can fill in those templates with **controlled terms** whenever possible
- Until we create **technology** that will make it easy for investigators to annotate their datasets in standardized, searchable ways
- Until we recognize the importance of creating FAIR data from the very beginning

CEDAR
CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL