CEDAR: Promoting FAIRness at the Source

Mark A. Musen, M.D., Ph.D
Stanford University
musen@stanford.edu

CEDAR

CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL

Systems to evaluate data FAIRness have had difficulty finding an audience

- Scientists really don’t want a FAIR “report card”
- No one wants to hear about problems with datasets that have *already* been uploaded to a repository
- There is no fully computable solution to the question of whether a dataset is FAIR in the first place

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

Scientists have no direct control over repository infrastructure

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

FAIR principles depend on community standards that are not objectively computable

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

	A	B	C	E	F	G
1	# Use this template for 3' or whole Gene expression studies when summarization probe set data will be provided as **CHP files**.					
2	# Do **NOT** submit CHP files unless they are relevant to your analysis (instead, use the Matrix table option to submit the relevant data, e.g. **Bioconduct**					
3	# Incomplete submissions will be returned. Click the **Metadata Example** tab below to view a completed worksheet					
4	# A complete submission will consist of: (1) a completed metadata worksheet, (2) the CHP files, and (3) the original CEL files.					
5	**# Field names (in blue on this page) should not be edited. Hover over cells containing field names to view field content guidelines or,**					
6	**# CLICK HERE for Field Content Guidelines Web page.**					
7						
8	SERIES					
9	# This section describes the overall					
10						
11	title					
12	summary					
13	summary					
14	overall design					
15	contributor					
16	contributor					
17						
18	SAMPLES					
19	**# The Sample names in the first column are arbitrary but they must match the column headers of the Matrix table (see next worksheet).**					
20						
21	Sample name	title	CHP file	source name	organism	characteristics: tag
22	SAMPLE 1					
23	SAMPLE 2					
24	SAMPLE 3					
25	SAMPLE 4					
26	SAMPLE 5					
27	SAMPLE 6					
28	SAMPLE 7					
29	SAMPLE 8					
30	SAMPLE 9					
31	SAMPLE X					
32						
33						
34	PROTOCOLS					
35	# This section includes protocols and fields which are common to all Samples.					
36	# Protocols which are applicable to specific Samples or specific channels should be included in additional columns of the SAMPLES section instead.					
37						
38	growth protocol					
39	treatment protocol					
40	extract protocol					
41	label protocol					
42	hyb protocol					

Unique title (less than 120 characters) that describes the overall study.

"Firstname,Initial,Lastname". Example: "John,H,Smith" or "Jane,Doe".

Unique title that describes the Sample. We suggest that you use the convention: [biomaterial]-[condition(s)]-[replicate number], e.g., Muscle_exercised_60min_rep2.

Replace 'tag' with a biosource characteristic (e.g. "gender", "strain", "tissue", "developmental stage", "tumor stage", etc), and then enter the value for each sample beneath (e.g. "female", "129SV", "brain", "embryo", etc). You may add additional characteristics columns to this template (see 'Metadata Example' spreadsheet).

[Optional] Describe the conditions that were used to grow or maintain organisms or cells prior to extract preparation.

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!

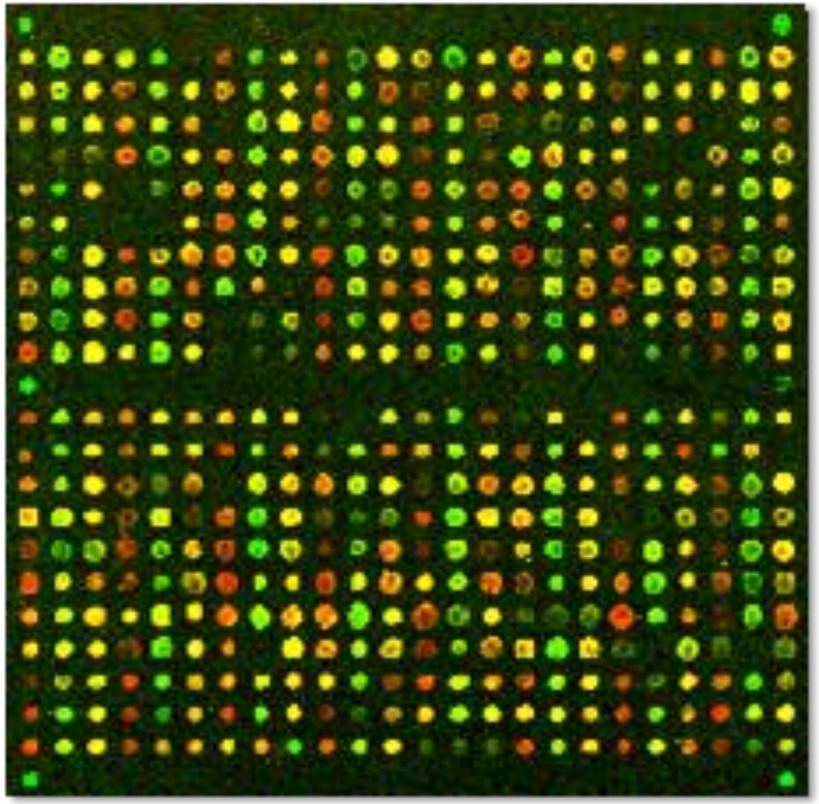
age
Age
AGE
`Age
age (after birth)
age (in years)
age (y)
age (year)
age (years)
Age (years)
Age (Years)
age (yr)
age (yr-old)
age (yrs)
Age (yrs)

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

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

F2: Data are described with rich metadata

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

Don't even try to measure FAIRness. Make data FAIR from the beginning!

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

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

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

CEDAR

Search

All / Users / Mark A. Musen

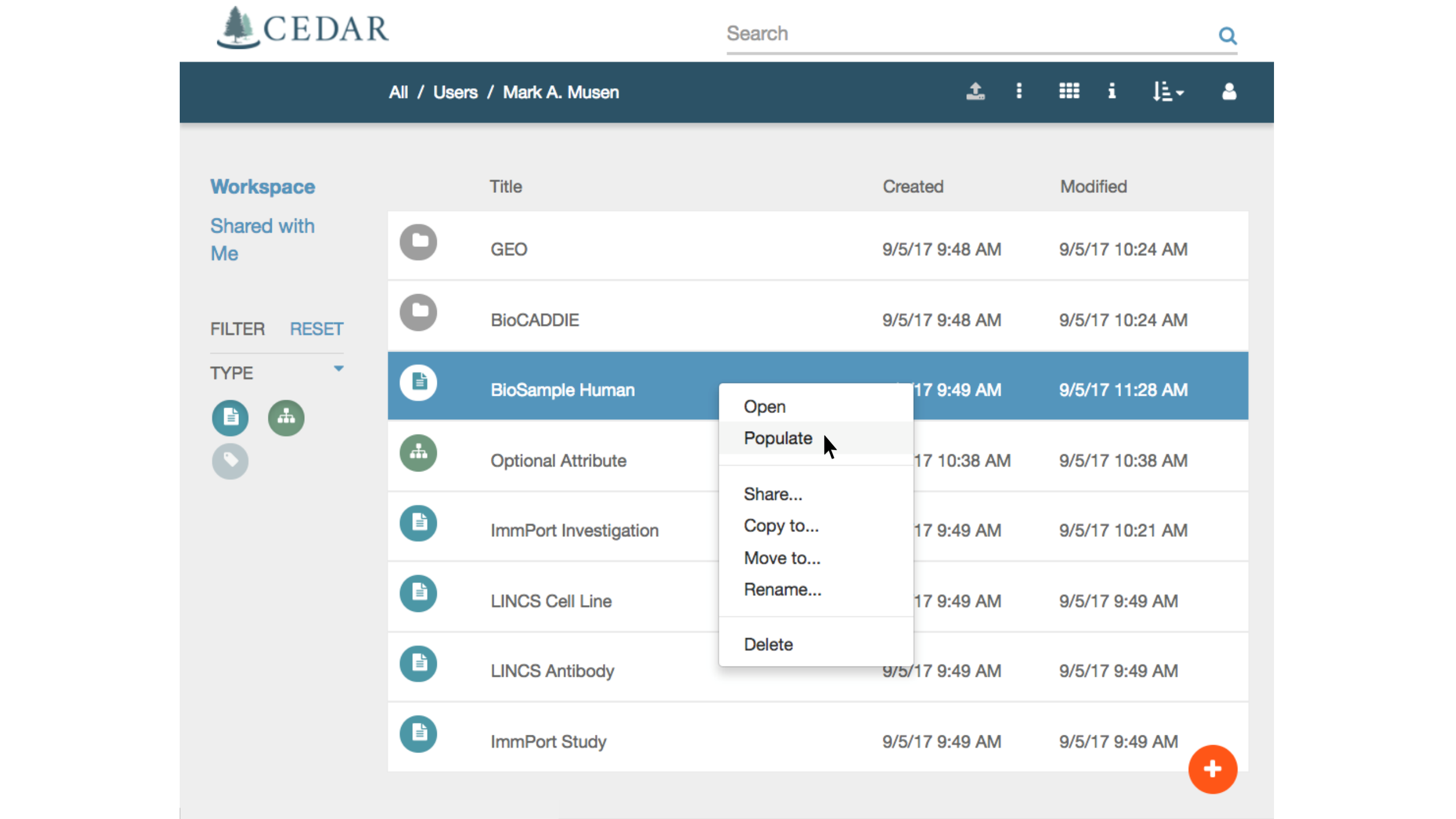
Workspace

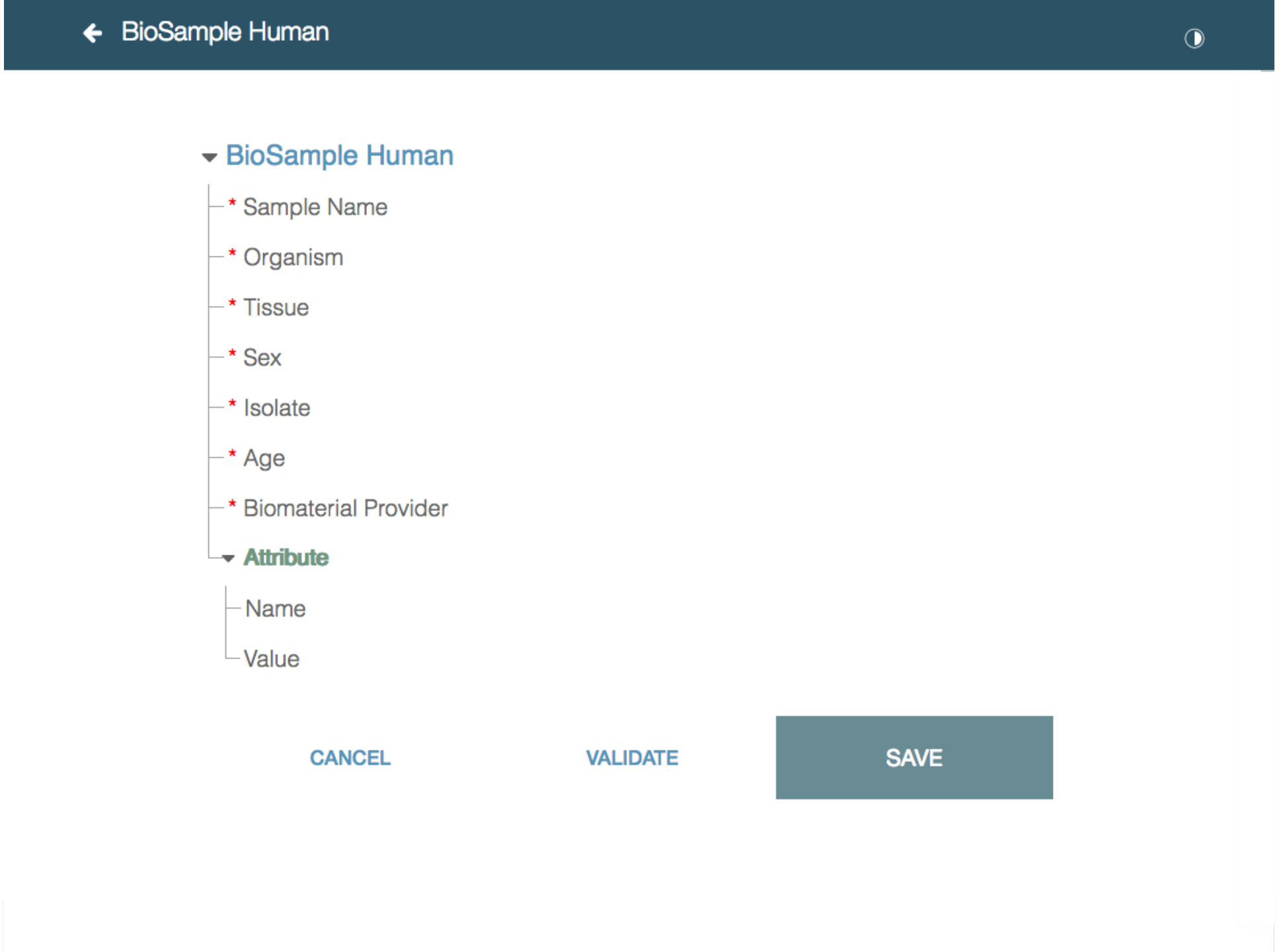
Shared with Me

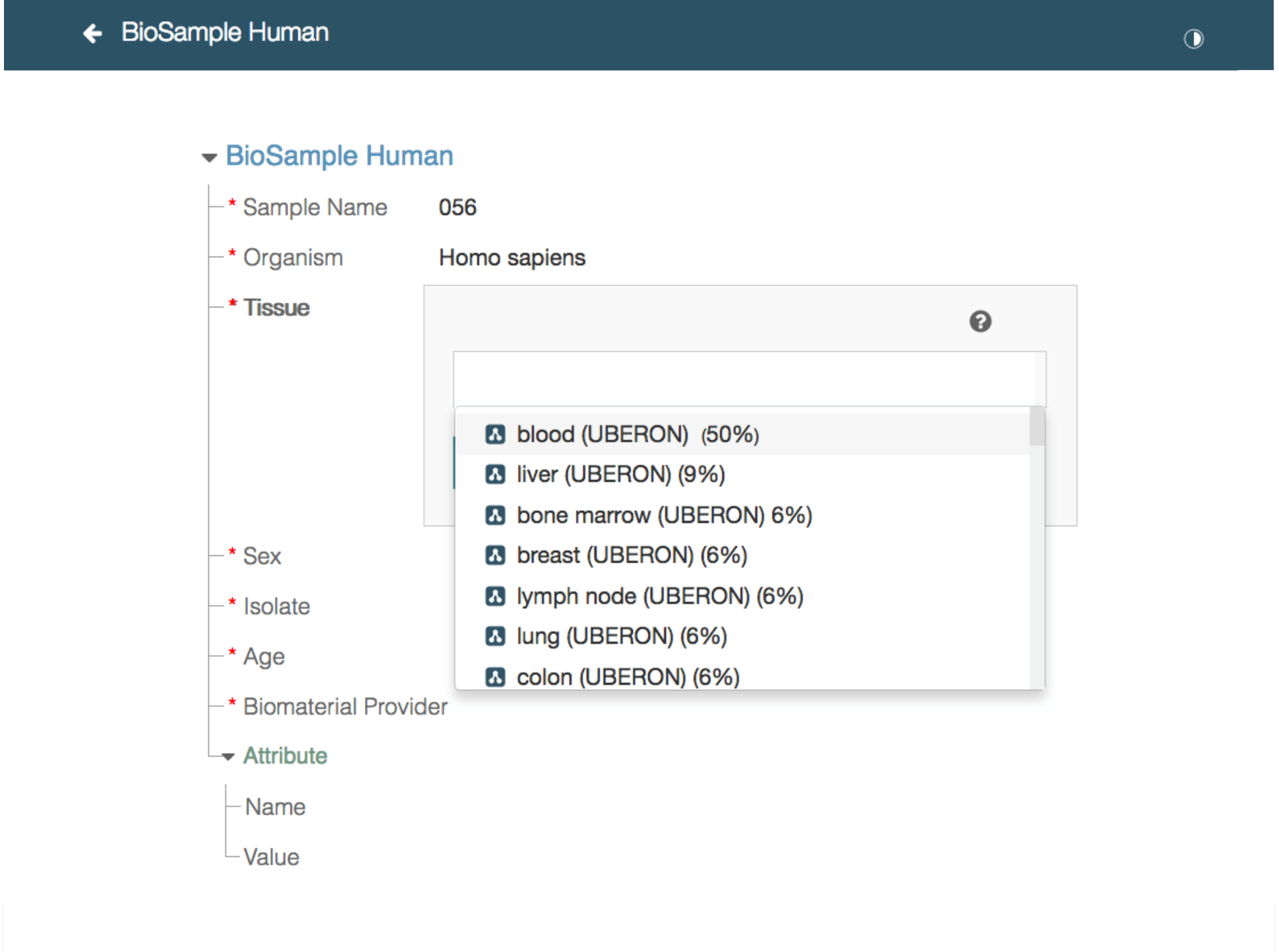
FILTER RESET

TYPE

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


CEDAR
Search
All / Users / Mark A. Musen
Workspace
Shared with Me
FILTER
RESET
TYPE
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 11:28 AM
Optional Attribute
9/5/17 10:38 AM
ImmPort Investigation
9/5/17 10:21 AM
LINCS Cell Line
9/5/17 9:49 AM
LINCS Antibody
9/5/17 9:49 AM
ImmPort Study
9/5/17 9:49 AM
9/5/17 9:49 AM
Open
Populate
Share...
Copy to...
Move to...
Rename...
Delete


BioSample Human
BioSample Human
* Sample Name
* Organism
* Tissue
* Sex
* Isolate
* Age
* Biomaterial Provider
Attribute
Name
Value
CANCEL
VALIDATE
SAVE


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

Projects that are adopting CEDAR

- COVID research in the Netherlands
- COVID research in the US (RADx)
- Neurobiology research in the UK (VFB)
- Tissue-mapping research in the US (HuBMAP)
- Cell-signaling research in the US (LINCS)
- Genomics research in the US (IDG)

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 uniform structure
- **Technology** to make it easy to author good metadata in the first place
- **Procedures** to create community-based standards in the first place

Metadata for Machines Workshops

- Are intensive 1–3 day invited, highly participatory sessions
- Historically, have been hosted by GO FAIR Organization
- Lead groups of scientists to consensus regarding essential metadata fields
  - for different areas of science
  - for different kinds of experiments
- Ultimately result in new CEDAR metadata templates

Online data will never be FAIR

- Until we standardize metadata structure using common **templates**
- 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