CEDAR: Enhancing Open Science Through Standard Metadata

Mark A. Musen
Stanford University
musen@stanford.edu

BMIR
Stanford Center for
Biomedical Informatics Research

SCIENTIFIC DATA

Amended: Addendum

OPEN

SUBJECT CATEGORIES
» Research data
» Publication characteristics

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson *et al.*[#]

There is an urgent need to improve the infrastructure supporting the reuse of scholarly data. A diverse set of stakeholders—representing academia, industry, funding agencies, and scholarly publishers—have come together to design and jointly endorse a concise and measureable set of principles that we refer to as the FAIR Data Principles. The intent is that these may act as a guideline for those wishing to enhance the reusability of their data holdings. Distinct from peer initiatives that focus on the human scholar, the FAIR Principles put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals. This Comment is the first formal publication of the FAIR Principles, and includes the rationale behind them, and some exemplar implementations in the community.

Received: 10 December 2015
Accepted: 12 February 2016
Published: 15 March 2016

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

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

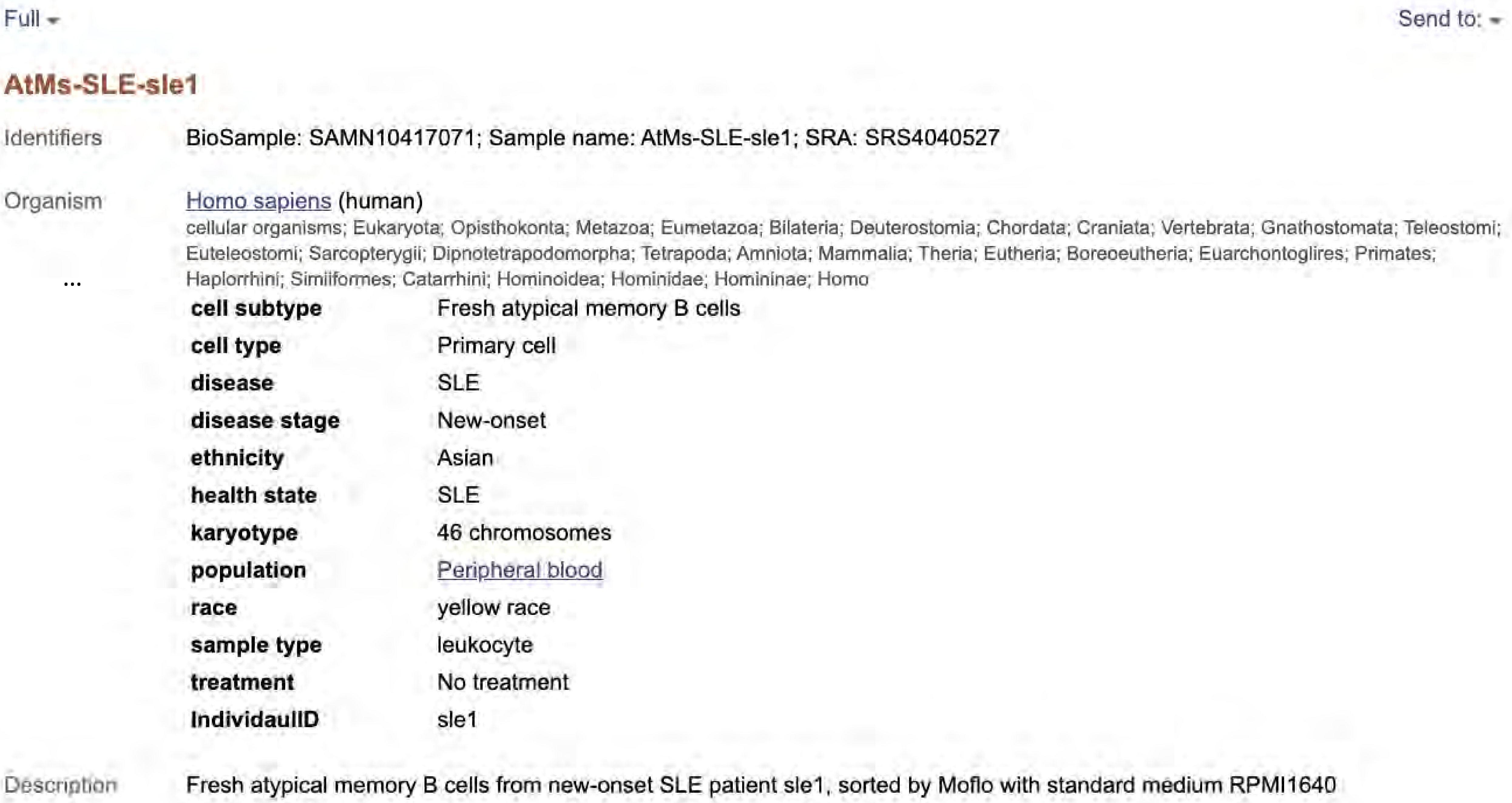
R1.1: (Meta)data are released with a clear and accessible data usage license

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 or delivering products, not leaving a legacy of reusable data
- Sponsors or managers may require data sharing, but they may not encourage the use of their own funds to pay for it
- Creating good metadata to describe data sets is unbearably hard



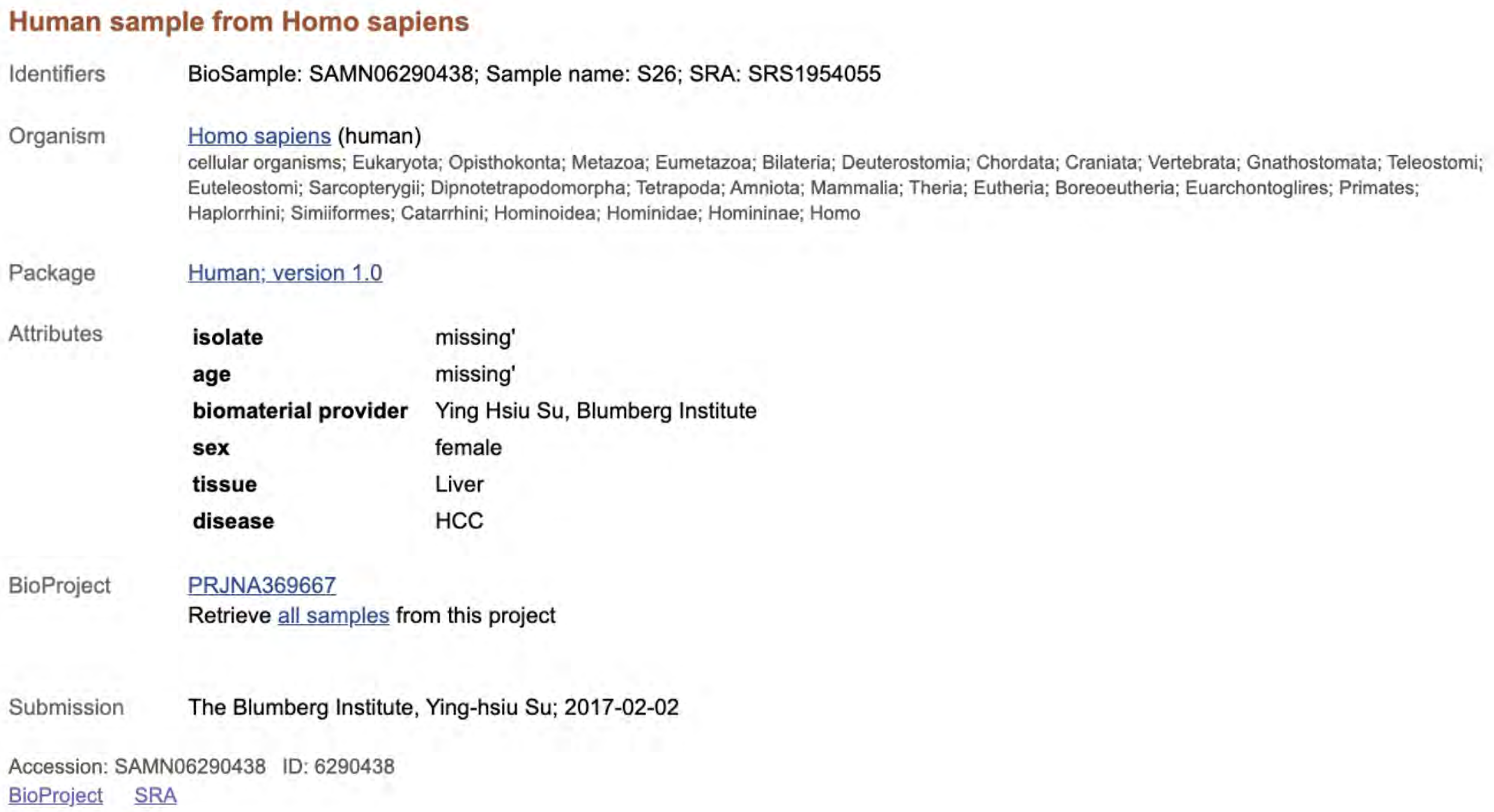
AtMs-SLE-sle1

Identifiers	BioSample: SAMN10417071; Sample name: AtMs-SLE-sle1; SRA: SRS4040527
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

...

cell subtype	Fresh atypical memory B cells
cell type	Primary cell
disease	SLE
disease stage	New-onset
ethnicity	Asian
health state	SLE
karyotype	46 chromosomes
population	Peripheral blood
race	yellow race
sample type	leukocyte
treatment	No treatment
IndividaulID	sle1

Description	Fresh atypical memory B cells from new-onset SLE patient sle1, sorted by Moflo with standard medium RPMI1640



Human sample from Homo sapiens

Identifiers	BioSample: SAMN06290438; Sample name: S26; SRA: SRS1954055
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

Attributes

isolate	missing'
age	missing'
biomaterial provider	Ying Hsiu Su, Blumberg Institute
sex	female
tissue	Liver
disease	HCC

BioProject	PRJNA369667 Retrieve all samples from this project
Submission	The Blumberg Institute, Ying-hsiu Su; 2017-02-02

Accession: SAMN06290438   ID: 6290438

BioProject   SRA

Sample from Homo sapiens

Identifiers	BioSample: SAMEA7571649; SRA: ERS5328271
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

Attributes

sample name	MIBC-Pat_4
collected by	Jena University Hospital
collection date	2010-01
sample type	MIBC
sex	w
ENA first public	2021-01-06
ENA last update	2020-11-13
ENA-CHECKLIST	ERC000011
External Id	SAMEA7571649
INSDC center alias	Jena University Hospital
INSDC center name	Jena University Hospital
INSDC first public	2021-01-06T17:11:48Z
INSDC last update	2020-11-13T09:13:33Z
INSDC status	public
SRA accession	ERS5328271
Submitter Id	MIBC-Pat_4

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

NCBI *BioSample* Metadata are Dreadful!

- 73% of “Boolean” metadata values are not actually *true* or *false*
  - *nonsmoker*, *former-smoker*
- 26% of “integer” metadata values cannot be parsed into integers
  - *JM52*, *UVPgt59.4*, *pig*
- 68% of metadata entries that are supposed to represent terms from biomedical ontologies do not actually do so.
  - *presumed normal*, *wild_type*

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

- **Ontologies** to provide controlled terms
- **Reporting guidelines** 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

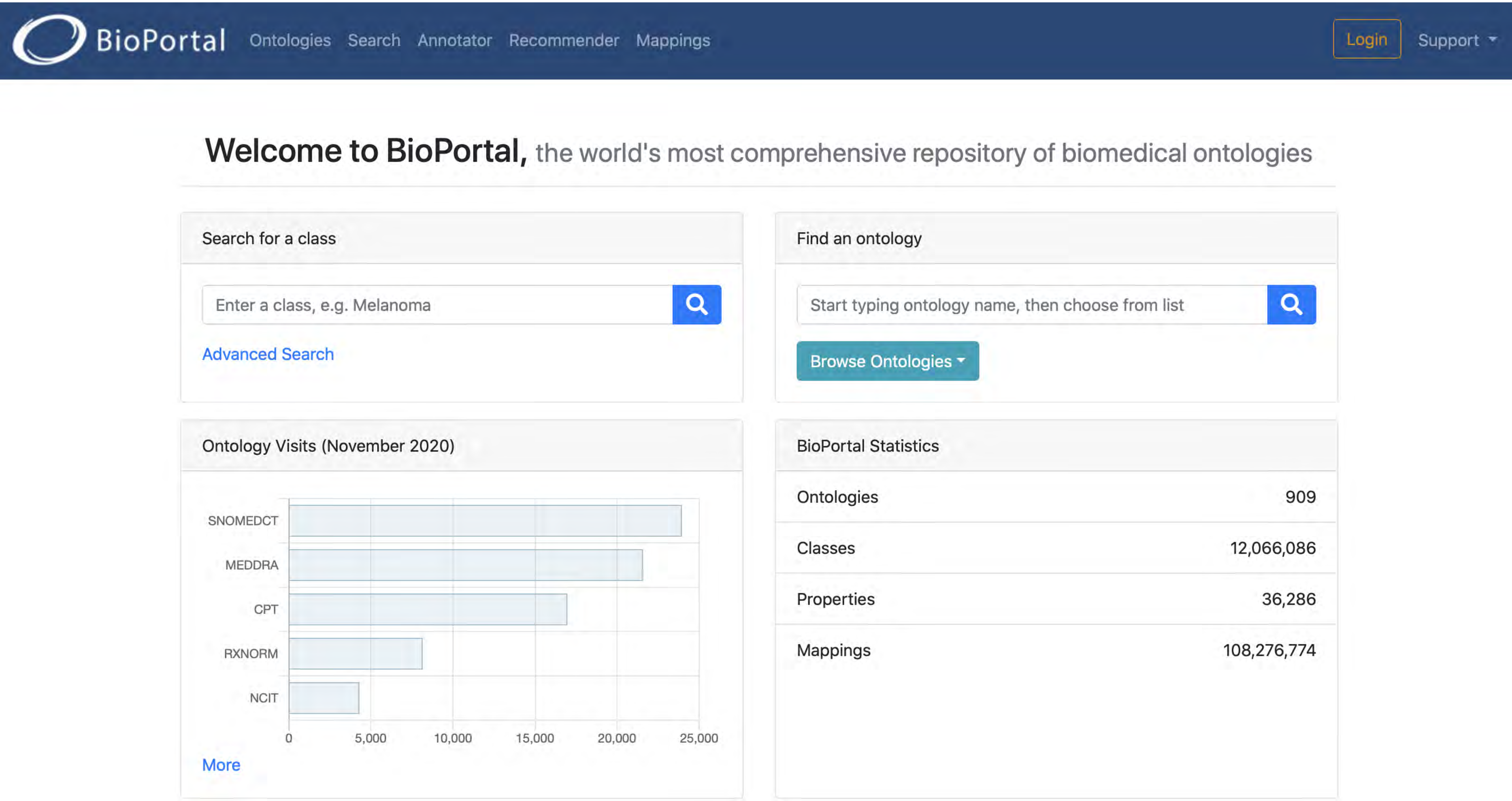
If we want to have FAIR data, we need good metadata. Good metadata need:

- **Ontologies** to provide controlled terms
- **Reporting guidelines** 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

Good metadata need ontologies!

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



Welcome to BioPortal, the world's most comprehensive repository of biomedical ontologies

Search for a class

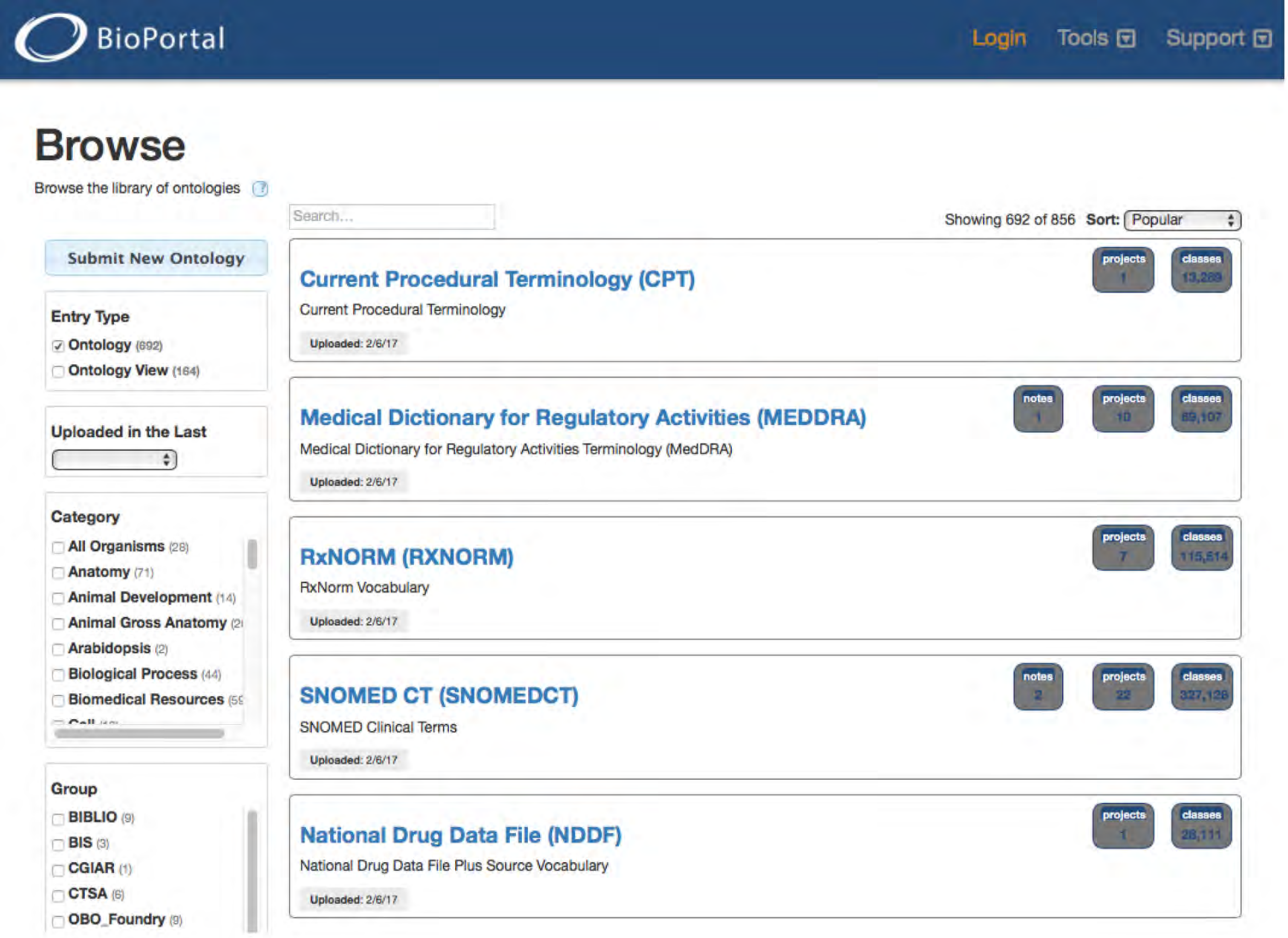
Find an ontology

Ontology Visits (November 2020)

BioPortal Statistics

Ontologies	909
Classes	12,066,086
Properties	36,286
Mappings	108,276,774

https://bioportal.bioontology.org



Browse

Browse the library of ontologies

Search...

Showing 692 of 856 **Sort:** Popular

Submit New Ontology

Entry Type

- ☑ **Ontology** (692)
- ☐ **Ontology View** (164)

Uploaded in the Last

Category

- ☐ **All Organisms** (28)
- ☐ **Anatomy** (71)
- ☐ **Animal Development** (14)
- ☐ **Animal Gross Anatomy** (2
- ☐ **Arabidopsis** (2)
- ☐ **Biological Process** (44)
- ☐ **Biomedical Resources** (5
- ☐ **Cell**

Group

- ☐ **BIBLIO** (9)
- ☐ **BIS** (3)
- ☐ **CGIAR** (1)
- ☐ **CTSA** (6)
- ☐ **OBO_Foundry** (9)

Current Procedural Terminology (CPT)

Current Procedural Terminology

Uploaded: 2/6/17

projects 1 · classes 13,289

Medical Dictionary for Regulatory Activities (MEDDRA)

Medical Dictionary for Regulatory Activities Terminology (MedDRA)

Uploaded: 2/6/17

notes 1 · projects 10 · classes 69,107

RxNORM (RXNORM)

RxNorm Vocabulary

Uploaded: 2/6/17

projects 7 · classes 115,614

SNOMED CT (SNOMEDCT)

SNOMED Clinical Terms

Uploaded: 2/6/17

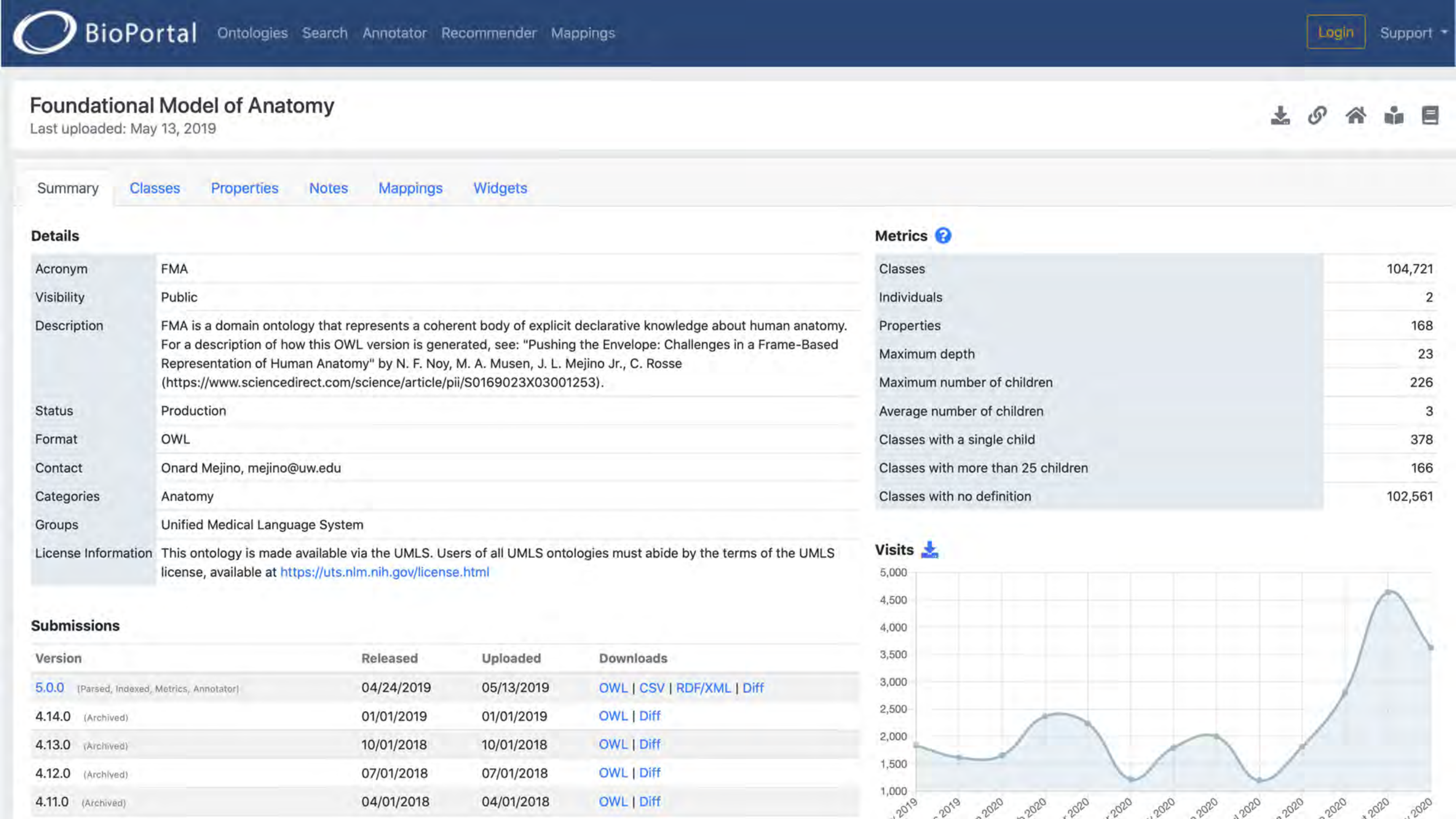
notes 2 · projects 22 · classes 327,128

National Drug Data File (NDDF)

National Drug Data File Plus Source Vocabulary

Uploaded: 2/6/17

projects 1 · classes 28,111



Foundational Model of Anatomy

Last uploaded: May 13, 2019

Details

Acronym	FMA
Visibility	Public
Description	FMA is a domain ontology that represents a coherent body of explicit declarative knowledge about human anatomy. For a description of how this OWL version is generated, see: "Pushing the Envelope: Challenges in a Frame-Based Representation of Human Anatomy" by N. F. Noy, M. A. Musen, J. L. Mejino Jr., C. Rosse (https://www.sciencedirect.com/science/article/pii/S0169023X03001253).
Status	Production
Format	OWL
Contact	Onard Mejino, mejino@uw.edu
Categories	Anatomy
Groups	Unified Medical Language System
License Information	This ontology is made available via the UMLS. Users of all UMLS ontologies must abide by the terms of the UMLS license, available at https://uts.nlm.nih.gov/license.html

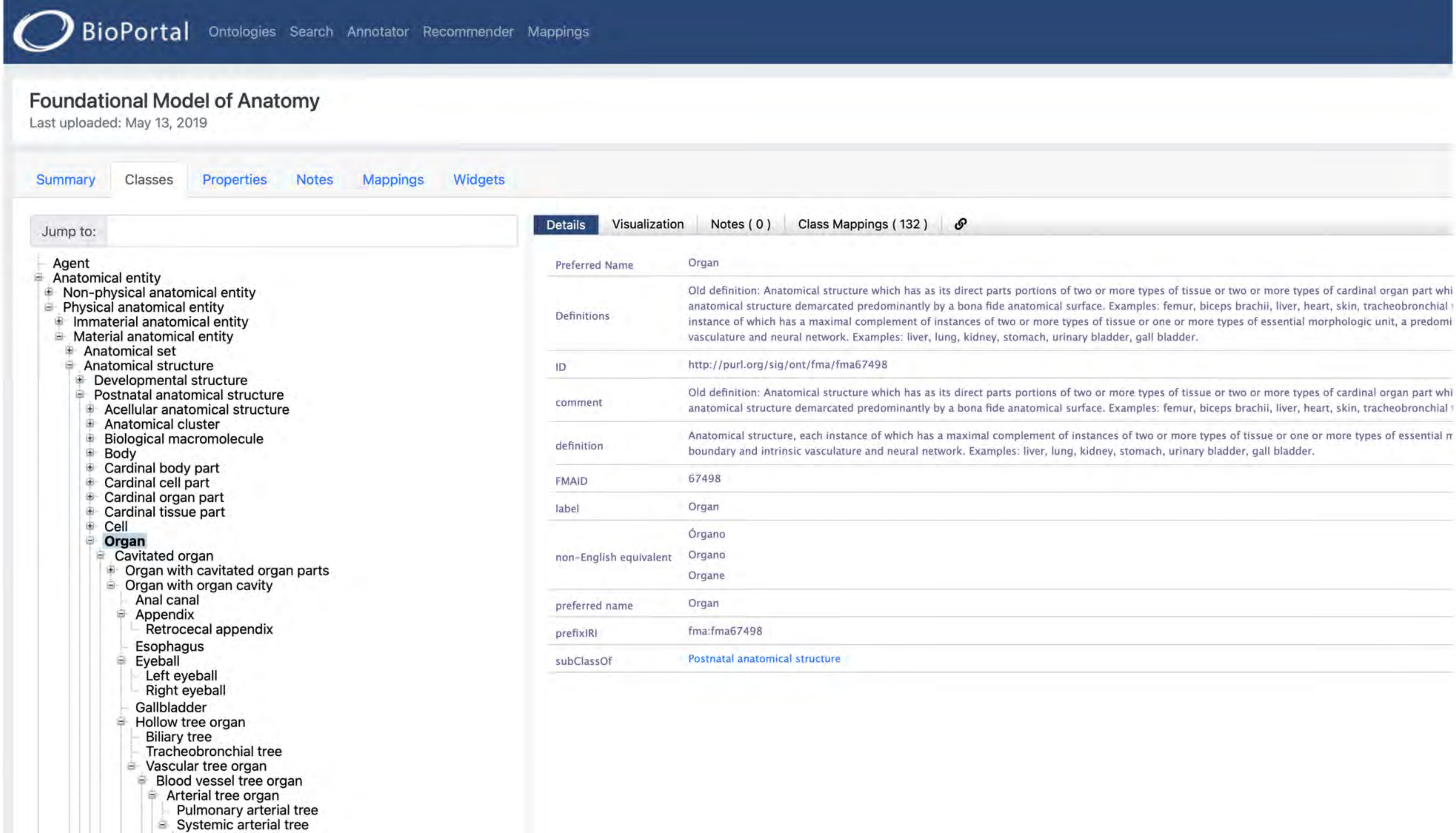
Metrics

Classes	104,721
Individuals	2
Properties	168
Maximum depth	23
Maximum number of children	226
Average number of children	3
Classes with a single child	378
Classes with more than 25 children	166
Classes with no definition	102,561

Submissions

Version	Released	Uploaded	Downloads
5.0.0 (Parsed, Indexed, Metrics, Annotator)	04/24/2019	05/13/2019	OWL | CSV | RDF/XML | Diff
4.14.0 (Archived)	01/01/2019	01/01/2019	OWL | Diff
4.13.0 (Archived)	10/01/2018	10/01/2018	OWL | Diff
4.12.0 (Archived)	07/01/2018	07/01/2018	OWL | Diff
4.11.0 (Archived)	04/01/2018	04/01/2018	OWL | Diff

Visits



Foundational Model of Anatomy

Last uploaded: May 13, 2019

- Agent
- Anatomical entity
  - Non-physical anatomical entity
  - Physical anatomical entity
    - Immaterial anatomical entity
    - Material anatomical entity
      - Anatomical set
      - Anatomical structure
        - Developmental structure
        - Postnatal anatomical structure
          - Acellular anatomical structure
          - Anatomical cluster
          - Biological macromolecule
          - Body
          - Cardinal body part
          - Cardinal cell part
          - Cardinal organ part
          - Cardinal tissue part
          - Cell
          - **Organ**
            - Cavitated organ
              - Organ with cavitated organ parts
              - Organ with organ cavity
                - Anal canal
                - Appendix
                  - Retrocecal appendix
                - Esophagus
                - Eyeball
                  - Left eyeball
                  - Right eyeball
                - Gallbladder
                - Hollow tree organ
                  - Biliary tree
                  - Tracheobronchial tree
                  - Vascular tree organ
                    - Blood vessel tree organ
                      - Arterial tree organ
                        - Pulmonary arterial tree
                        - Systemic arterial tree

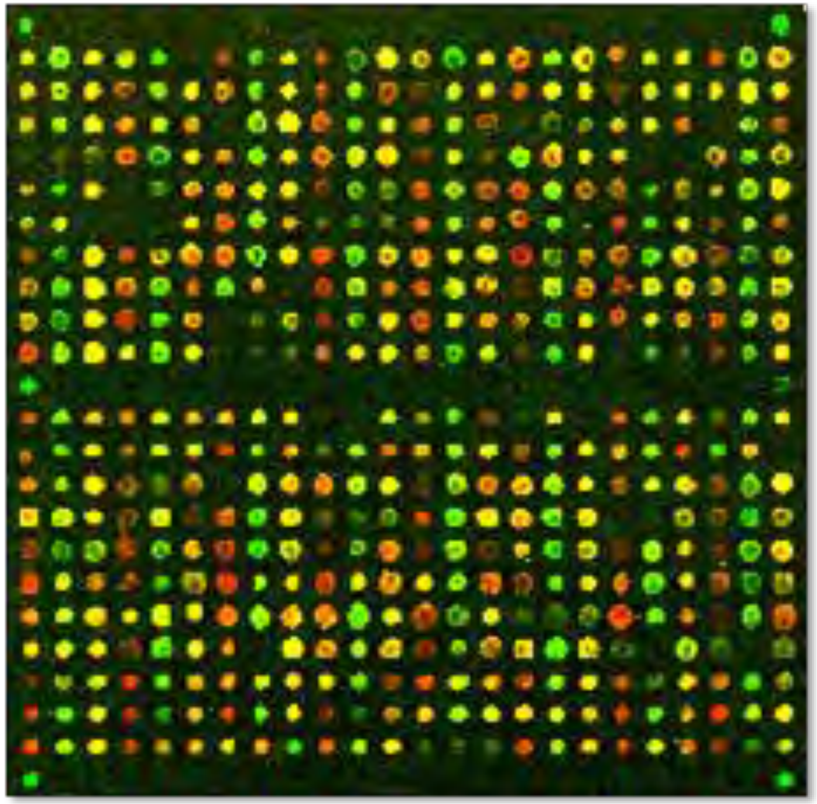
Preferred Name	Organ
Definitions	Old definition: Anatomical structure which has as its direct parts portions of two or more types of tissue or two or more types of cardinal organ part whi anatomical structure demarcated predominantly by a bona fide anatomical surface. Examples: femur, biceps brachii, liver, heart, skin, tracheobronchial instance of which has a maximal complement of instances of two or more types of tissue or one or more types of essential morphologic unit, a predomi vasculature and neural network. Examples: liver, lung, kidney, stomach, urinary bladder, gall bladder.
ID	http://purl.org/sig/ont/fma/fma67498
comment	Old definition: Anatomical structure which has as its direct parts portions of two or more types of tissue or two or more types of cardinal organ part whi anatomical structure demarcated predominantly by a bona fide anatomical surface. Examples: femur, biceps brachii, liver, heart, skin, tracheobronchial
definition	Anatomical structure, each instance of which has a maximal complement of instances of two or more types of tissue or one or more types of essential m boundary and intrinsic vasculature and neural network. Examples: liver, lung, kidney, stomach, urinary bladder, gall bladder.
FMAID	67498
label	Organ
non-English equivalent	Órgano Organo Organe
preferred name	Organ
prefixIRI	fma:fma67498
subClassOf	Postnatal anatomical structure

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

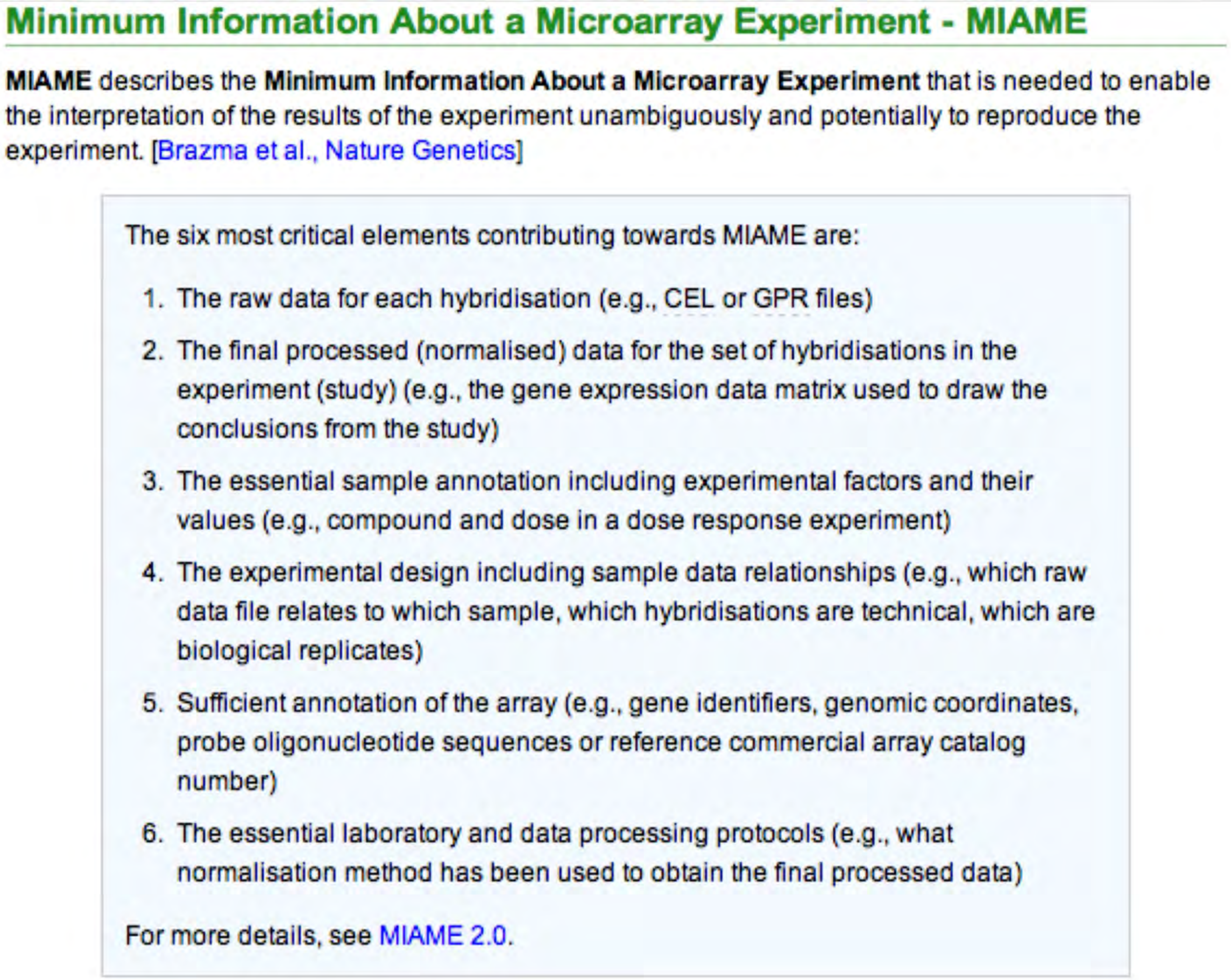
- **Ontologies** to provide controlled terms
- **Reporting guidelines** to provide a standardized structure for the metadata components
- **Technology** to make it easy to author good metadata in the first place
- **Procedures** to create community-based standards in the first place

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.

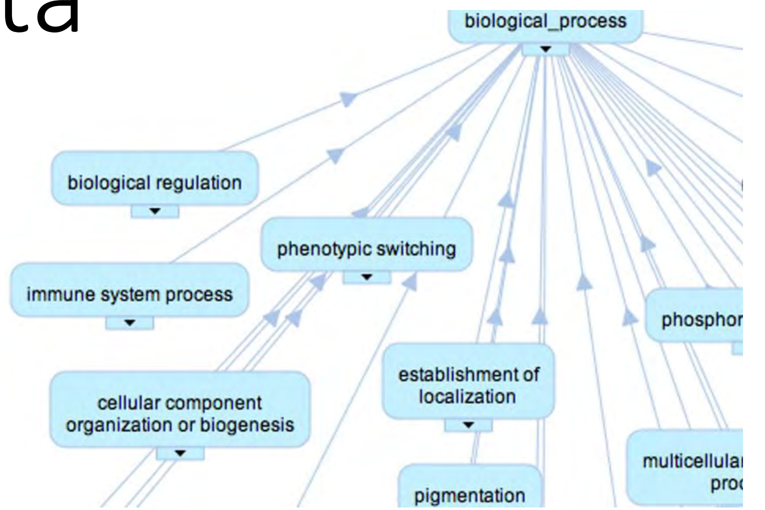
But it didn’t stop with MIAME!

- 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)

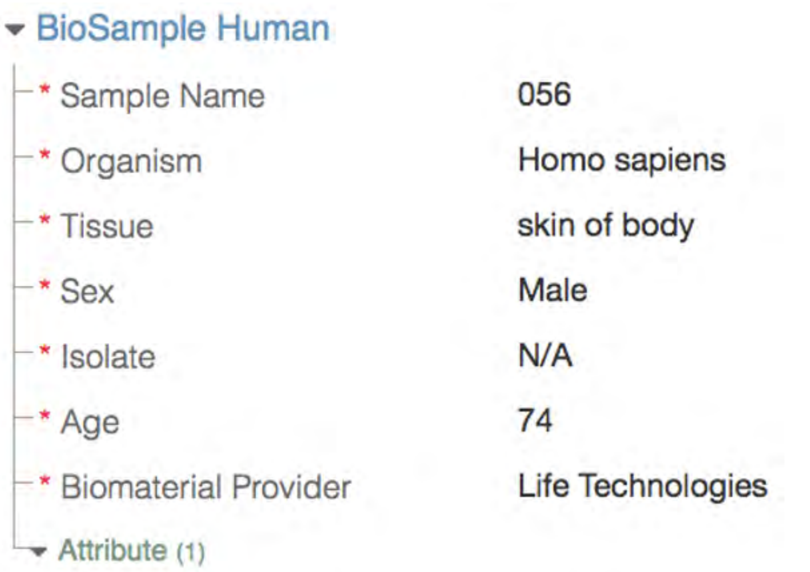
These are exactly the kinds of community standards
that we need to structure metadata!

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)

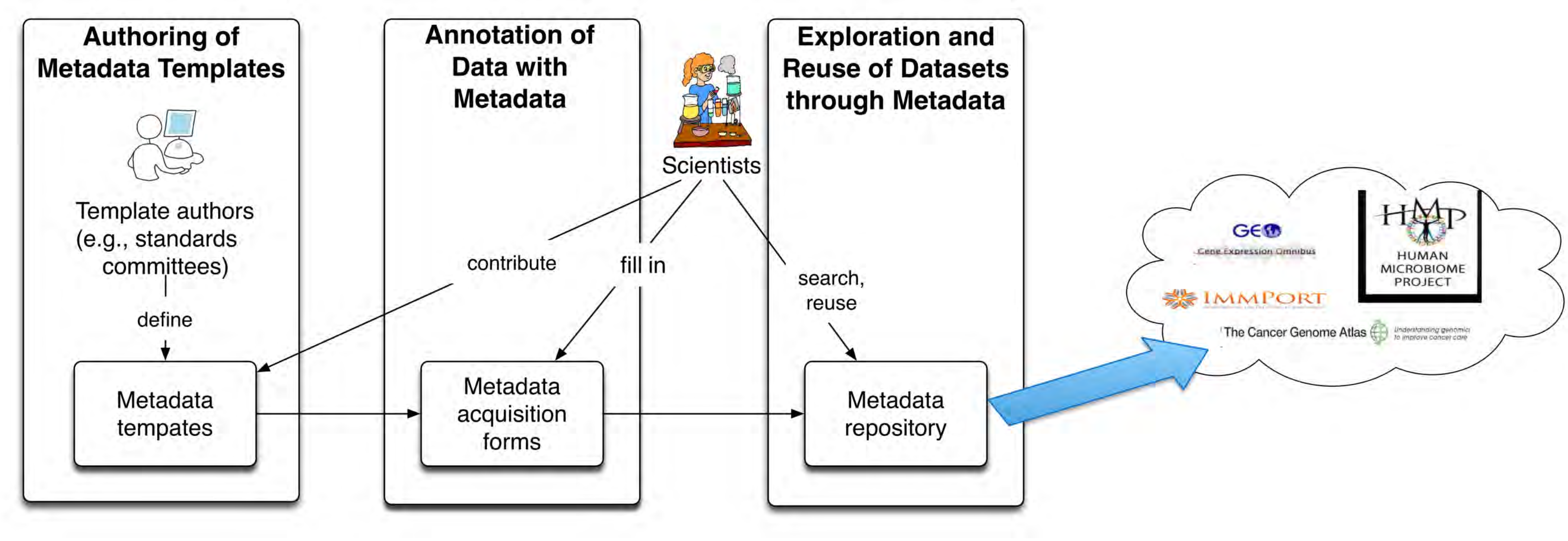
If we want to have FAIR data, we need good metadata. Good metadata need:

- **Ontologies** to provide controlled terms
- **Reporting guidelines** 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

Technology for better metadata

1. **CEDAR Workbench:** An editor that helps researchers to create standards-compliant metadata from
    - Ontologies
    - Reporting guidelines
2. **CEDAR Metadata Validator:** A system that validates spreadsheet-entered metadata against CEDAR templates

The **CEDAR** Workbench



Authoring of Metadata Templates

Template authors (e.g., standards committees)

define

Metadata tempates

Annotation of Data with Metadata

Metadata acquisition forms

Scientists

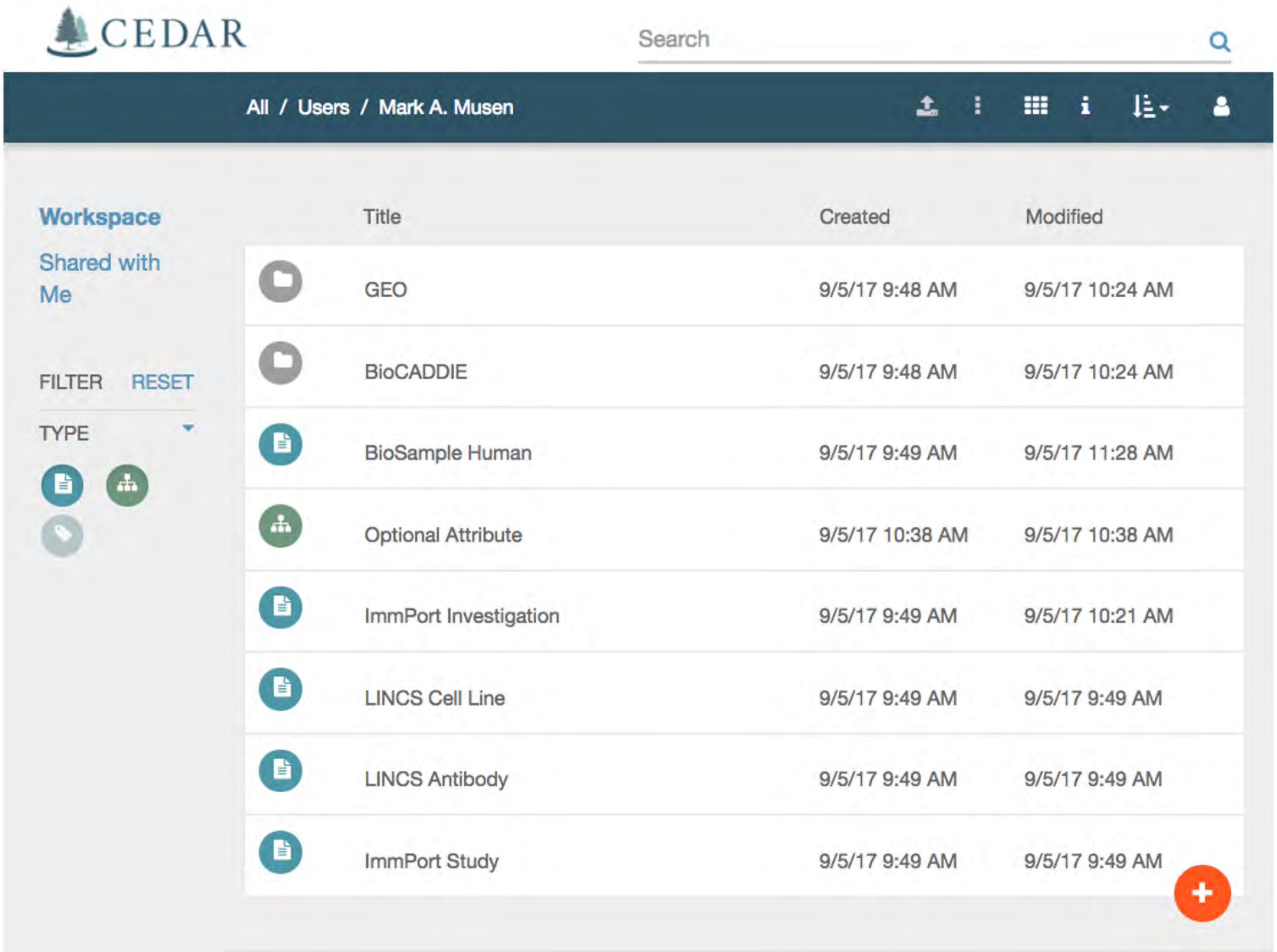
contribute

fill in

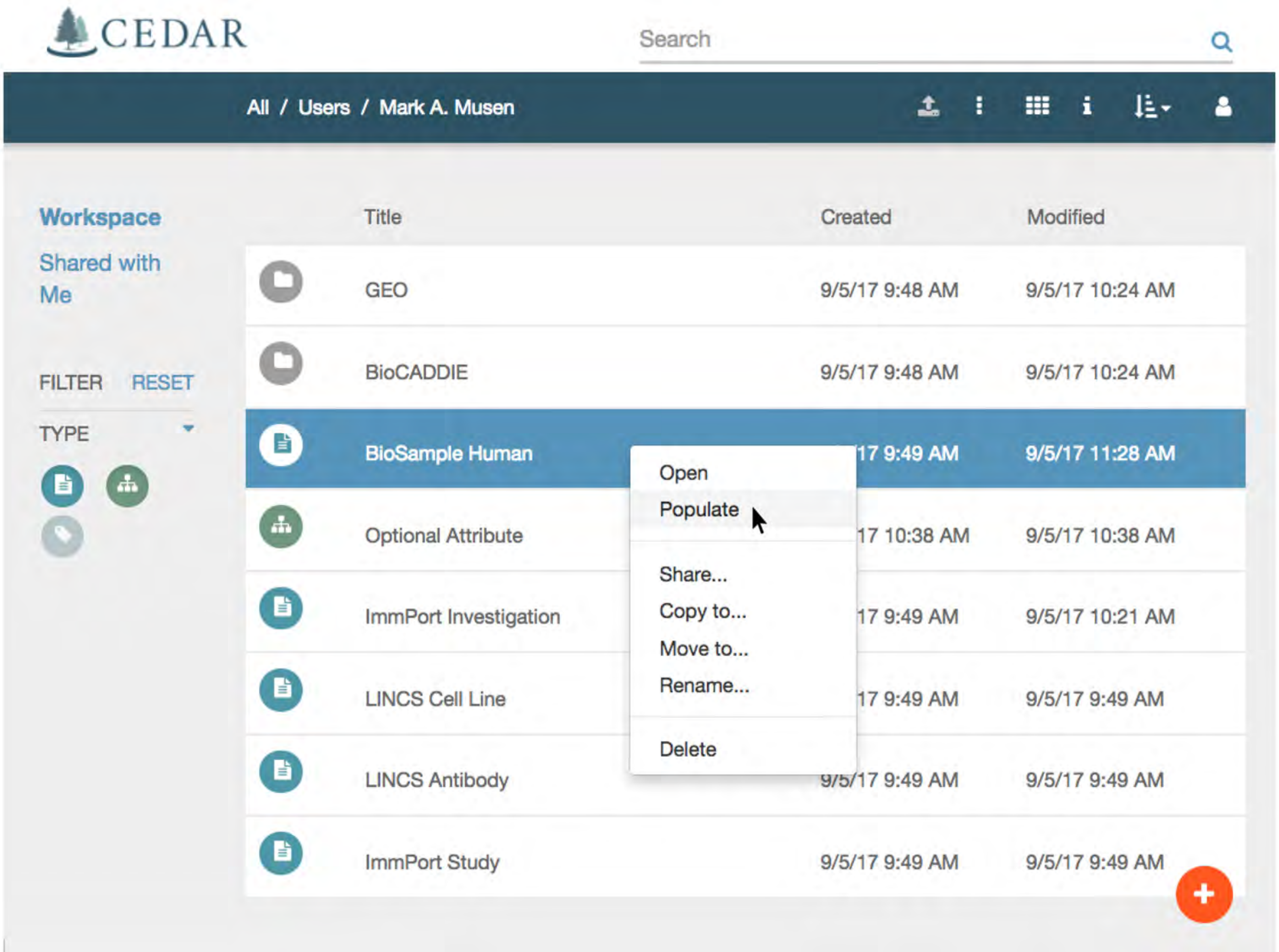
search, reuse

Exploration and Reuse of Datasets through Metadata

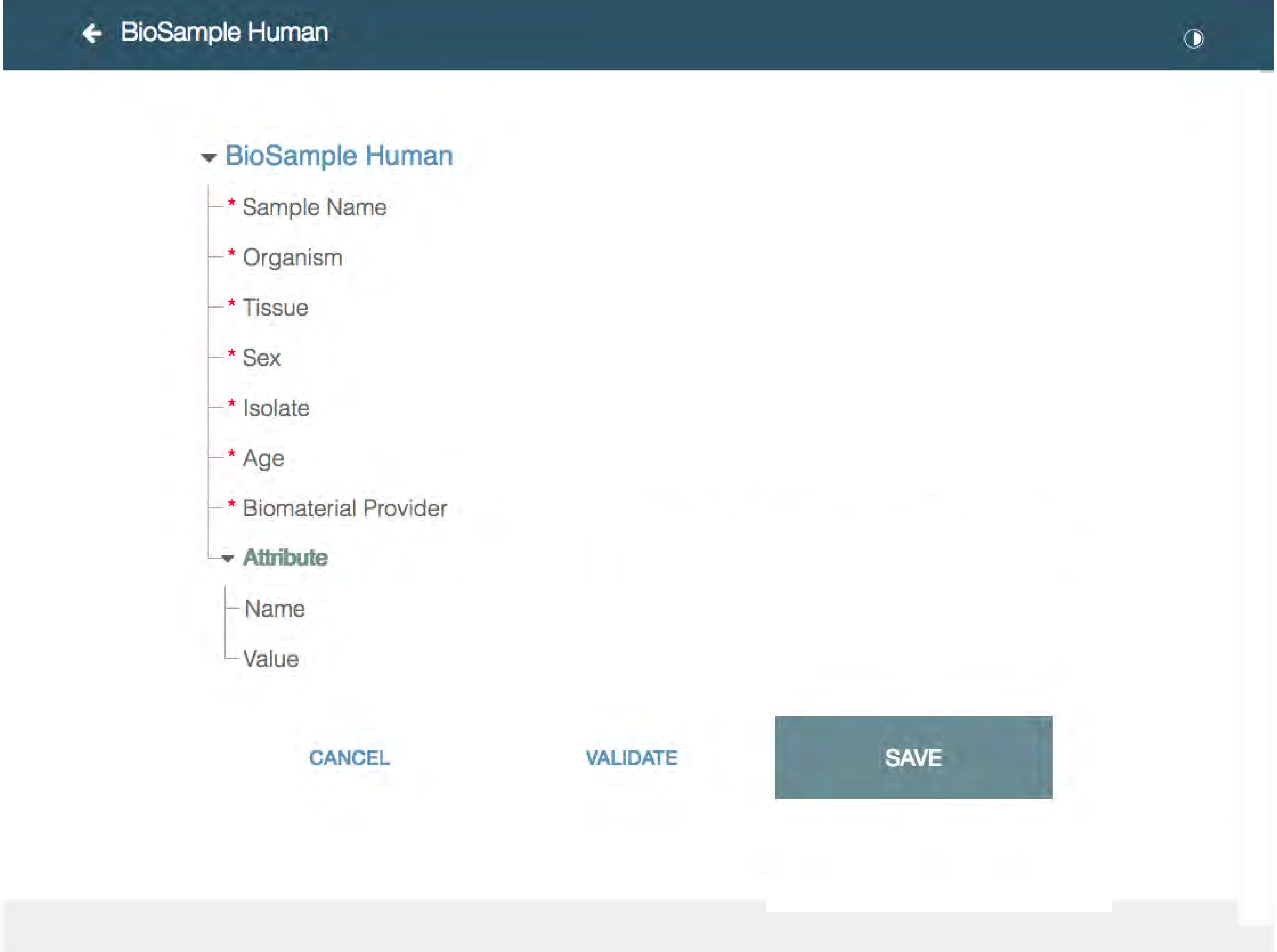
Metadata repository



	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
品	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



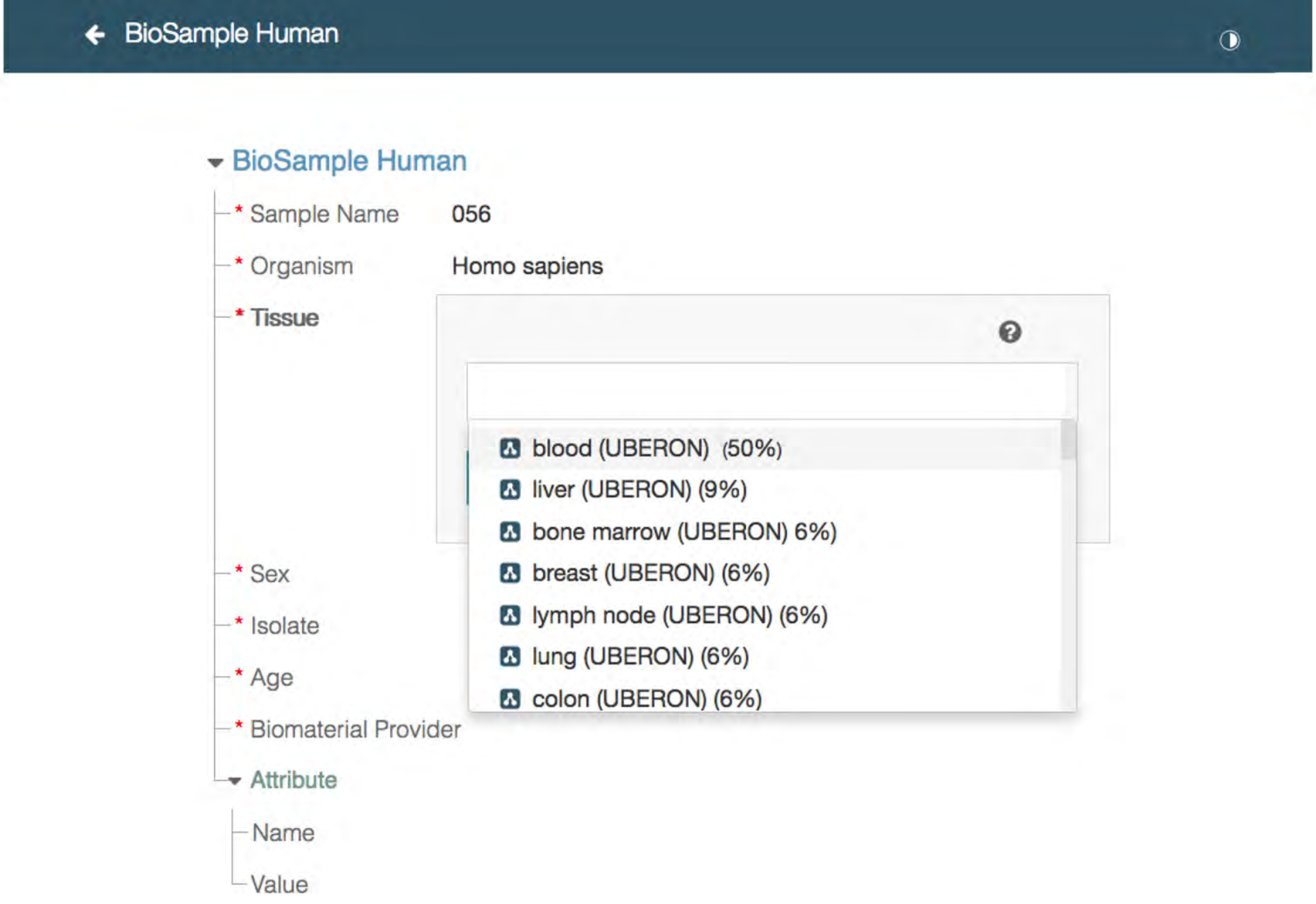
Title	Created	Modified
GEO	9/5/17 9:48 AM	9/5/17 10:24 AM
BioCADDIE	9/5/17 9:48 AM	9/5/17 10:24 AM
BioSample Human	17 9:49 AM	9/5/17 11:28 AM
Optional Attribute	17 10:38 AM	9/5/17 10:38 AM
ImmPort Investigation	17 9:49 AM	9/5/17 10:21 AM
LINCS Cell Line	17 9:49 AM	9/5/17 9:49 AM
LINCS Antibody	9/5/17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

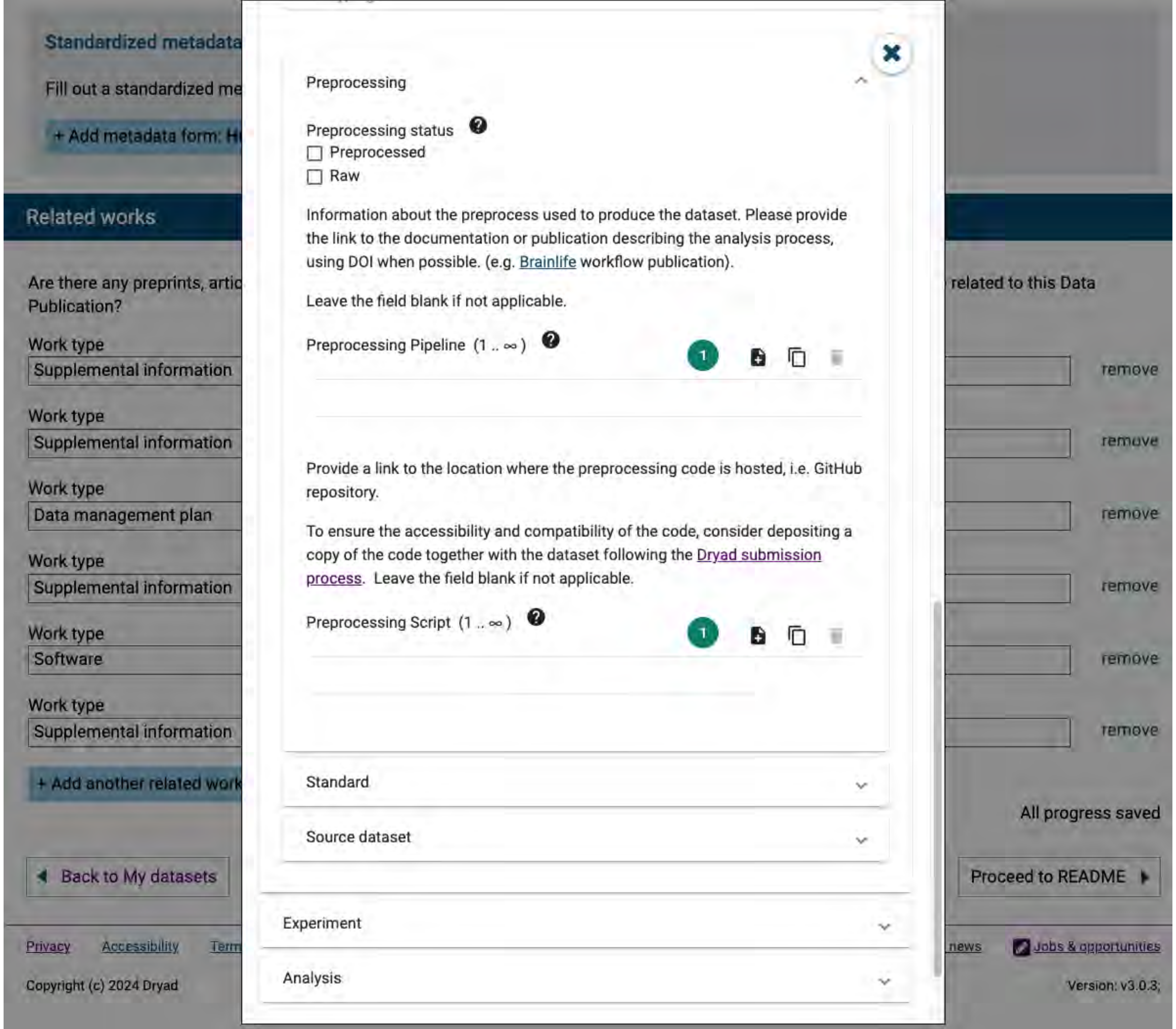


BioSample Human

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

CEDAR Metadata Editor in the **Dryad** Platform



Preprocessing

Preprocessing status ?

☐ Preprocessed
☐ Raw

Information about the preprocess used to produce the dataset. Please provide the link to the documentation or publication describing the analysis process, using DOI when possible. (e.g. Brainlife workflow publication).

Leave the field blank if not applicable.

Preprocessing Pipeline (1 .. ∞) ?

Provide a link to the location where the preprocessing code is hosted, i.e. GitHub repository.

To ensure the accessibility and compatibility of the code, consider depositing a copy of the code together with the dataset following the Dryad submission process. Leave the field blank if not applicable.

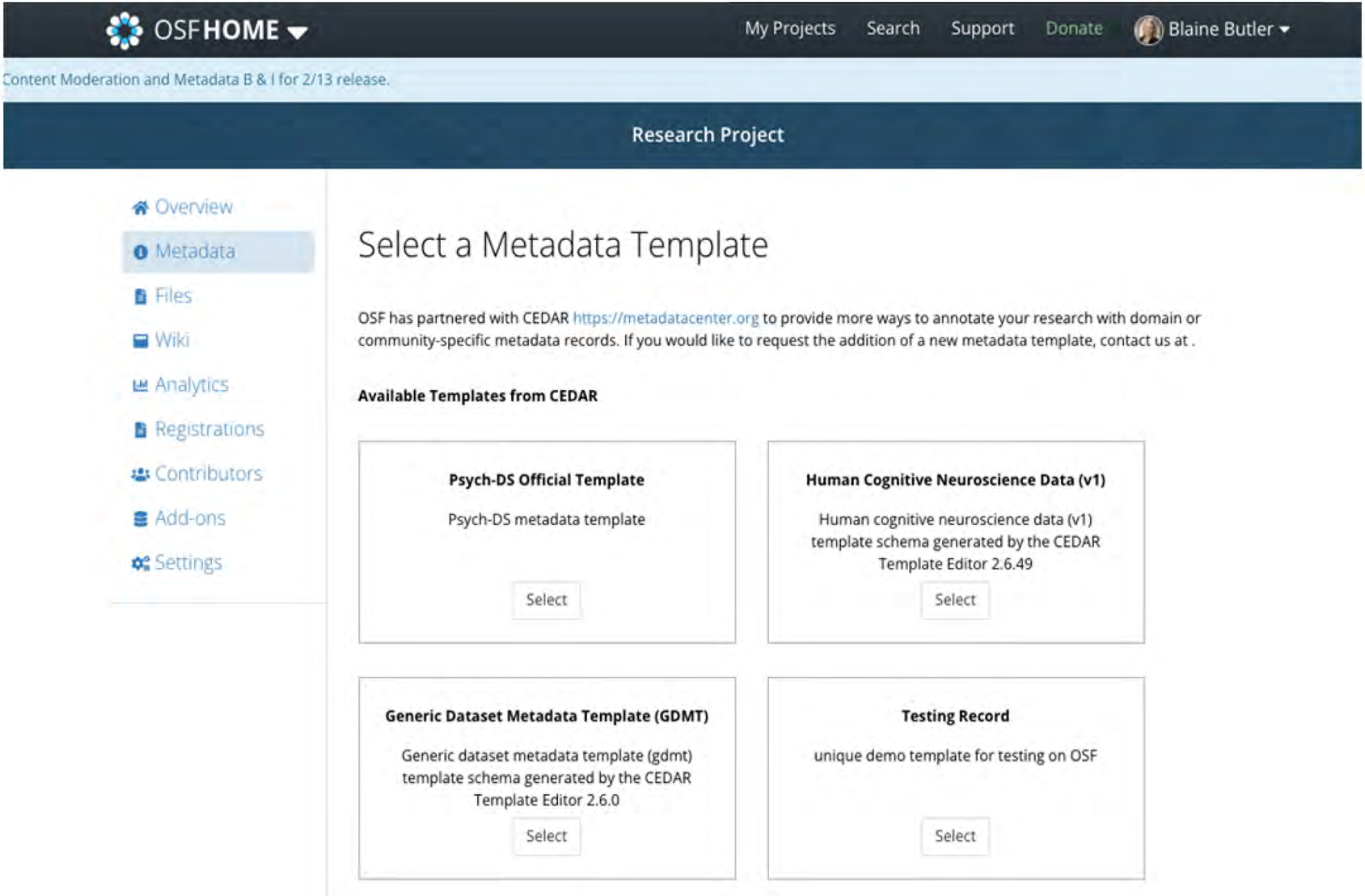
Preprocessing Script (1 .. ∞) ?

Standard

Source dataset

Experiment

Analysis



Content Moderation and Metadata B & I for 2/13 release.

Select a Metadata Template

OSF has partnered with CEDAR https://metadatacenter.org to provide more ways to annotate your research with domain or community-specific metadata records. If you would like to request the addition of a new metadata template, contact us at .

Available Templates from CEDAR

Psych-DS Official Template

Psych-DS metadata template

Human Cognitive Neuroscience Data (v1)

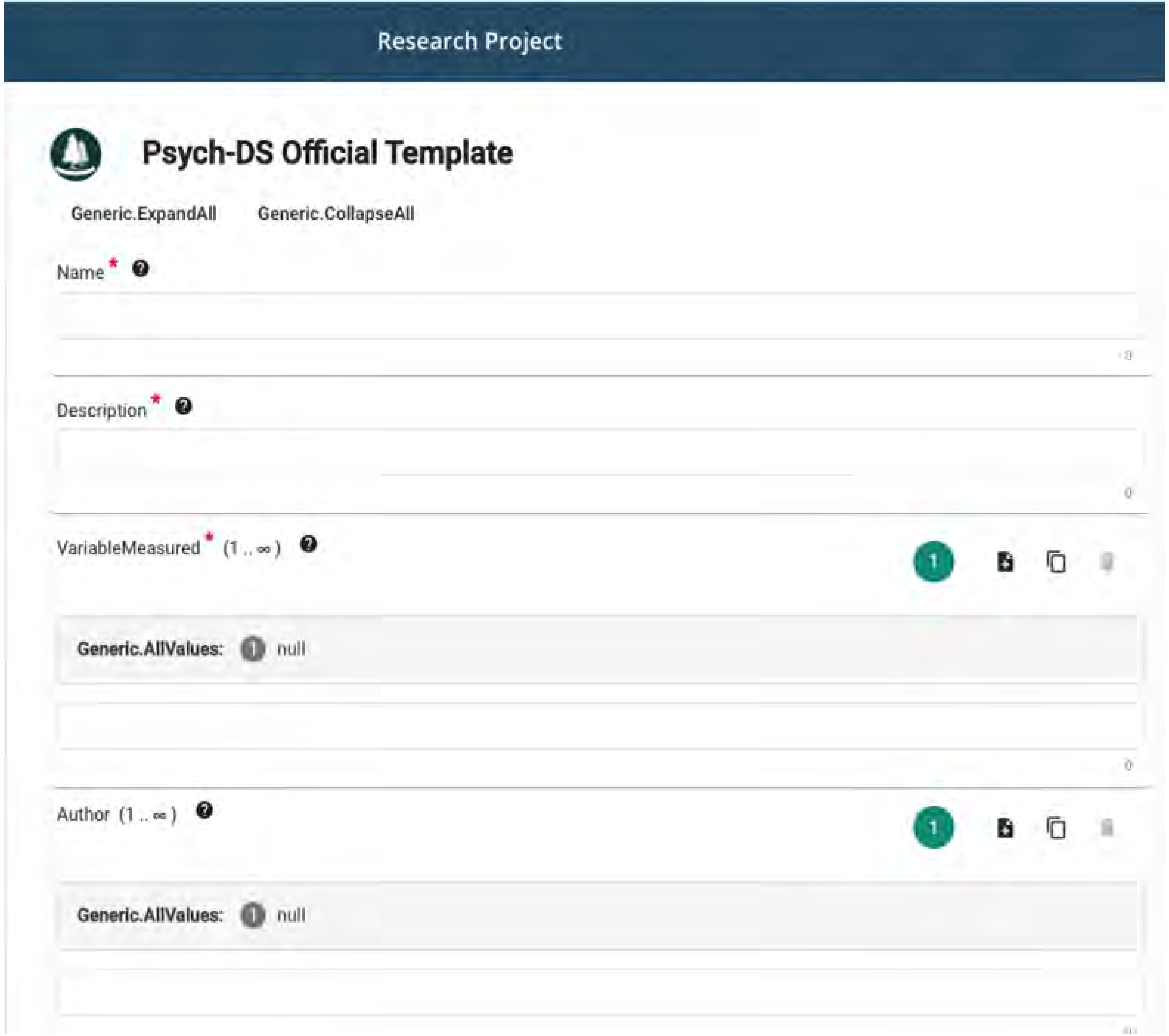
Human cognitive neuroscience data (v1) template schema generated by the CEDAR Template Editor 2.6.49

Generic Dataset Metadata Template (GDMT)

Generic dataset metadata template (gdmt) template schema generated by the CEDAR Template Editor 2.6.0

Testing Record

unique demo template for testing on OSF


Research Project

Psych-DS Official Template

Name *

Description *

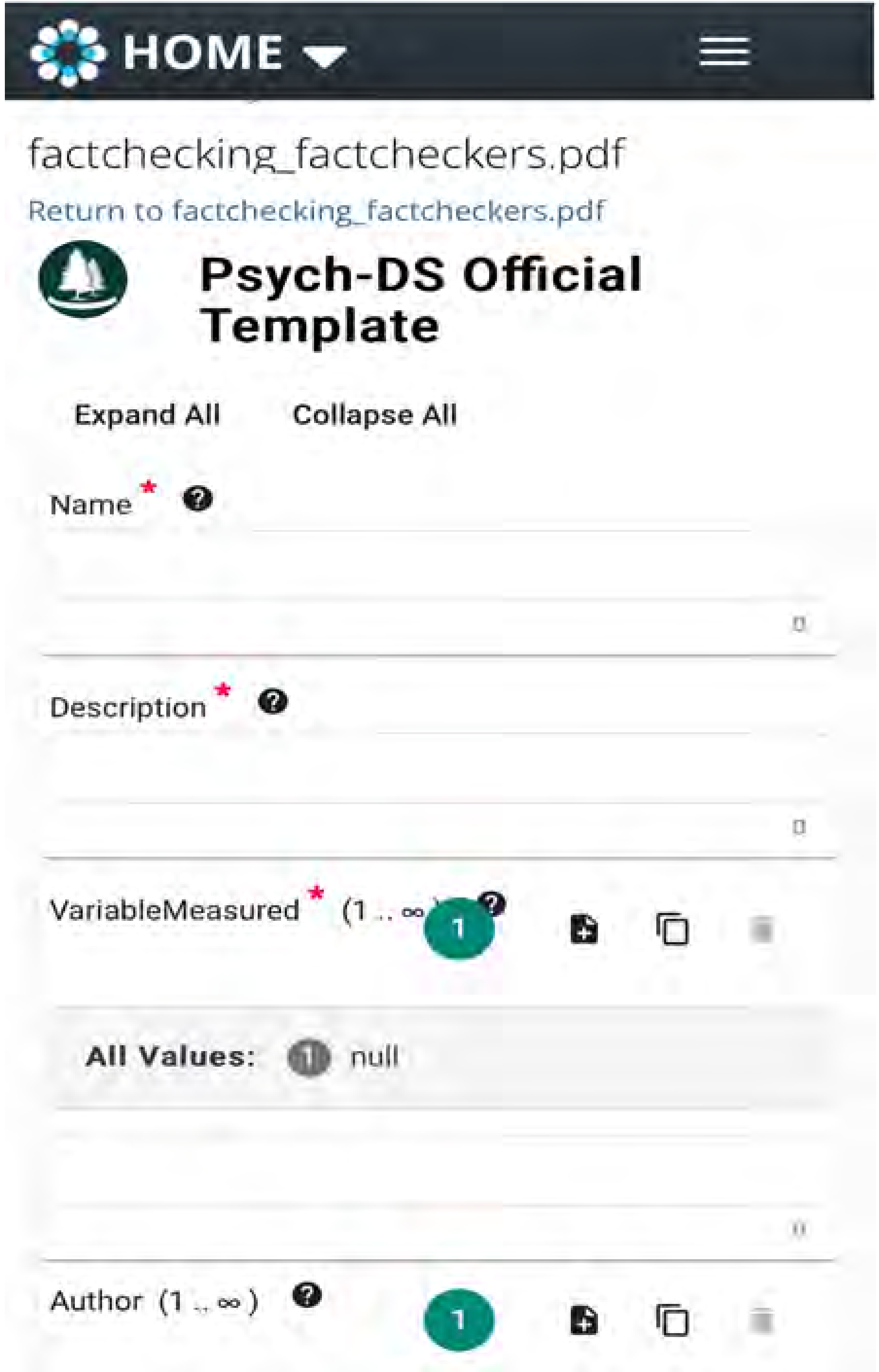
VariableMeasured * (1 .. ∞)

Generic.AllValues: 1 null

Author (1 .. ∞)

Generic.AllValues: 1 null

CEDAR Metadata Editor in the **Open Science Framework** Web Platform



factchecking_factcheckers.pdf

Return to factchecking_factcheckers.pdf

Psych-DS Official Template

Name *

Description *

VariableMeasured * (1 .. ∞)

All Values: 1 null

Author (1 .. ∞)

CEDAR Metadata Editor in the **Open Science Framework** App

Technology for better metadata

1. **CEDAR Workbench:** An editor that helps researchers to create standards-compliant metadata from
    - Ontologies
    - Reporting guidelines
2. **CEDAR Metadata Validator:** A system that validates spreadsheet-entered metadata against CEDAR templates

HuBMAP Donors Samples Datasets Other ▾ Atlas & Tools ▾ Resources ▾ User Profile ▾

Human BioMolecular Atlas Program

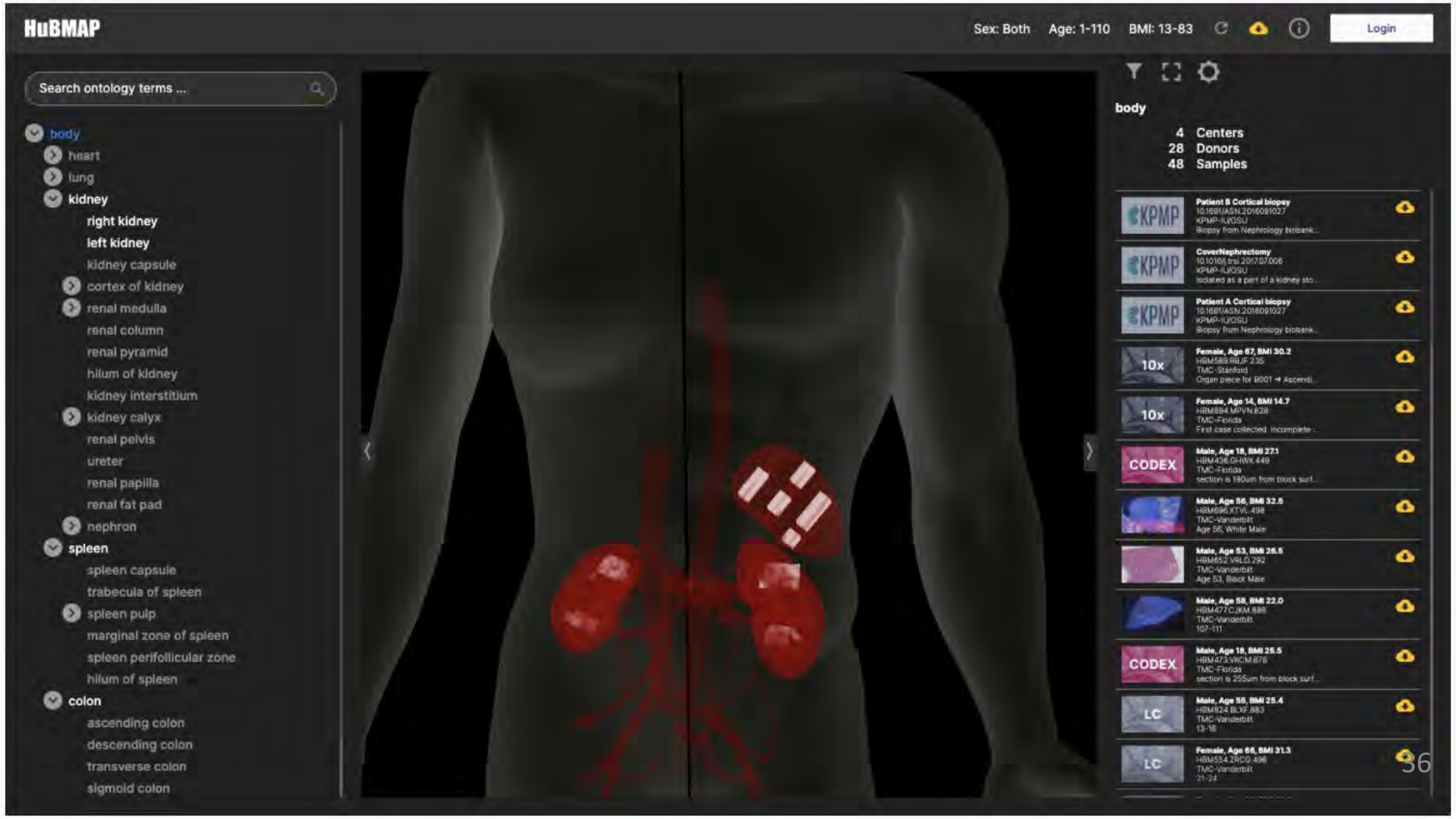
An open, global atlas of the human body at the cellular level

The HuBMAP Data Portal is the central resource for discovery, visualization, and download of single-cell tissue data generated by the consortium. A standardized data curation and processing workflow ensure that only high quality is released.

Navigate healthy human cells with the Common Coordinate Framework

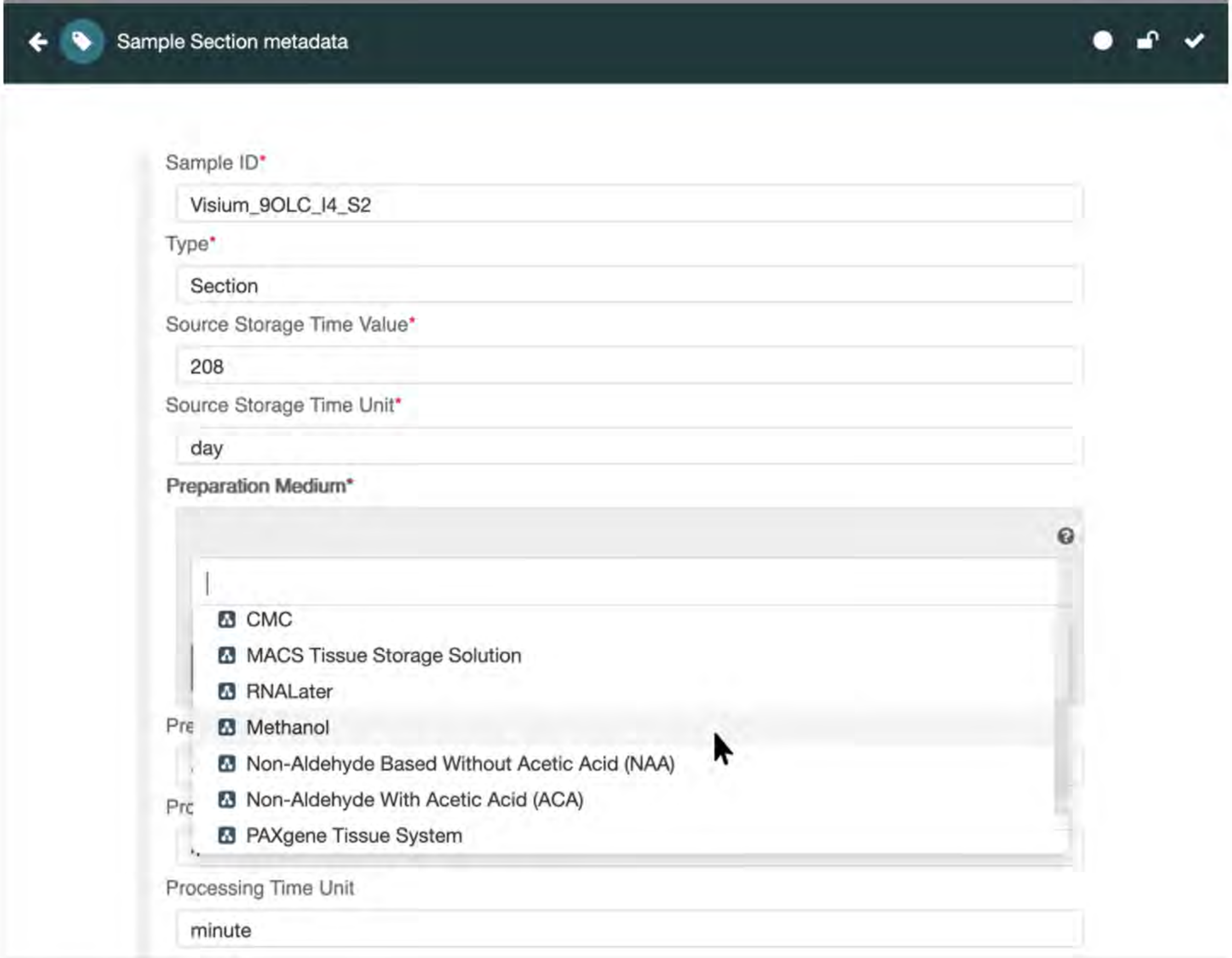
Interact with the human body data with the Anatomical Structures, Cell Types and Biomarkers (ASCT+B) Tables and CCF Ontology. Also explore two user interfaces: the Registration User Interface (RUI) for tissue data registration and Exploration User Interface (EUI) for semantic and spatial data.

Get Started



‹ ○ ● ○ ›

Screenshot



Sample Section metadata

Sample ID*: Visium_9OLC_I4_S2

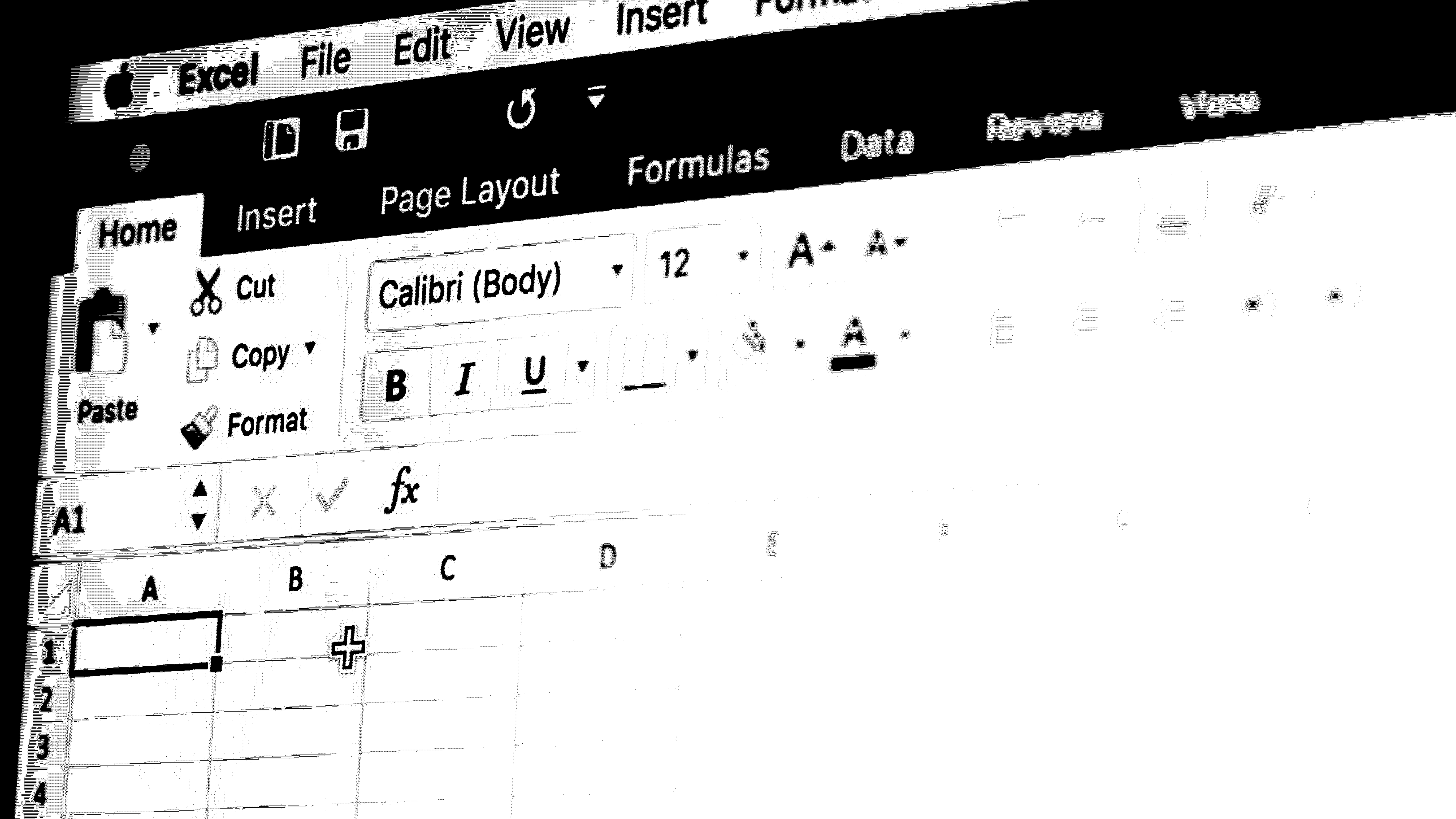
Type*: Section

Source Storage Time Value*: 208

Source Storage Time Unit*: day

Preparation Medium*

Processing Time Unit: minute

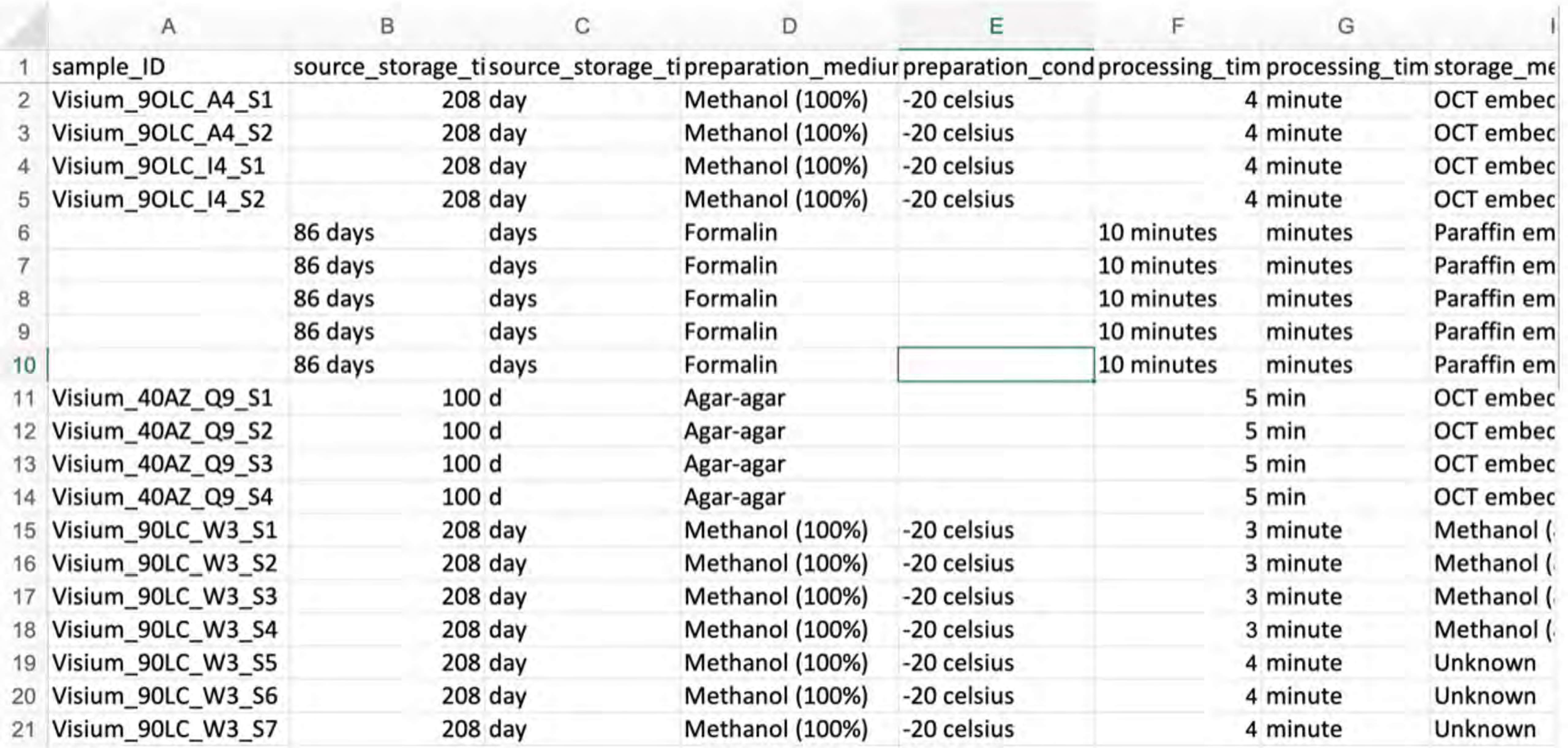

Excel
File
Edit
View
Insert
Home
Insert
Page Layout
Formulas
Data
Cut
Copy
Paste
Format
Calibri (Body)
12
B
I
U
A1
fx
A
B
C
D
1
2
3
4

Spreadsheets can’t enforce adherence to standards

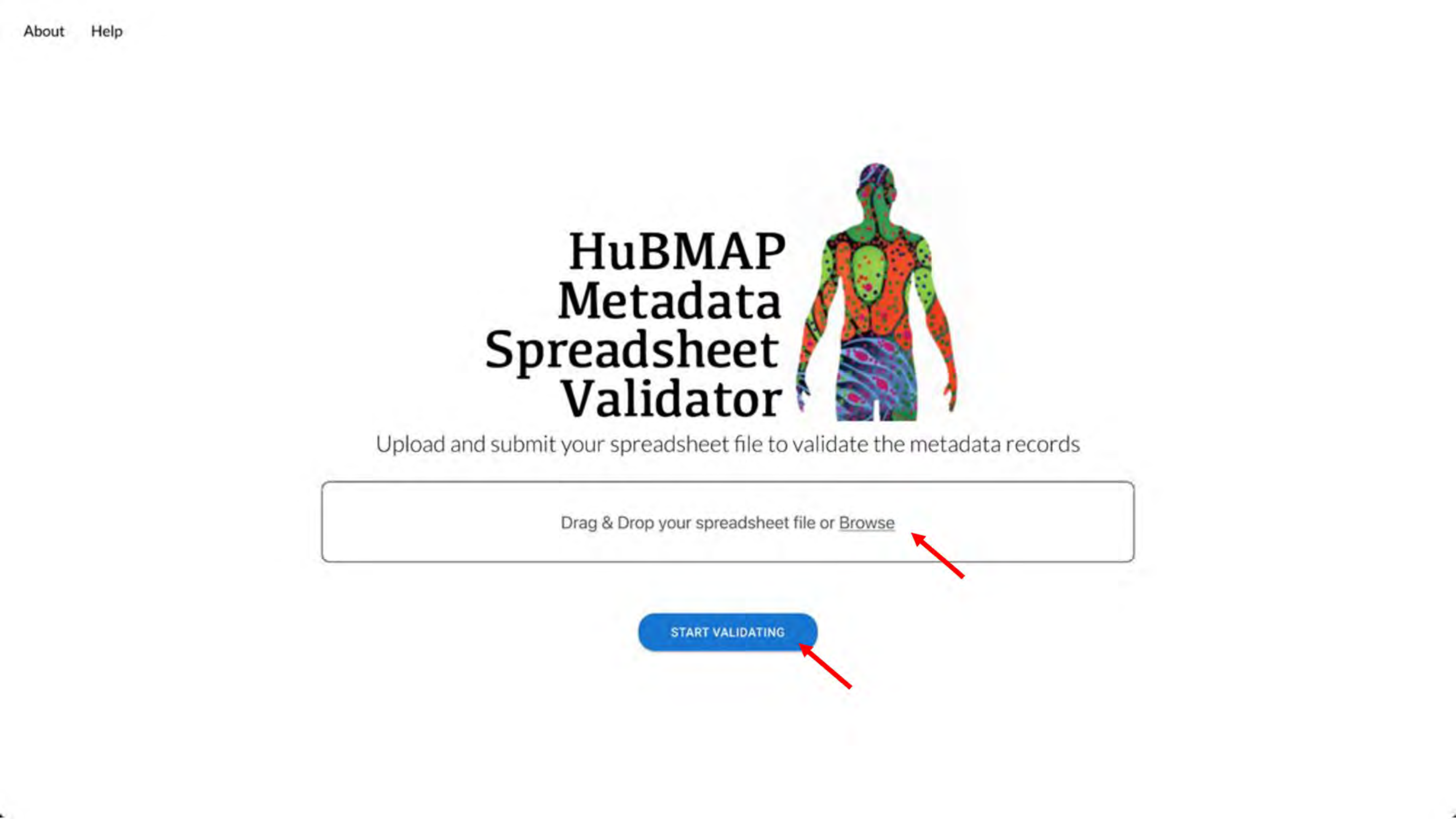
- Spreadsheets are prone to errors, such as missing fields and formatting problems
- Validation features in tools such as Excel are limited, allowing users to enter erroneous information
- Metadata ingestion processes based on spreadsheets need to anticipate and rectify such errors

Our Solution: A CEDAR-based approach that …

- Facilitates high-quality metadata creation and validation when using spreadsheets
- Takes advantage of:
    - CEDAR’s library of customizable metadata templates for reporting guidelines
    - Established controlled terminologies and ontologies



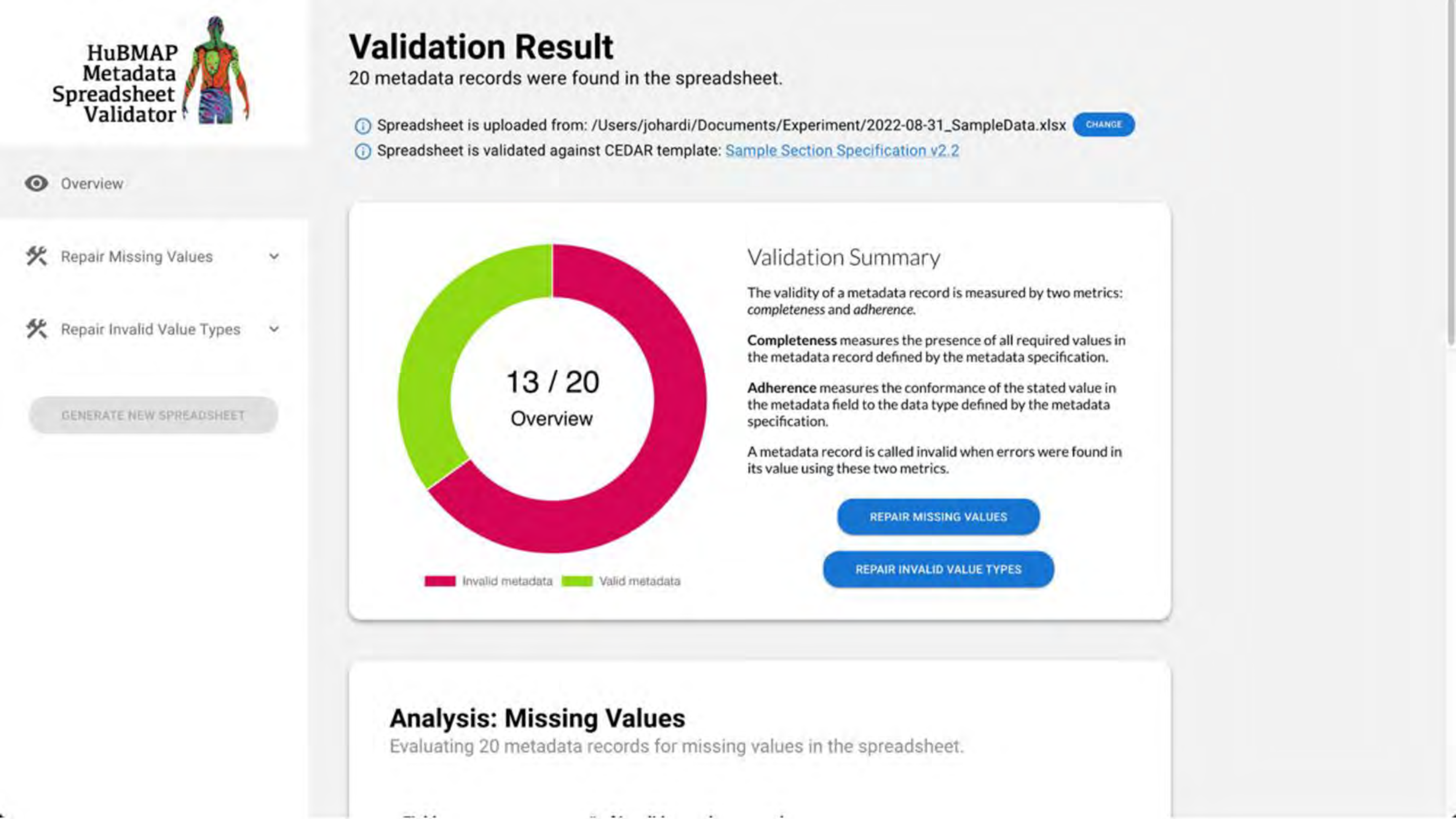
	A	B	C	D	E	F	G	
1	sample_ID	source_storage_ti	source_storage_ti	preparation_mediur	preparation_cond	processing_tim	processing_tim	storage_me
2	Visium_9OLC_A4_S1	208	day	Methanol (100%)	-20 celsius	4	minute	OCT embec
3	Visium_9OLC_A4_S2	208	day	Methanol (100%)	-20 celsius	4	minute	OCT embec
4	Visium_9OLC_I4_S1	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



Validation Result

20 metadata records were found in the spreadsheet.

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

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.*

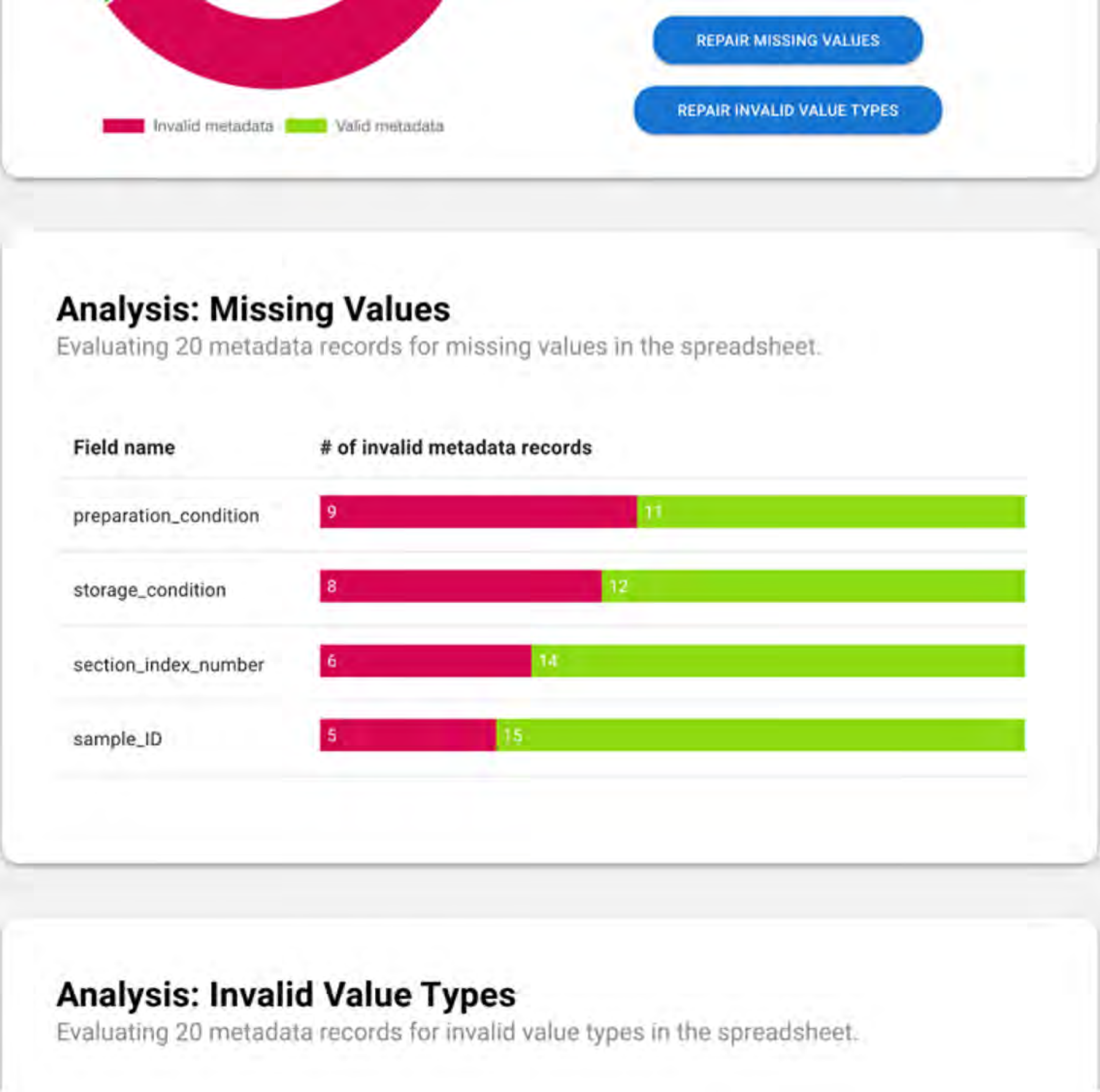
Completeness measures the presence of all required values in the metadata record defined by the metadata specification.

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.

Analysis: Missing Values

Evaluating 20 metadata records for missing values in the spreadsheet.



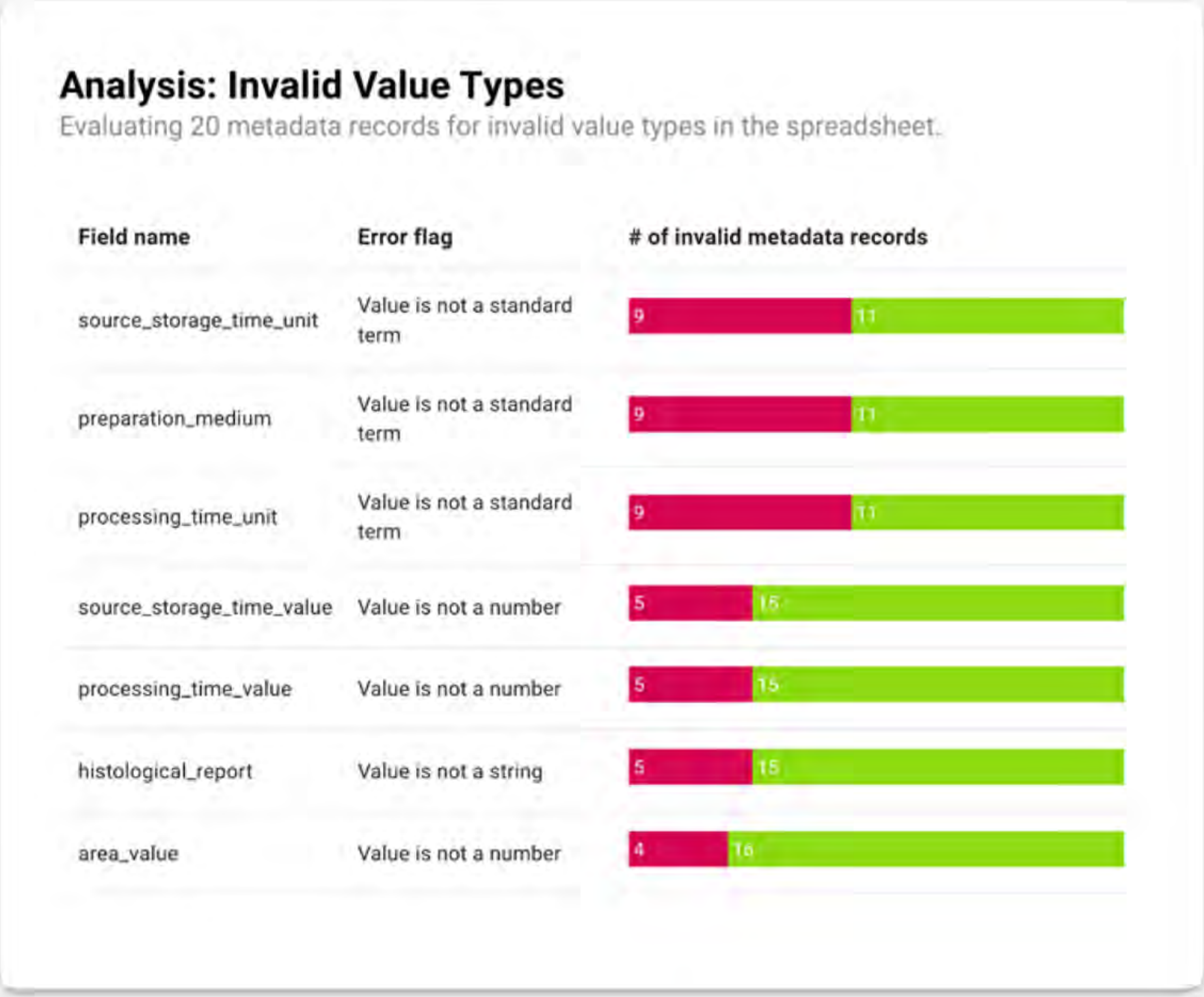
Analysis: Missing Values

Evaluating 20 metadata records for missing values in the spreadsheet.

Field name	# of invalid metadata records
preparation_condition	🟥 9 🟩 11
storage_condition	🟥 8 🟩 12
section_index_number	🟥 6 🟩 14
sample_ID	🟥 5 🟩 15

Analysis: Invalid Value Types

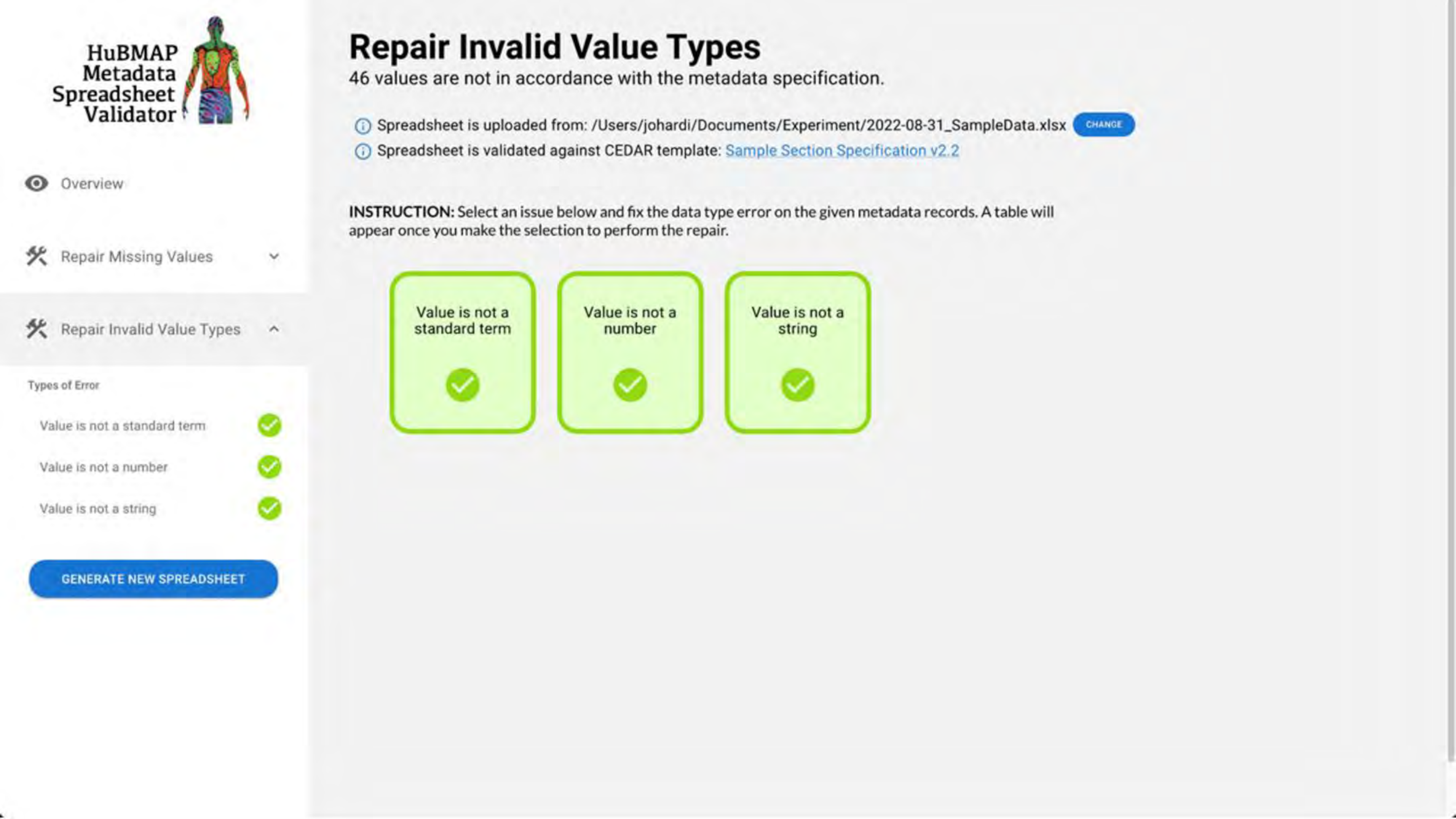
Evaluating 20 metadata records for invalid value types in the spreadsheet.



Analysis: Invalid Value Types

Evaluating 20 metadata records for invalid value types in the spreadsheet.

Field name	Error flag	# of invalid metadata records
source_storage_time_unit	Value is not a standard term	🟥 9 🟩 11
preparation_medium	Value is not a standard term	🟥 9 🟩 11
processing_time_unit	Value is not a standard term	🟥 9 🟩 11
source_storage_time_value	Value is not a number	🟥 5 🟩 15
processing_time_value	Value is not a number	🟥 5 🟩 15
histological_report	Value is not a string	🟥 5 🟩 15
area_value	Value is not a number	🟥 4 🟩 16



Repair Invalid Value Types

46 values are not in accordance with the metadata specification.

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

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

INSTRUCTION: Select an issue below and fix the data type error on the given metadata records. A table will appear once you make the selection to perform the repair.

Value is not a standard term ✓

Value is not a number ✓

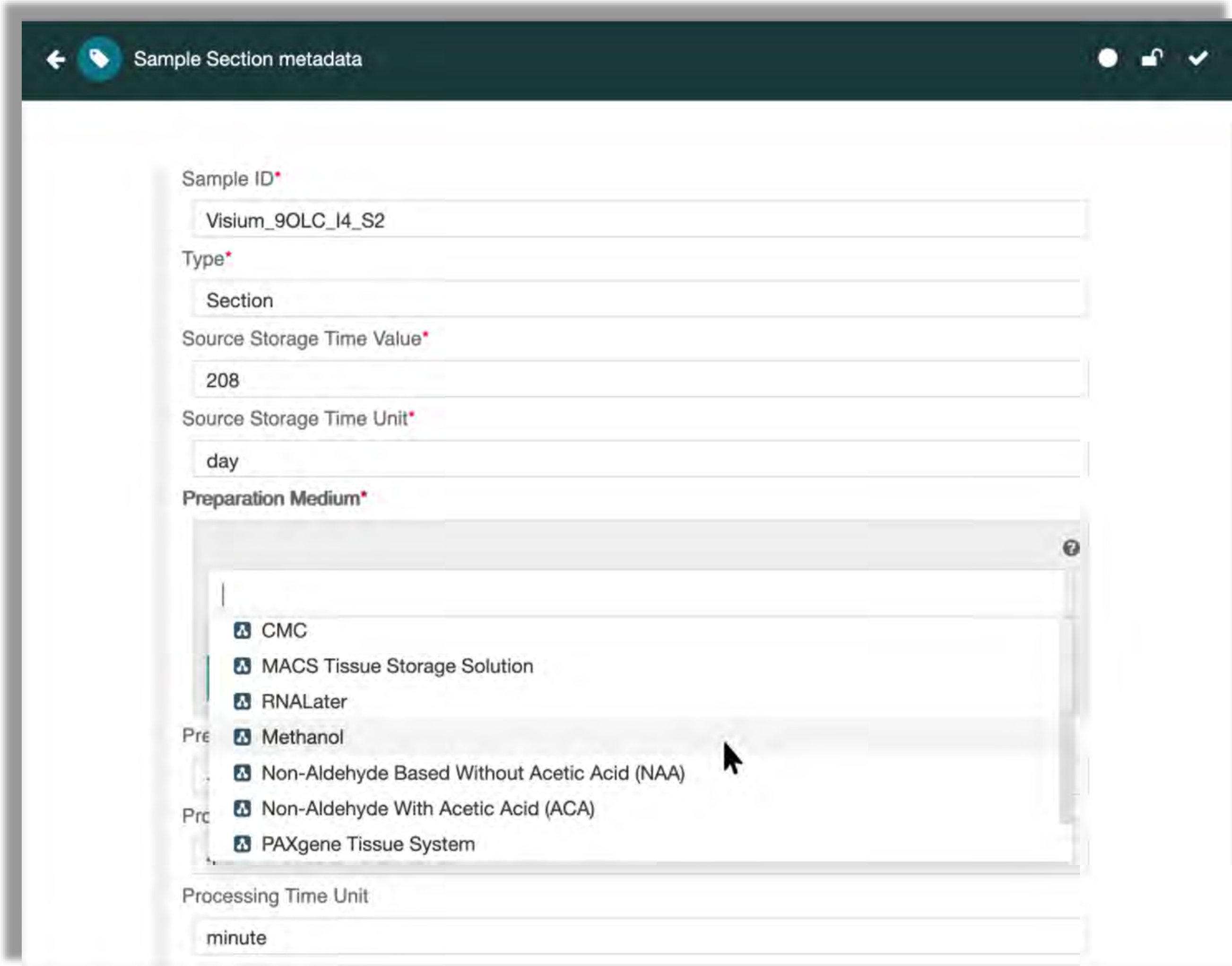
Value is not a string ✓

Technology for better metadata

1. **CEDAR Workbench:** An editor that helps researchers to create standards-compliant metadata from
    - Ontologies
    - Reporting guidelines
2. **CEDAR Metadata Validator:** A system that validates spreadsheet-entered metadata against CEDAR templates

A metadata template …

- Serves as a *knowledge base* of a scientific community's metadata preferences
- Captures those preferences in a *reusable, standardized form*
- Can be used by *people* to review, enhance, or build on those preferences
- Can be accessed by *machines* to assist in a variety of tasks



Sample Section metadata

Sample ID*: Visium_9OLC_I4_S2

Type*: Section

Source Storage Time Value*: 208

Source Storage Time Unit*: day

Preparation Medium*

Processing Time Unit: minute

Metadata for Machines Workshops

- Are intensive 1–3 day invited, highly participatory sessions
- Historically, have been hosted by the GO FAIR Organization
- Lead groups of scientists to consensus regarding
    - Ontologies
    - Reporting guidelines

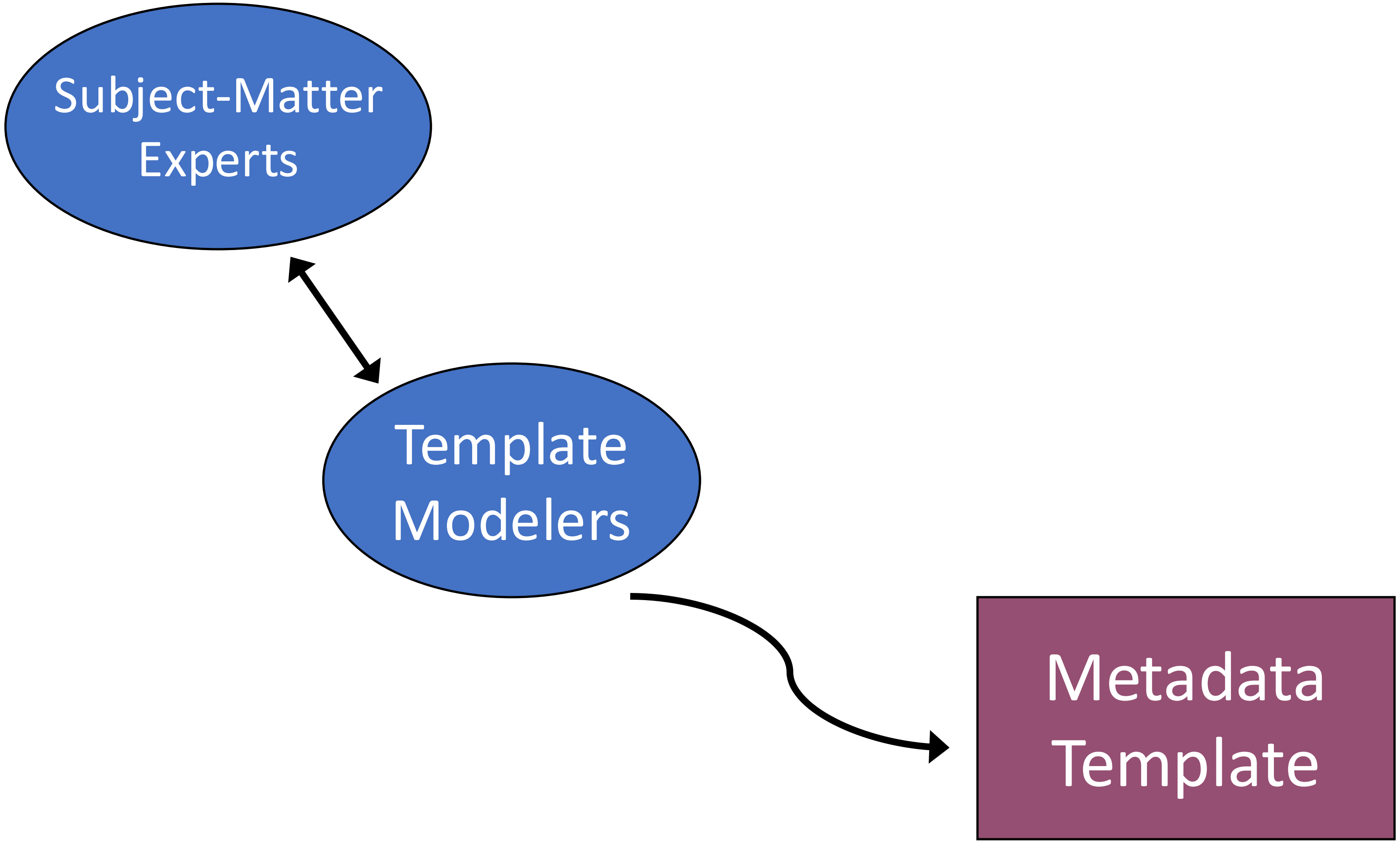
  for different
    - Areas of science
    - Classes of experiments
- Ultimately result in new CEDAR metadata templates

The Netherlands Organization for Health Research and Development

- Has hosted Metadata for Machines workshops to develop metadata templates and controlled terminologies needed for all its funded research related to COVID
- Uses CEDAR to create the metadata templates during the workshops
- Mandates the use of these metadata templates *as a condition of funding*
- Is now expanding the use of M4Ms and standardized metadata into other areas of research that it supports

ZonMw

Building a CEDAR metadata template



Subject-Matter Experts

Template Modelers

Metadata Template

CEDAR metadata templates communicate community standards in a reusable manner

- Capture knowledge of a professional community in machine-readable form (in our case, knowledge of preferred metadata standards)
- Help the community to codify its knowledge in a public, inspectable, editable place
- Ideally, can plug-and-play with a variety of software systems
- Allow the community—and others outside the community—to reuse the knowledge over and over again

A CEDAR templates are like cartridges that can plug into a variety of applications to communicate preferences regarding community-based metadata standards.

Making data FAIR requires good metadata; making good metadata requires:

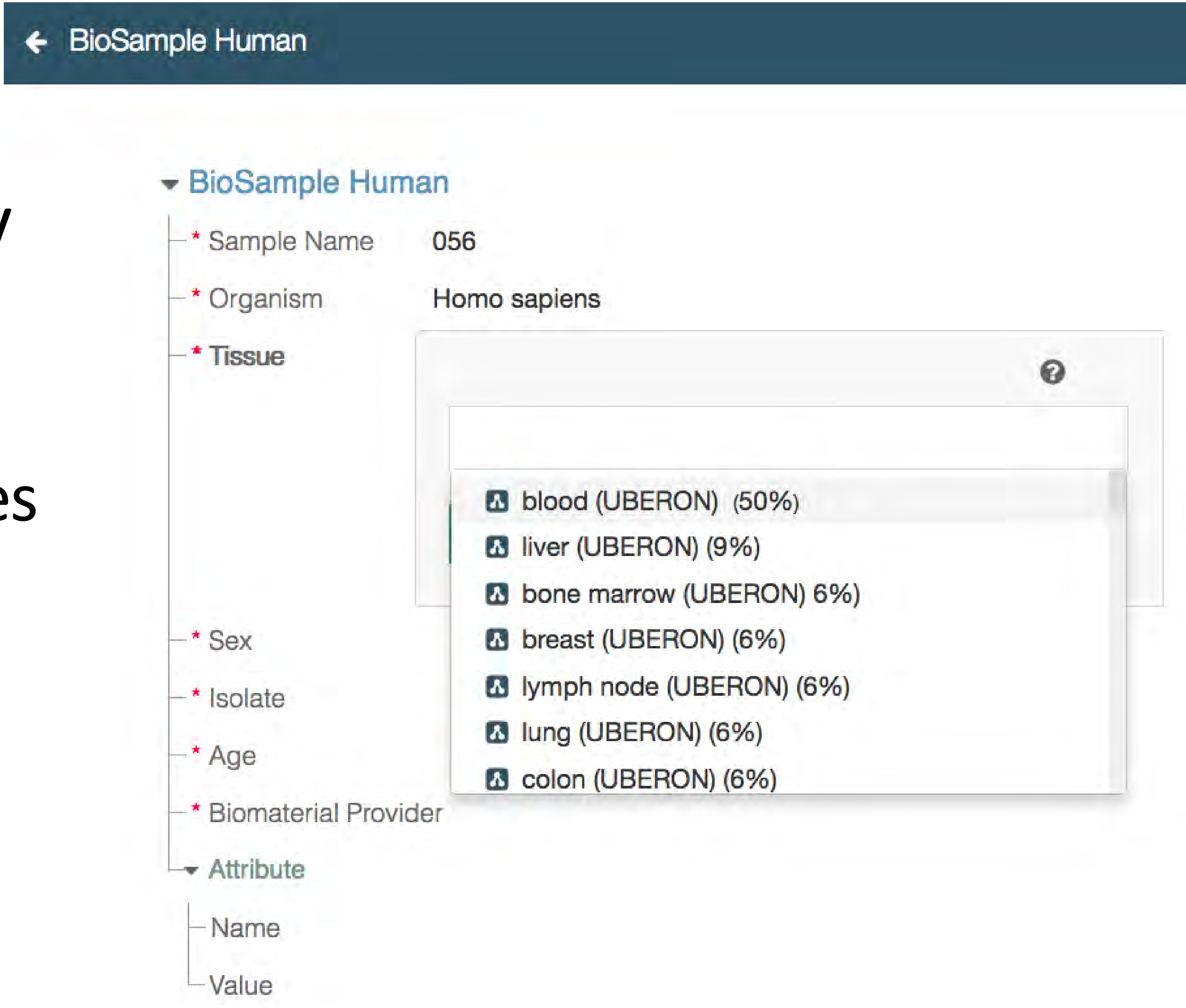
- Community-endorsed metadata standards for all areas of science
- Technology
  - Like **CEDAR**,
    to help create standards-adherent metadata in the first place
  - Like the **CEDAR Metadata Validator**
    to help improve metadata entered from spreadsheets
- A concerted effort on the part of funders, publishers, professional societies, and investigators to stimulate the creation of the standards needed to advance science

Data will not be FAIR until …

- Funding agencies enforce their requirements
- Publishers demand it
- Investigators feel peer pressure
- Academic institutions deem the sharing of FAIR data to be an essential component of scholarship
- Professional societies take the lead in developing community-based standards for their constituencies

In the meantime, semantic technology remains the key to

- Making data FAIR
- Enabling third parties to find and access other people's data
- Making new discoveries through data reuse



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