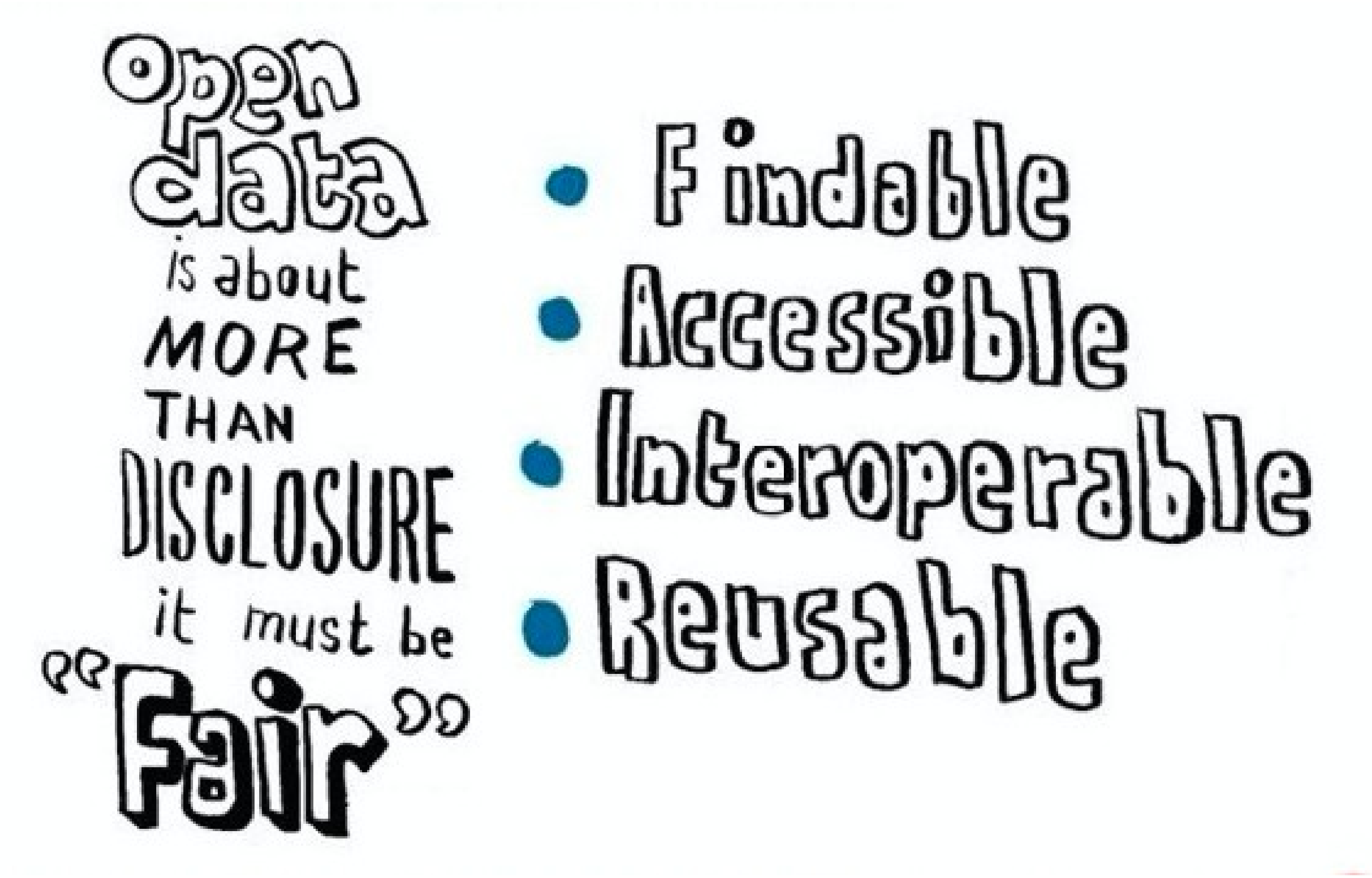
The Center for Expanded Data Annotation and Retrieval:

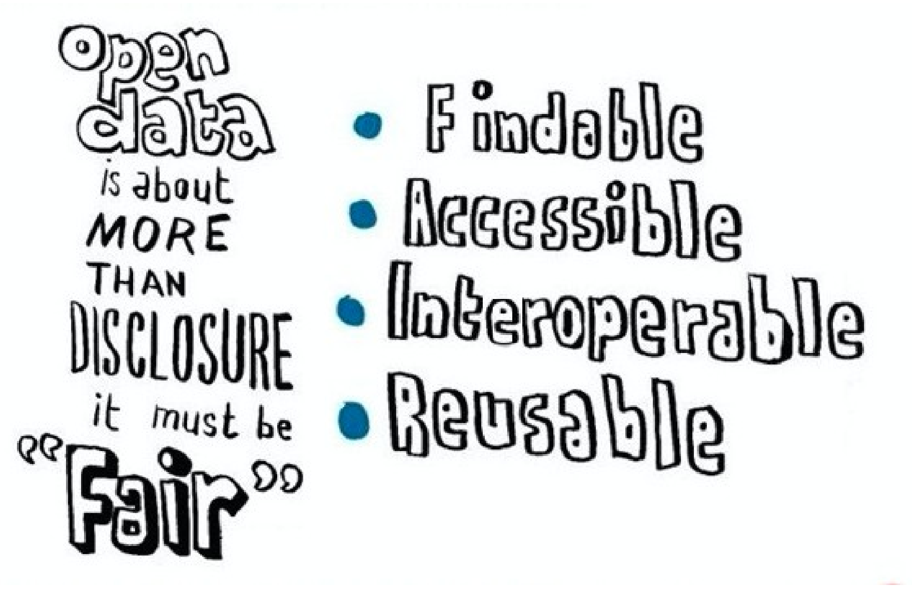
Better Metadata Through Domain-Specific V

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

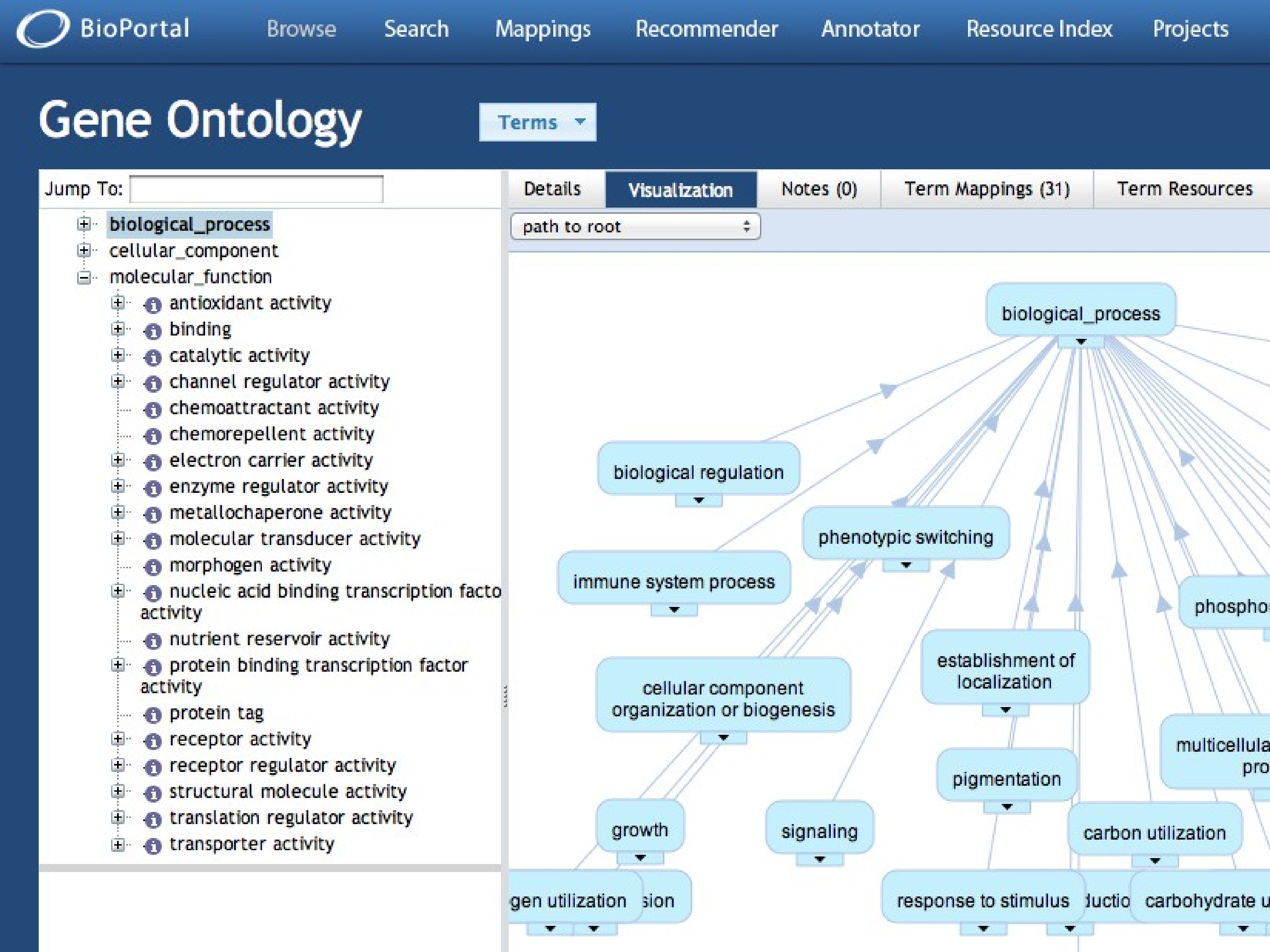
CEDAR

CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL


open data is about MORE THAN DISCLOSURE it must be "Fair"
• Findable
• Accessible
• Interoperable
• Reusable



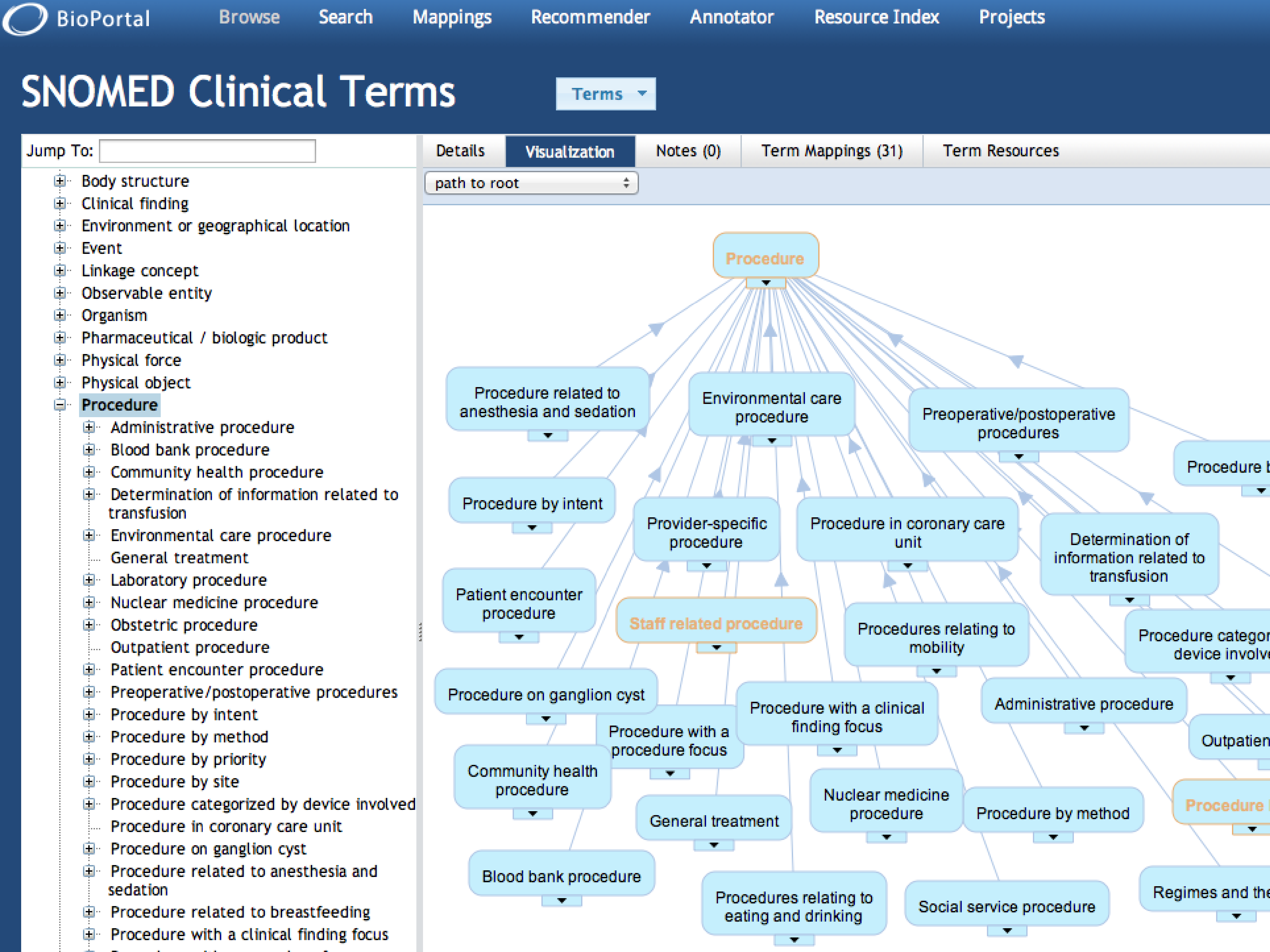
Problem #1: Describe what exists in the world completely and consistently using ontologies



Gene Ontology

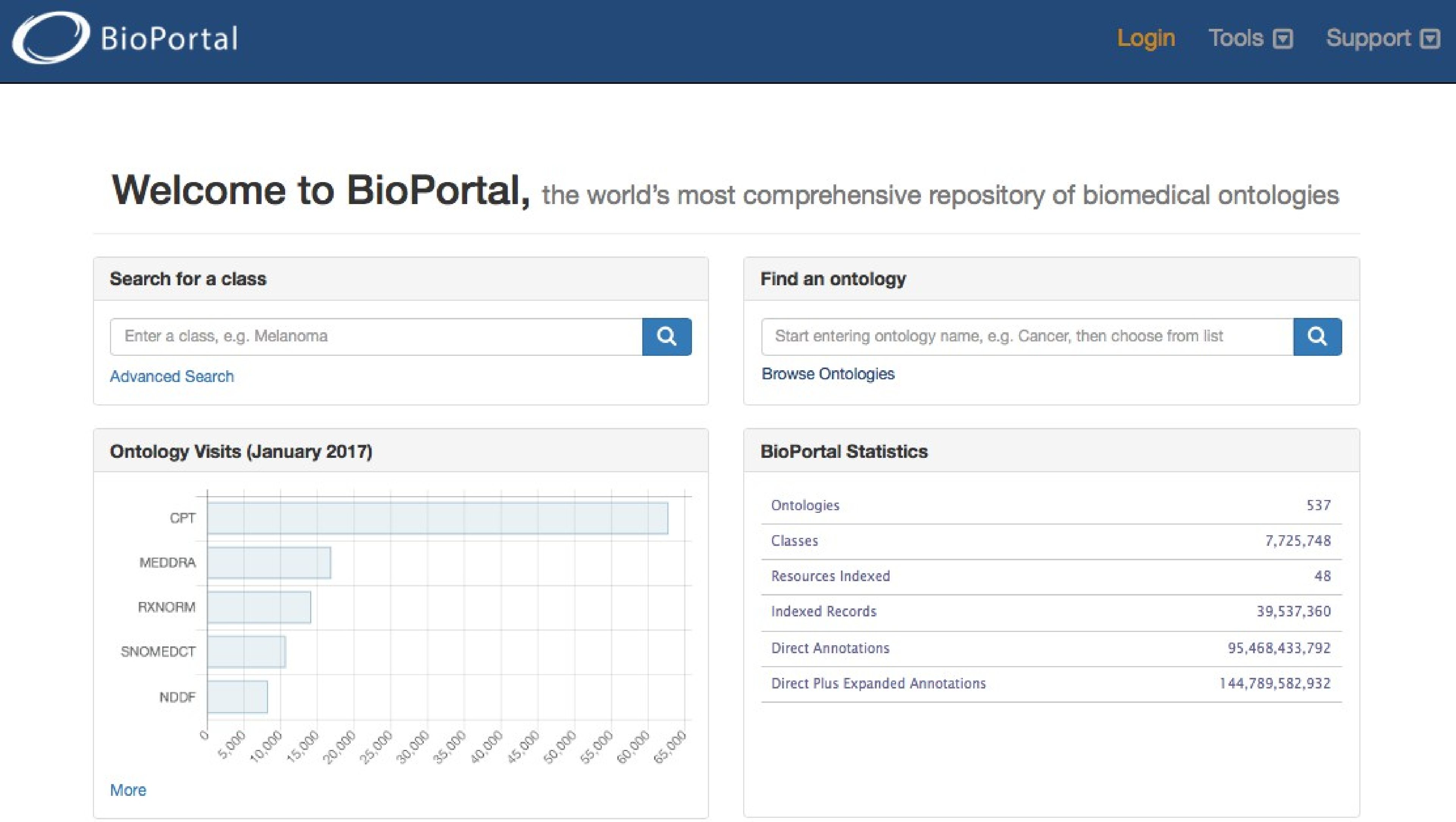
Jump To:

- biological_process
- cellular_component
- molecular_function
  - antioxidant activity
  - binding
  - catalytic activity
  - channel regulator activity
  - chemoattractant activity
  - chemorepellent activity
  - electron carrier activity
  - enzyme regulator activity
  - metallochaperone activity
  - molecular transducer activity
  - morphogen activity
  - nucleic acid binding transcription facto activity
  - nutrient reservoir activity
  - protein binding transcription factor activity
  - protein tag
  - receptor activity
  - receptor regulator activity
  - structural molecule activity
  - translation regulator activity
  - transporter activity


BioPortal
Browse
Search
Mappings
Recommender
Annotator
Resource Index
Projects
Terms
Jump To:
Details
Visualization
Notes (0)
Term Mappings (31)
Term Resources
path to root
Procedure
Procedure related to anesthesia and sedation
Environmental care procedure
Preoperative/postoperative procedures
Procedure b
Procedure by intent
Provider-specific procedure
Procedure in coronary care unit
Determination of information related to transfusion
Patient encounter procedure
Staff related procedure
Procedures relating to mobility
Procedure categor device involve
Procedure on ganglion cyst
Procedure with a clinical finding focus
Administrative procedure
Procedure with a procedure focus
Outpatien
Community health procedure
Nuclear medicine procedure
Procedure by method
Procedure
General treatment
Blood bank procedure
Procedures relating to eating and drinking
Social service procedure
Regimes and the

SNOMED Clinical Terms

- Body structure
- Clinical finding
- Environment or geographical location
- Event
- Linkage concept
- Observable entity
- Organism
- Pharmaceutical / biologic product
- Physical force
- Physical object
- Procedure
  - Administrative procedure
  - Blood bank procedure
  - Community health procedure
  - Determination of information related to transfusion
  - Environmental care procedure
  - General treatment
  - Laboratory procedure
  - Nuclear medicine procedure
  - Obstetric procedure
  - Outpatient procedure
  - Patient encounter procedure
  - Preoperative/postoperative procedures
  - Procedure by intent
  - Procedure by method
  - Procedure by priority
  - Procedure by site
  - Procedure categorized by device involved
  - Procedure in coronary care unit
  - Procedure on ganglion cyst
  - Procedure related to anesthesia and sedation
  - Procedure related to breastfeeding
  - Procedure with a clinical finding focus



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

Search for a class

Find an ontology

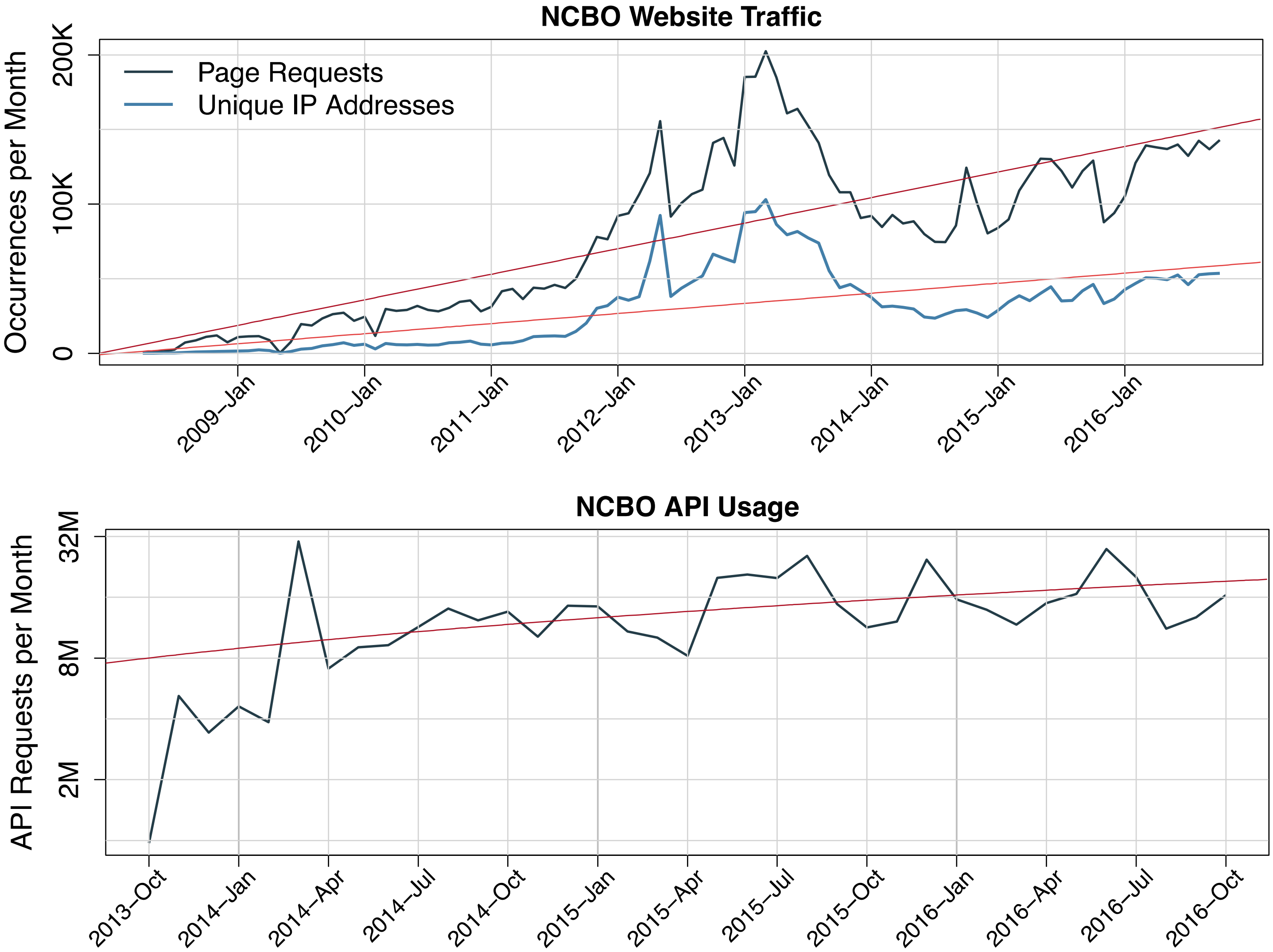
Start entering ontology name, e.g. Cancer, then choose from list
Browse Ontologies

Ontology Visits (January 2017)

BioPortal Statistics

Ontologies	537
Classes	7,725,748
Resources Indexed	48
Indexed Records	39,537,360
Direct Annotations	95,468,433,792
Direct Plus Expanded Annotations	144,789,582,932

http://bioportal.bioontology.org


NCBO Website Traffic
Occurrences per Month
0
100K
200K
Page Requests
Unique IP Addresses
2009–Jan
2010–Jan
2011–Jan
2012–Jan
2013–Jan
2014–Jan
2015–Jan
2016–Jan
NCBO API Usage
API Requests per Month
2M
8M
32M
2013–Oct
2014–Jan
2014–Apr
2014–Jul
2014–Oct
2015–Jan
2015–Apr
2015–Jul
2015–Oct
2016–Jan
2016–Apr
2016–Jul
2016–Oct

Who is using NCBO technology?

openMDR

Laboratree alpha

OpenClinica® Open Source for Clinical Research

OASIS ONTOLOGIE

Technische Universität Braunschweig

SilicoTryp

RAPID

Resourc

abase
Integrated Rice Science Database

SEEK Beta

WIKIPEDIA The Free Encyclopedia

Gene Wiki
From Wikipedia, the free encyclopedia

University

BRISSKit
Funded by BRISSkit

MAVIR Consorcio MAVIR

OntoFactory

University of Stavanger

Ontological Discovery Environment

University of Leicester

EMBL-EBI European Bioinformatics Institute

eurreca
EURopean micronutrient RECommendations Aligned

isacreator

GWAS CENTRAL

Synapse BETA

ODiSSea ★★★★★ (5.0 based
SciVerse Hub

i2b2 WIKI HOM Home

Galaxy Tool Shed

iPlant Collaborative™

MGH MASSACHUSETTS GENERAL HOSPITAL

BIRN

BioScholar

CTSA Clinical & Translational Science Awards®
Translating Discoveries to Medical Practice

CBiB Centre de Bioinformatique de Bordeaux

ONTOLOGY DEVELOPMENT • INFORMATION EXTRACTION

BIO SITEMAPS EDITOR

DOMEO ANNOTATION TOOLKIT

UNIVERSITY OF AMSTERDAM

THE UNIVERSITY OF TEXAS MDAnderson ~~Cancer~~ Center
Making Cancer History®

MGI

OntoFinder

UNIVERSITY OF AMSTERDAM

BUCK INSTITUTE FOR RESEARCH ON AGING

MICAD Molecular Imaging & Contrast Age

STANFORD SCHOOL of MEDICINE

Stanford Center for Clinical Informatics

university of groningen

Pacific Northwest NATIONAL LABORATORY

BioLit

RIKEN

ontogrator BETA

caNanoLab

IPB Leibniz Institute of Plant Biochemistry
Foundation under German Law

IO INFORMATICS

MIND INFORMATICS

MAYO CLINIC

S3DB

eleMAP

Kunnskapsegget

Neural Electro Magnetic Ontologies

sswap.info beta
Simple Semantic Web Architecture and Protocol
iPlant Collaborative™

Modularize

MERB v 1.1 CBiB • Contact

eagle-i consortium

SE-mantic
M-edical
I-mage
A-nnotation

Resource Discovery System

iCAT

DBCLS Database Center for Life Science

ontocat

SimTK

UTHealth™ The University of Texas

<caTag

RED FLY

The DXBrain Project

nbic netherlands bioinformatics centre

protégé

the Gene Ontology

Sage BIONETWORKS

Tufts UNIVERSITY

RSNA Informatics™

RadLex®

Imperial College London

dbNP

Nutritional Phenotype Database
Developer's Corner

NIF

Qiime Quantitative Insights Into Microbial Ecology

NEBC NERC Environmental Bioinformatics Centre

University of Colorado Boulder

gsk GlaxoSmithKline

Microsoft® Research

BioDAG Builder

Ontology Add-in for Word

ols2owl

The Office of the National Coordinator for Health Information Technology

Case-Based Reasoning System

Linking Open Drug Data (LODD)

The University Of Sheffield.

carcinoGENOMICS
a Project of the European Union

g2R

gminer

NMC Data Support Platform
a joint initiative of the NMC and NuGO, powered by NBIC

GEN2PHEN

cgfb

UNIVERSITÉ DE BORDEAUX

INRA

cancergrid

biosharing

Netherlands Toxicogenomics CENTRE

STANFORD ENGINEERING

NuGO Nutrigenomics Organisation

SIDR

UCSF

meducator world wide med

MEDICO RadSpeech

RightField

W3C

Institute of Bioinformatics

THESEUS

HARVARD UNIVERSITY

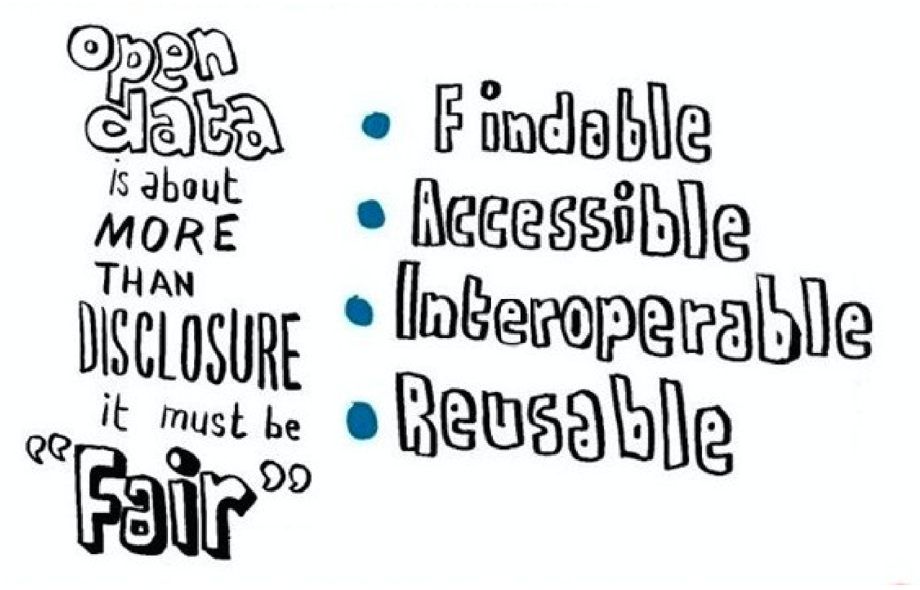
MG-RAST metagenomics analysis server

CardioVascular Research Grid

UNIVERSITAS BERGENSIS

University of Bergen

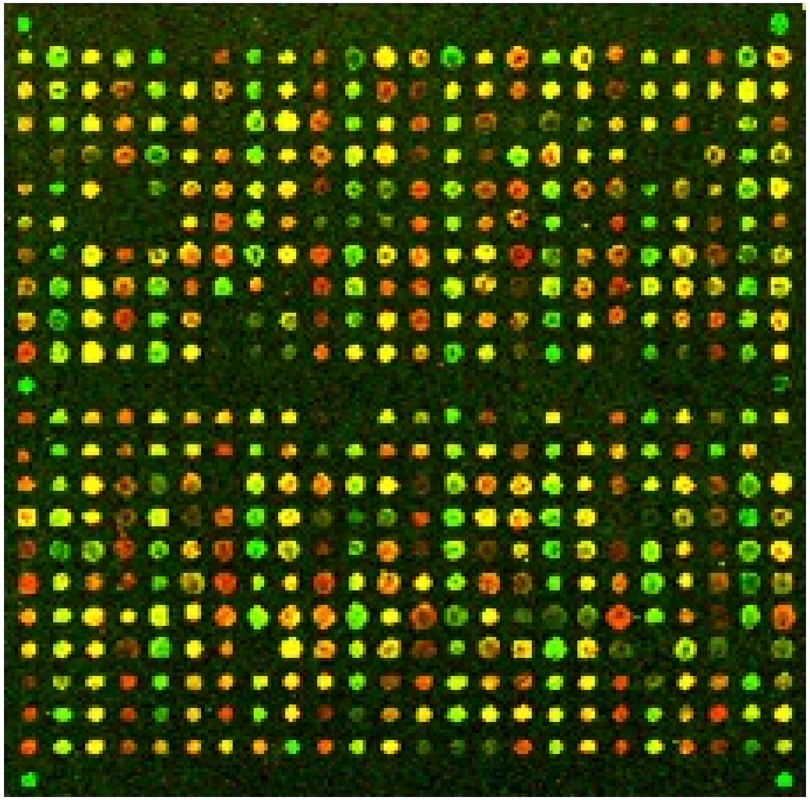
SCDE STEM CELL DISCOVERY ENGINE



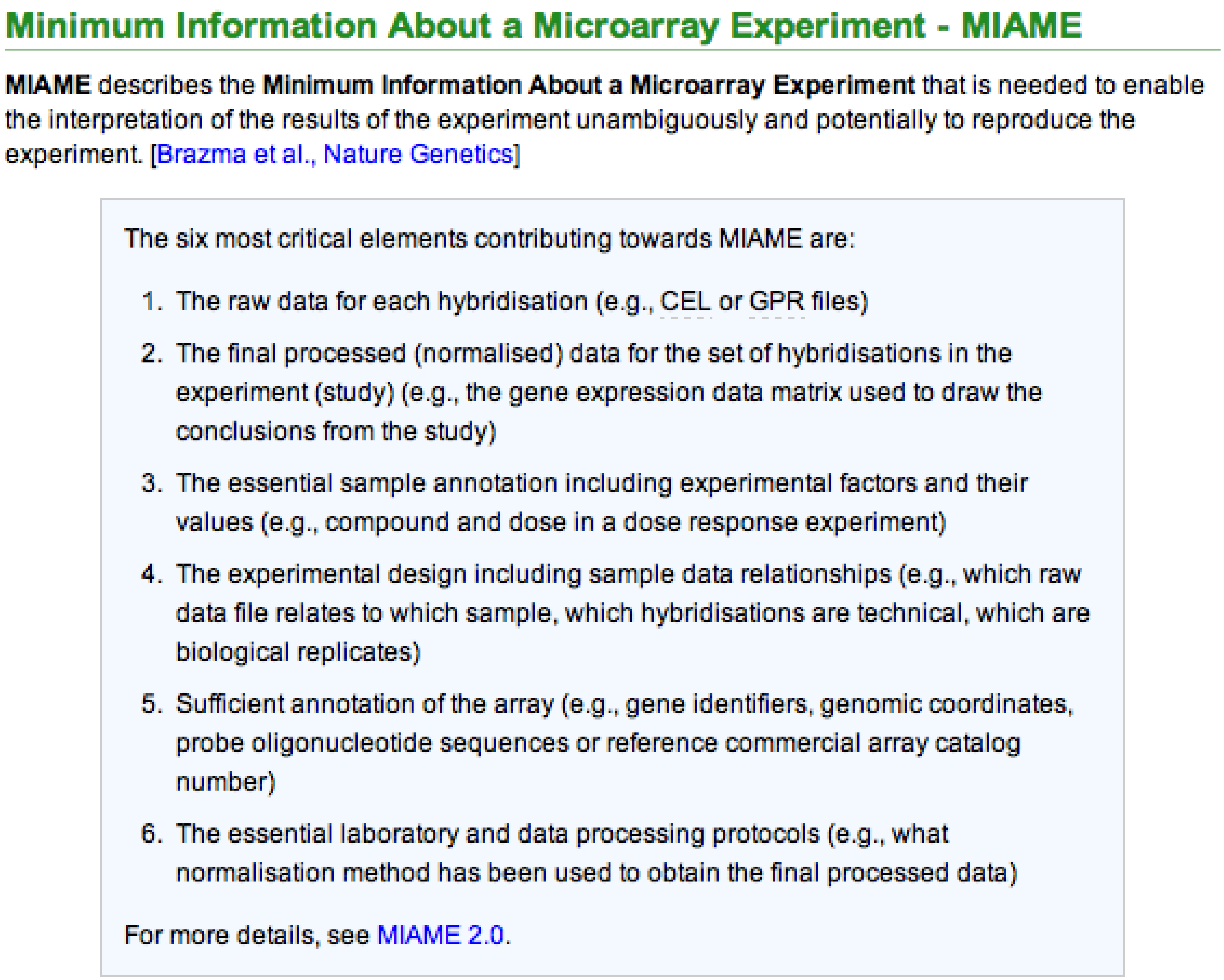
Problem #2: Describe properties of experiments completely and consistently

The microarray community took the lead in standardizing metadata

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



- DNA Microarra



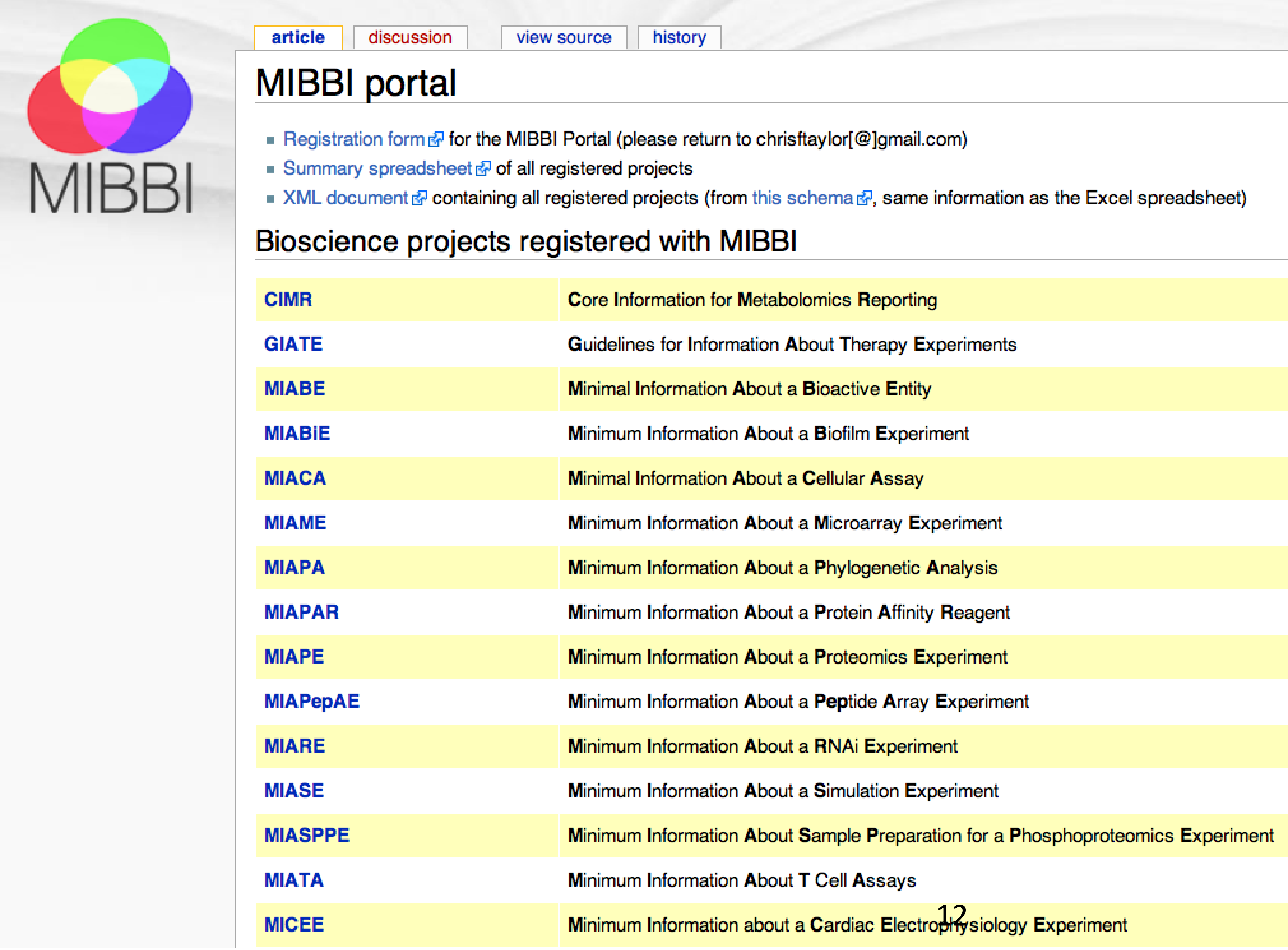
Minimum Information About a Microarray Experiment - MIAME

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.

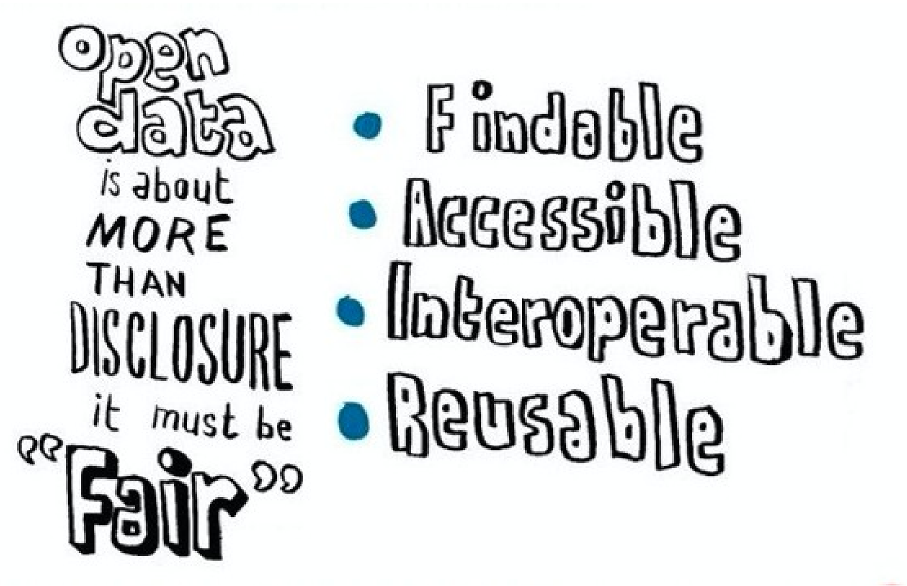


MIBBI portal

- Registration form for the MIBBI Portal (please return to chrisftaylor[@]gmail.com)
- Summary spreadsheet of all registered projects
- XML document containing all registered projects (from this schema, same information as the Excel spreadsheet)

Bioscience projects registered with MIBBI

CIMR	**C**ore **I**nformation for **M**etabolomics **R**eporting
GIATE	**G**uidelines for **I**nformation **A**bout **T**herapy **E**xperiments
MIABE	**M**inimal **I**nformation **A**bout a **B**ioactive **E**ntity
MIABiE	**M**inimum **I**nformation **A**bout a **Bi**ofilm **E**xperiment
MIACA	**M**inimal **I**nformation **A**bout a **C**ellular **A**ssay
MIAME	**M**inimum **I**nformation **A**bout a **M**icroarray **E**xperiment
MIAPA	**M**inimum **I**nformation **A**bout a **P**hylogenetic **A**nalysis
MIAPAR	**M**inimum **I**nformation **A**bout a **P**rotein **A**ffinity **R**eagent
MIAPE	**M**inimum **I**nformation **A**bout a **P**roteomics **E**xperiment
MIAPepAE	**M**inimum **I**nformation **A**bout a **Pep**tide **A**rray **E**xperiment
MIARE	**M**inimum **I**nformation **A**bout a **R**NAi **E**xperiment
MIASE	**M**inimum **I**nformation **A**bout a **S**imulation **E**xperiment
MIASPPE	**M**inimum **I**nformation **A**bout **S**ample **P**reparation for a **P**hosphoproteomics **E**xperiment
MIATA	**M**inimum **I**nformation **A**bout **T** Cell **A**ssays
MICEE	**M**inimum **I**nformation about a **C**ardiac **E**lectrophysiology **E**xperiment

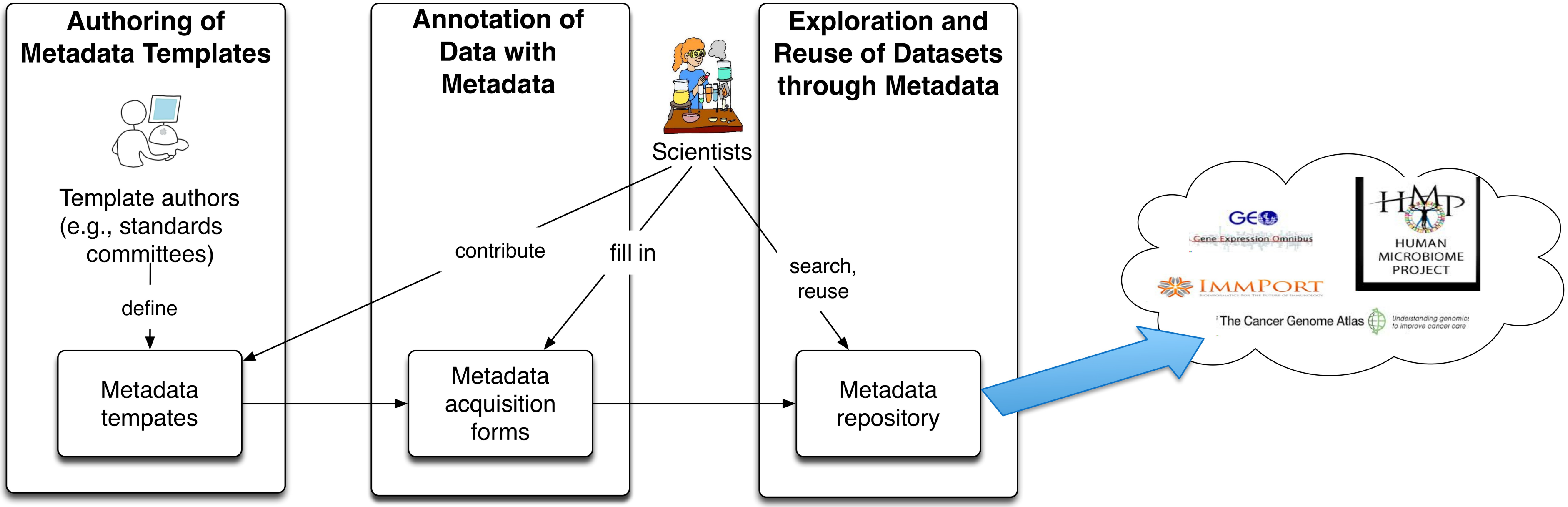


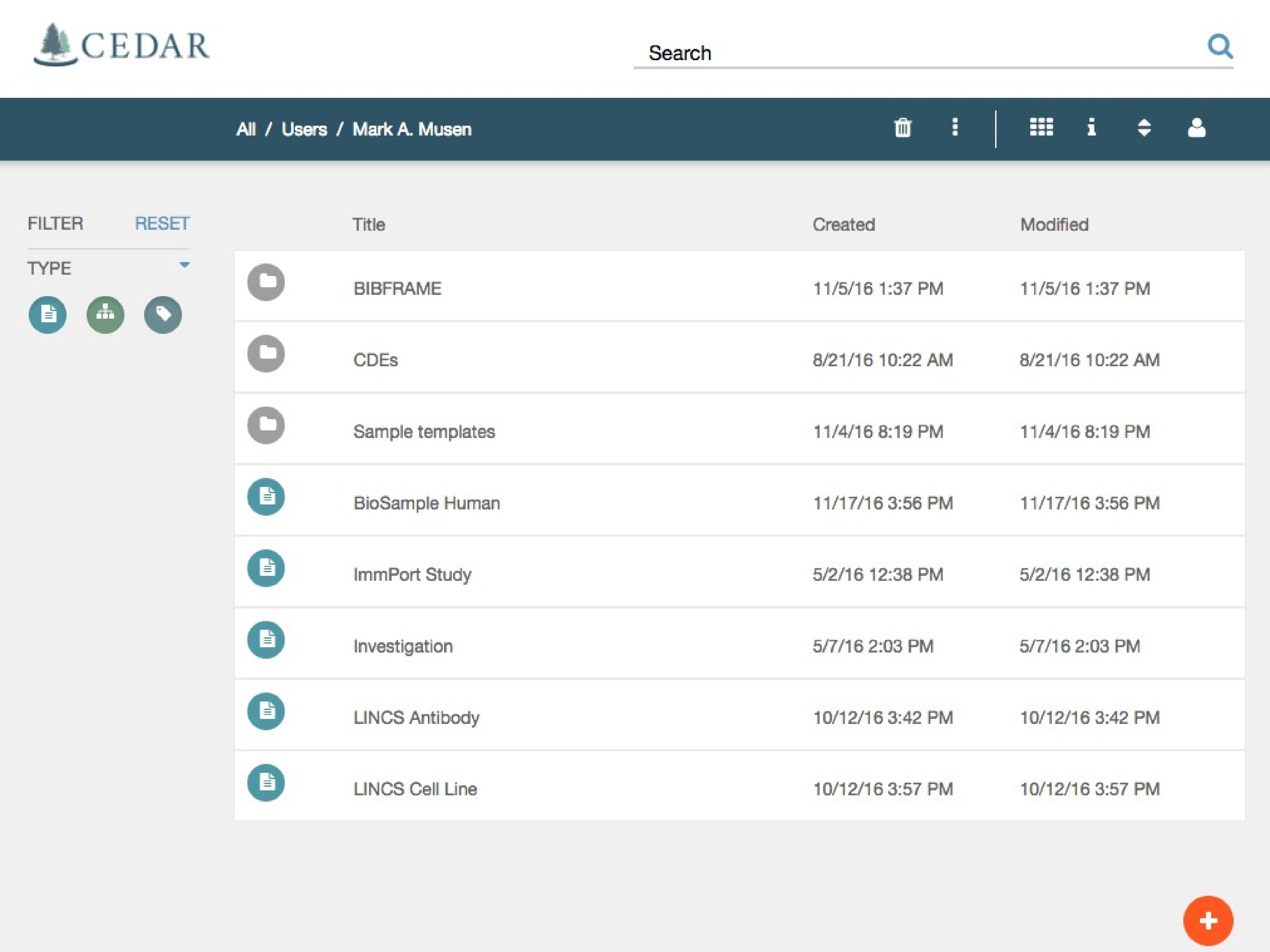
Problem #3: Make it palatable to describe experiments completely and consistently http://metadatacenter.org

CEDAR

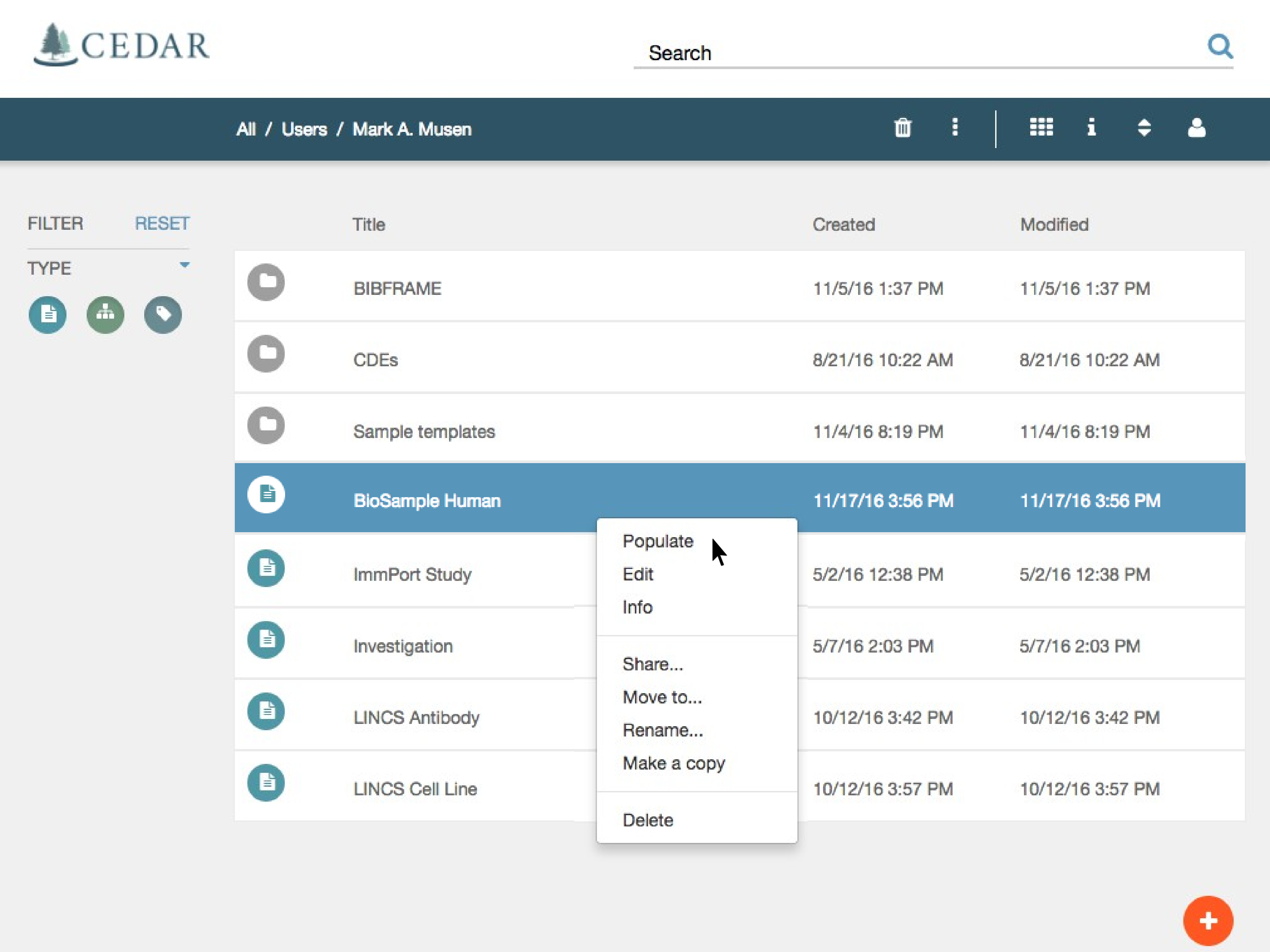
CENTER FOR EXPANDED DATA ANNOTATION
AND RETRIEVAL

The CEDAR Approach to Metadata

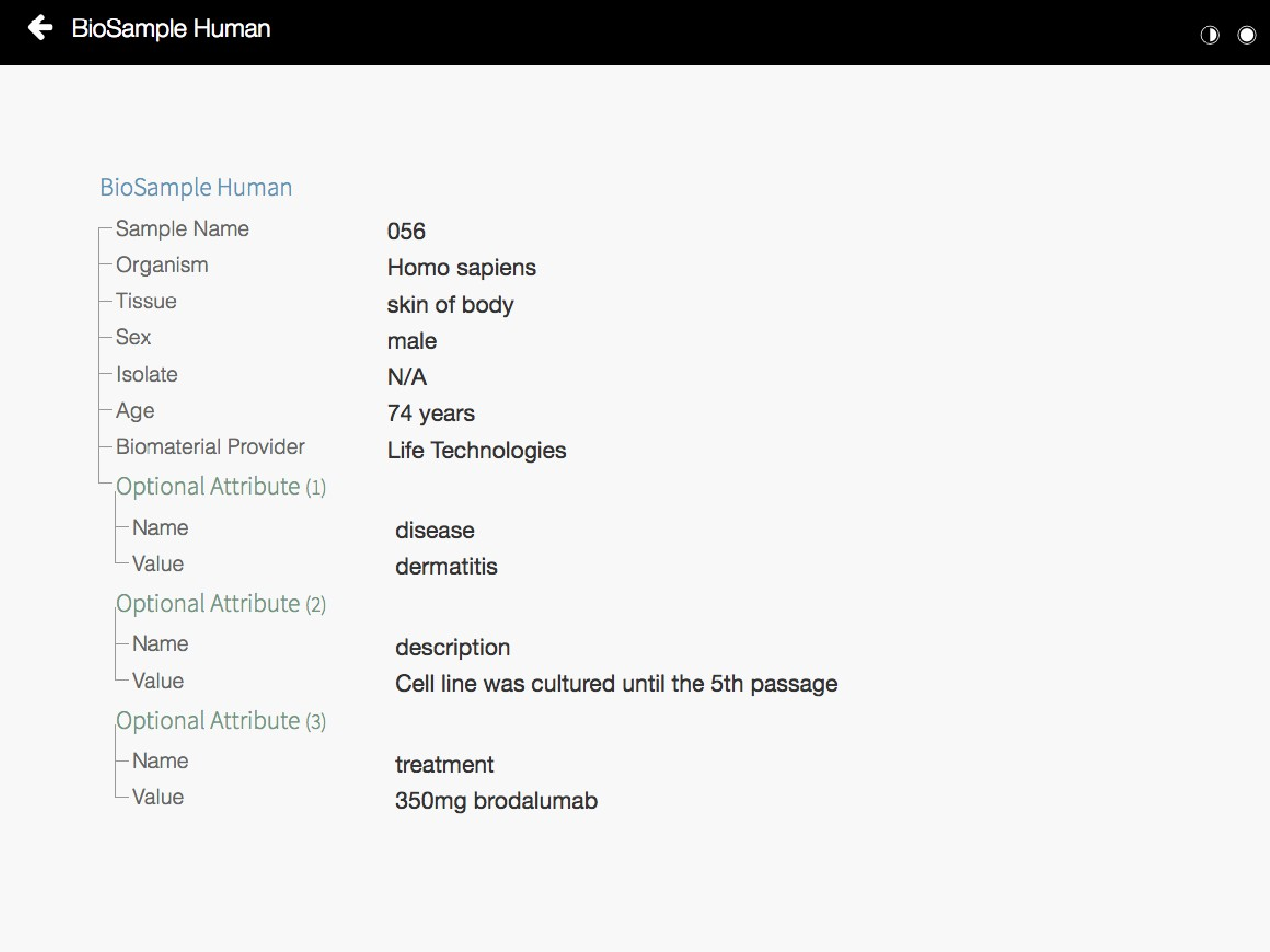

Authoring of Metadata Templates
Template authors (e.g., standards committees)
define
Metadata tempates
Annotation of Data with Metadata
contribute
fill in
Metadata acquisition forms
Scientists
Exploration and Reuse of Datasets through Metadata
search, reuse
Metadata repository
GEO
Gene Expression Omnibus
HMP
HUMAN MICROBIOME PROJECT
IMMPORT
The Cancer Genome Atlas
Understanding genomics to improve cancer care



	Title	Created	Modified
📁	BIBFRAME	11/5/16 1:37 PM	11/5/16 1:37 PM
📁	CDEs	8/21/16 10:22 AM	8/21/16 10:22 AM
📁	Sample templates	11/4/16 8:19 PM	11/4/16 8:19 PM
📄	BioSample Human	11/17/16 3:56 PM	11/17/16 3:56 PM
📄	ImmPort Study	5/2/16 12:38 PM	5/2/16 12:38 PM
📄	Investigation	5/7/16 2:03 PM	5/7/16 2:03 PM
📄	LINCS Antibody	10/12/16 3:42 PM	10/12/16 3:42 PM
📄	LINCS Cell Line	10/12/16 3:57 PM	10/12/16 3:57 PM



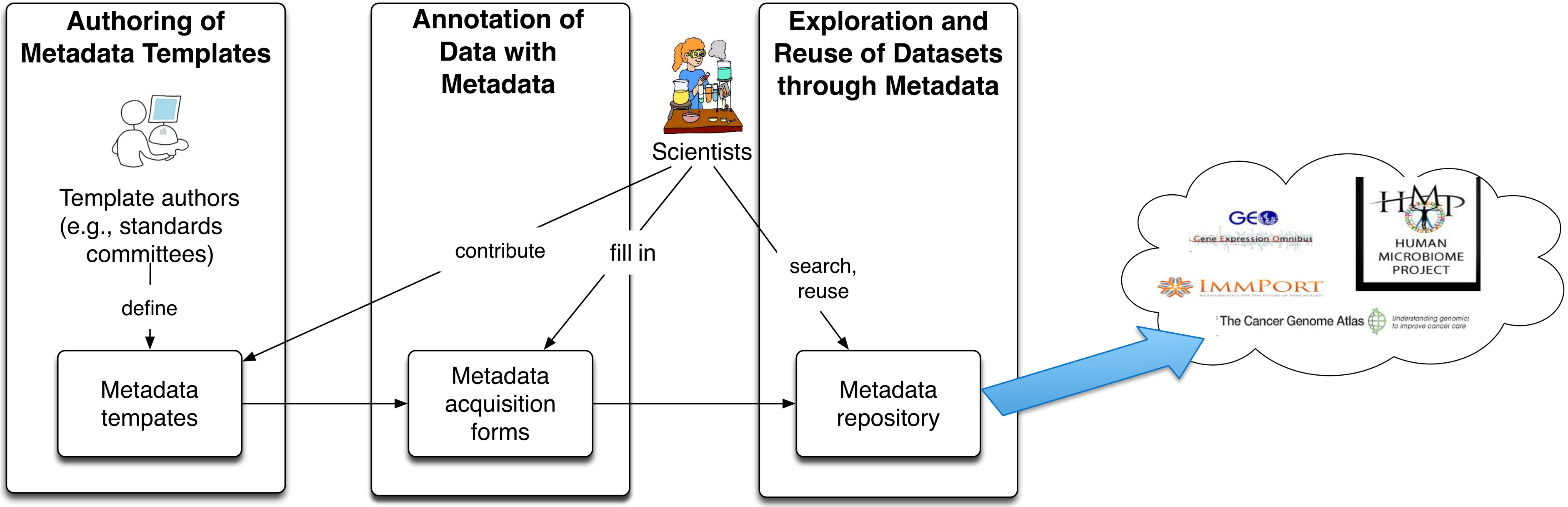
Title	Created	Modified
BIBFRAME	11/5/16 1:37 PM	11/5/16 1:37 PM
CDEs	8/21/16 10:22 AM	8/21/16 10:22 AM
Sample templates	11/4/16 8:19 PM	11/4/16 8:19 PM
BioSample Human	11/17/16 3:56 PM	11/17/16 3:56 PM
ImmPort Study	5/2/16 12:38 PM	5/2/16 12:38 PM
Investigation	5/7/16 2:03 PM	5/7/16 2:03 PM
LINCS Antibody	10/12/16 3:42 PM	10/12/16 3:42 PM
LINCS Cell Line	10/12/16 3:57 PM	10/12/16 3:57 PM



BioSample Human

Sample Name	056
Organism	Homo sapiens
Tissue	skin of body
Sex	male
Isolate	N/A
Age	74 years
Biomaterial Provider	Life Technologies
Optional Attribute (1)	
Name	disease
Value	dermatitis
Optional Attribute (2)	
Name	description
Value	Cell line was cultured until the 5th passage
Optional Attribute (3)	
Name	treatment
Value	350mg brodalumab

The CEDAR Approach to Metadata



Authoring of Metadata Templates

Template authors (e.g., standards committees)

define

Metadata tempates

Annotation of Data with Metadata

Metadata acquisition forms

Scientists

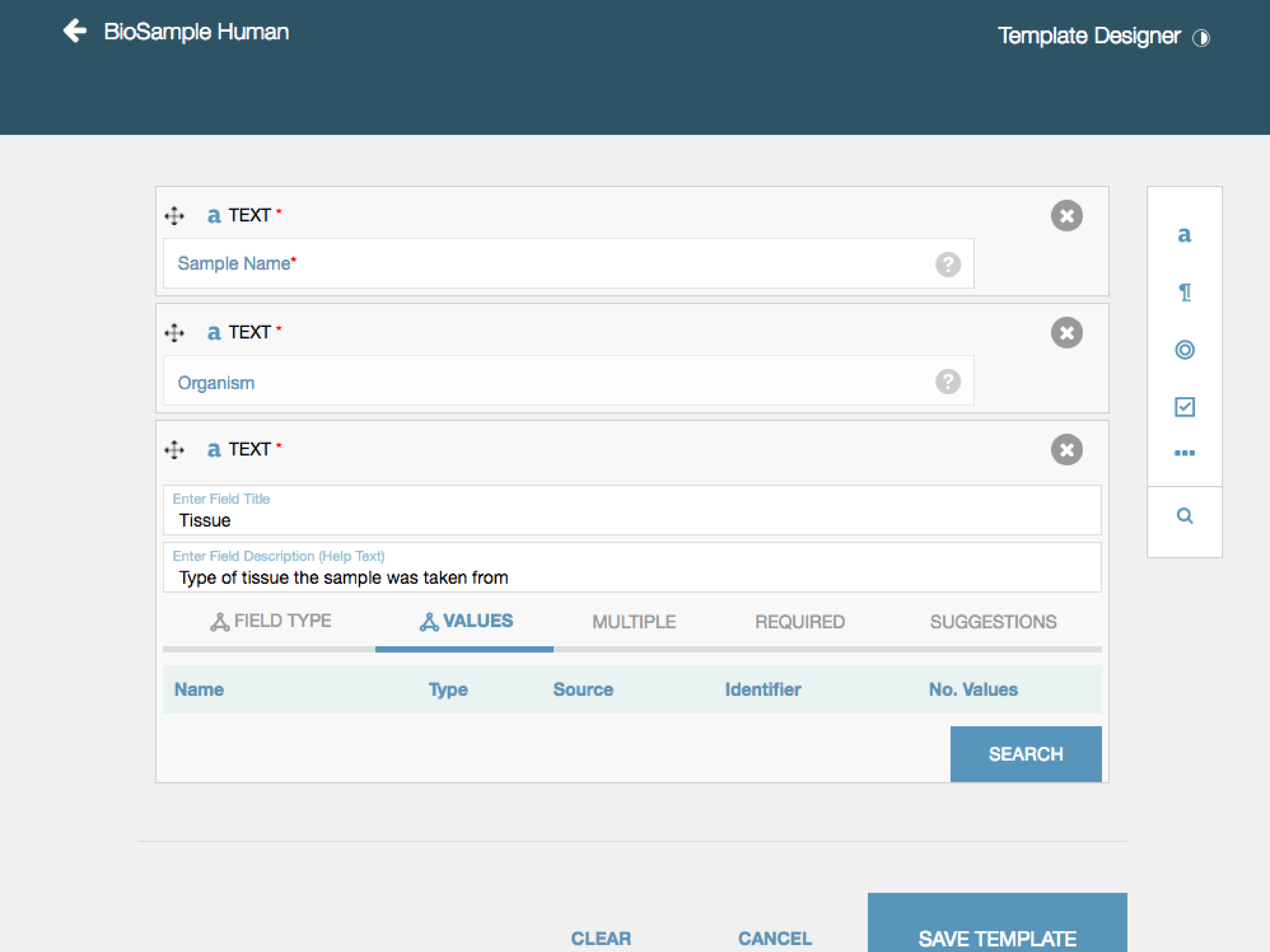
contribute

fill in

search, reuse

Exploration and Reuse of Datasets through Metadata

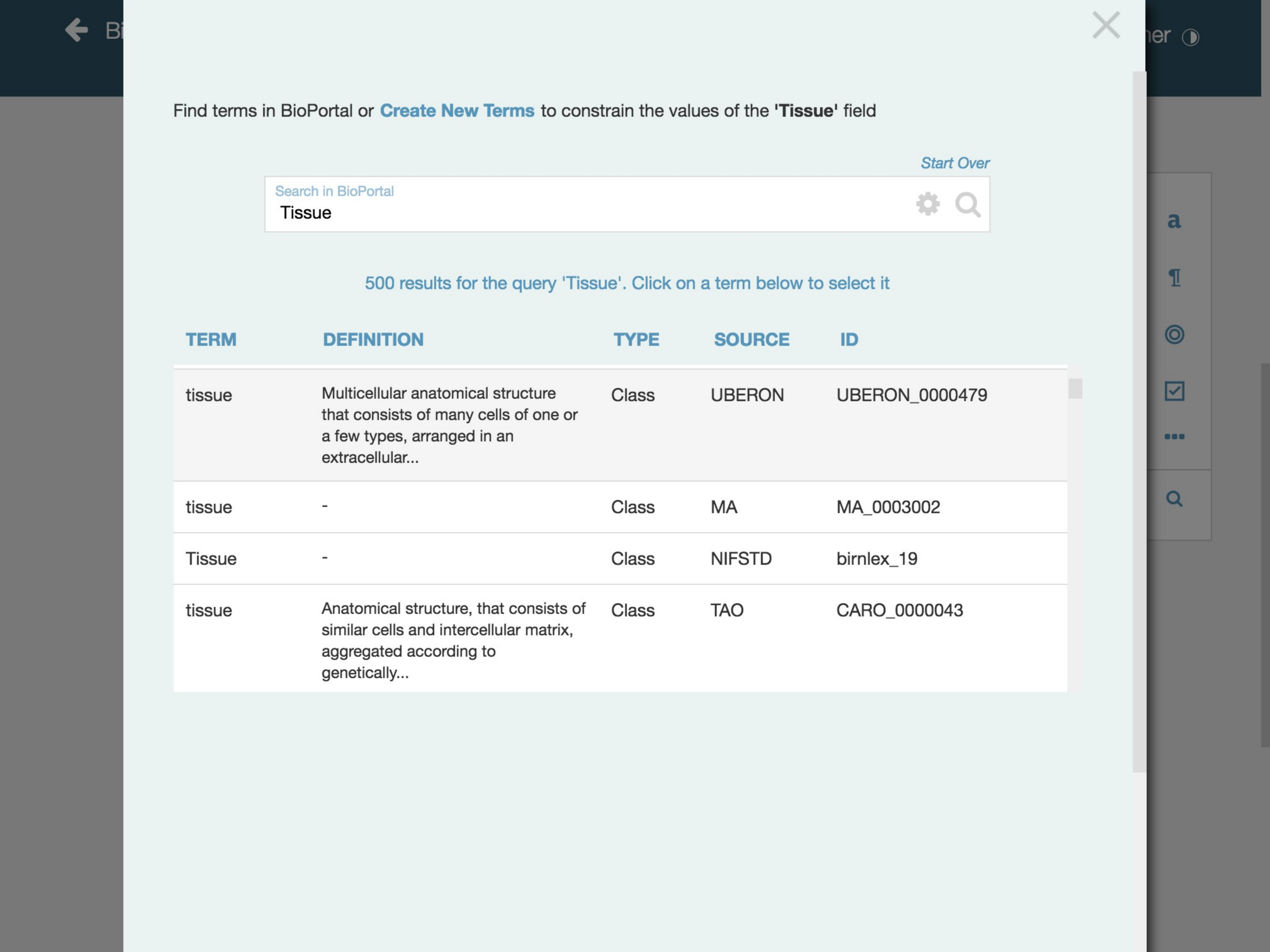
Metadata repository



Enter Field Title
Tissue

Enter Field Description (Help Text)
Type of tissue the sample was taken from

Name	Type	Source	Identifier	No. Values

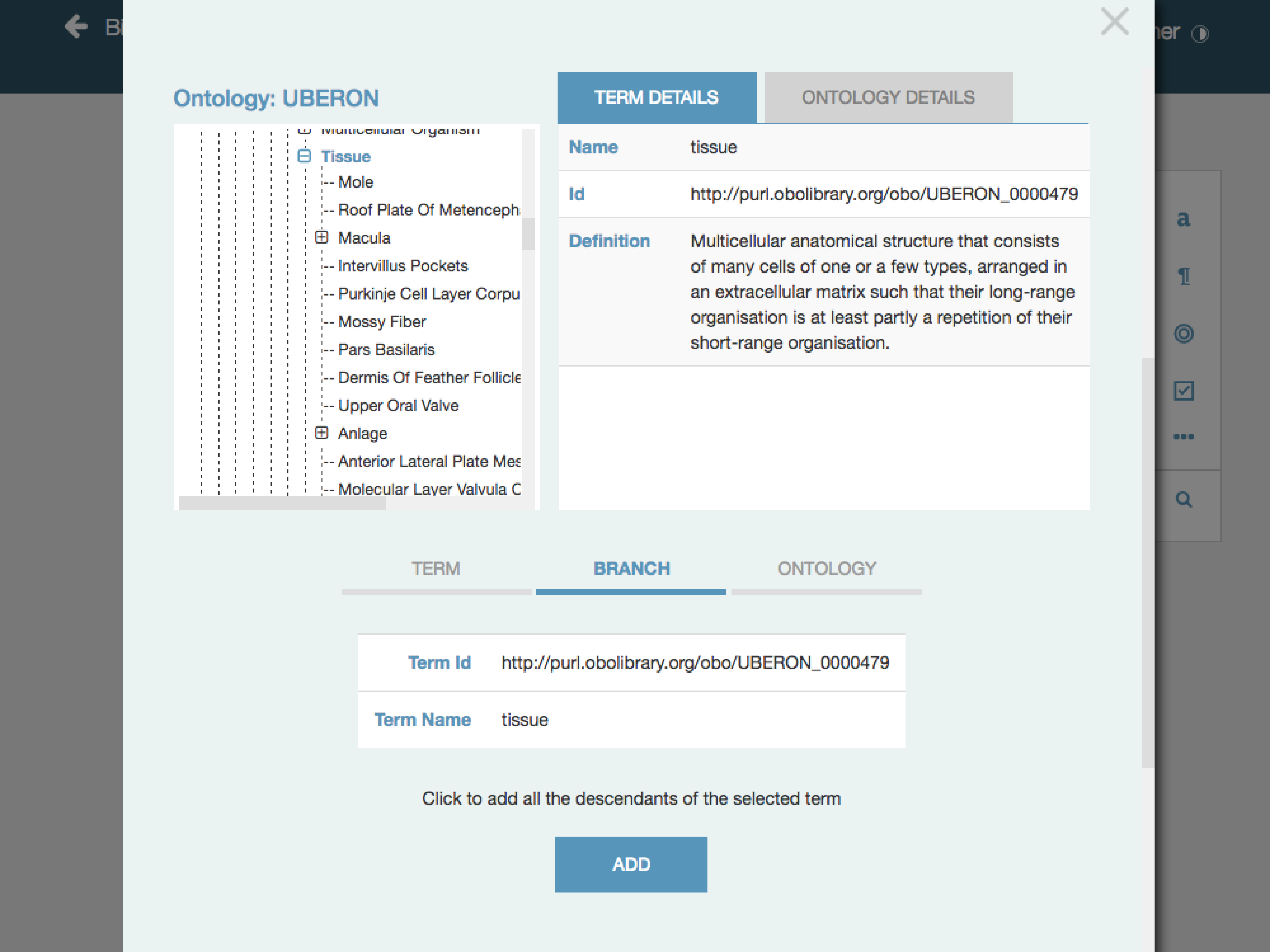


Find terms in BioPortal or **Create New Terms** to constrain the values of the **'Tissue'** field

Tissue

500 results for the query 'Tissue'. Click on a term below to select it

TERM	DEFINITION	TYPE	SOURCE	ID
tissue	Multicellular anatomical structure that consists of many cells of one or a few types, arranged in an extracellular...	Class	UBERON	UBERON_0000479
tissue	-	Class	MA	MA_0003002
Tissue	-	Class	NIFSTD	birnlex_19
tissue	Anatomical structure, that consists of similar cells and intercellular matrix, aggregated according to genetically...	Class	TAO	CARO_0000043



Ontology: UBERON

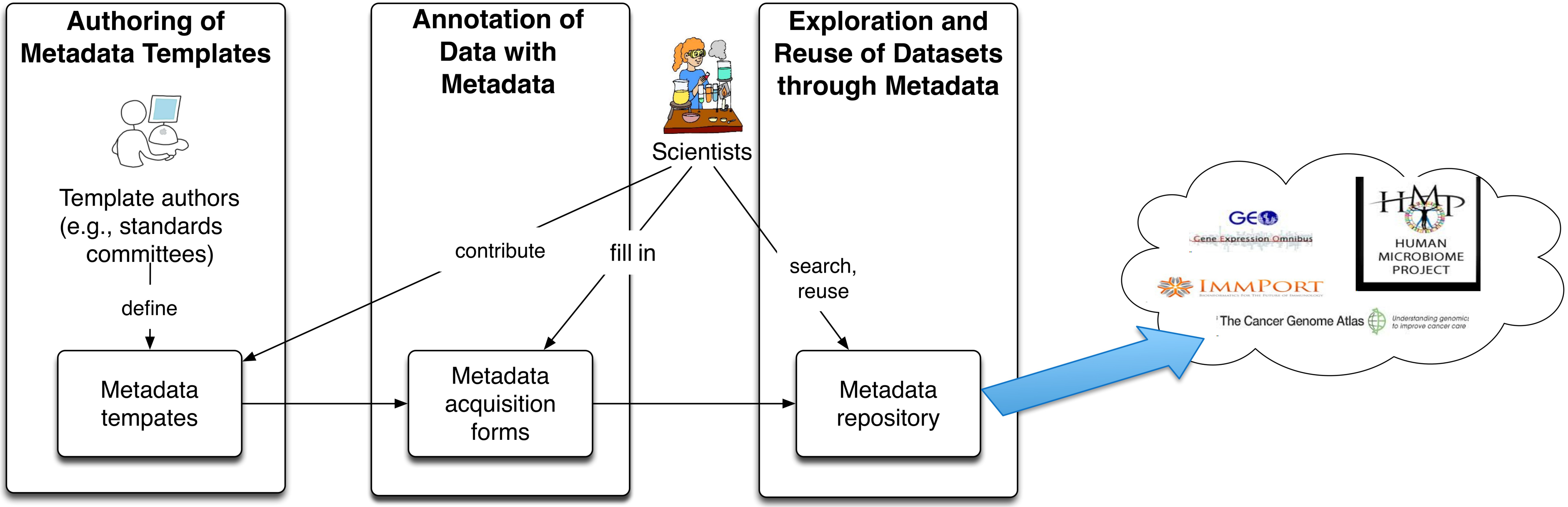
Multicellular Organism
⊟ Tissue
-- Mole
-- Roof Plate Of Metenceph
⊞ Macula
-- Intervillus Pockets
-- Purkinje Cell Layer Corpu
-- Mossy Fiber
-- Pars Basilaris
-- Dermis Of Feather Follicle
-- Upper Oral Valve
⊞ Anlage
-- Anterior Lateral Plate Mes
-- Molecular Layer Valvula C

Name	tissue
Id	http://purl.obolibrary.org/obo/UBERON_0000479
Definition	Multicellular anatomical structure that consists of many cells of one or a few types, arranged in an extracellular matrix such that their long-range organisation is at least partly a repetition of their short-range organisation.

Term Id	http://purl.obolibrary.org/obo/UBERON_0000479
Term Name	tissue

Click to add all the descendants of the selected term

The CEDAR Approach to Metadata



Authoring of Metadata Templates

Template authors (e.g., standards committees)

define

Metadata tempates

Annotation of Data with Metadata

Metadata acquisition forms

Scientists

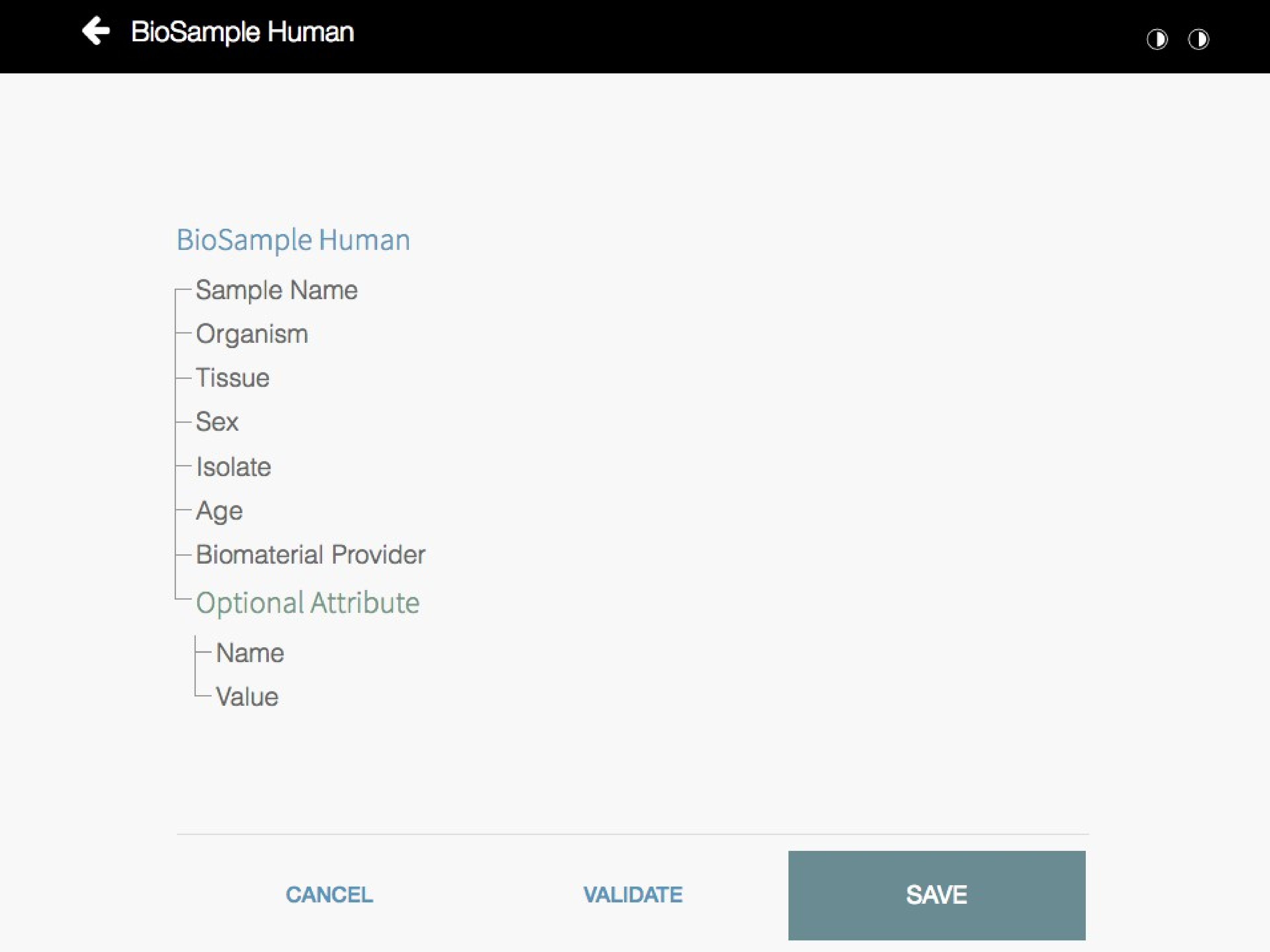
contribute

fill in

search, reuse

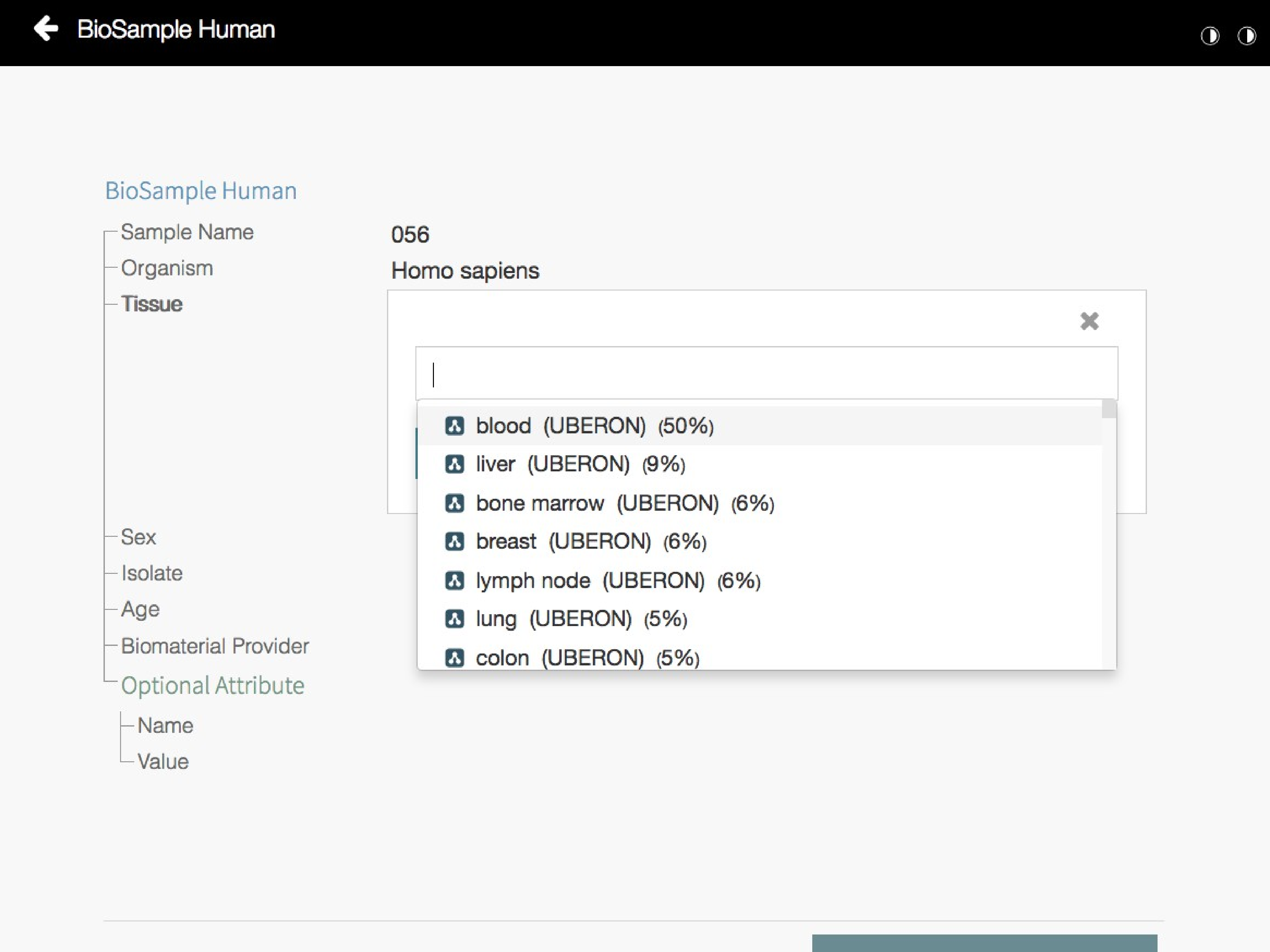
Exploration and Reuse of Datasets through Metadata

Metadata repository



BioSample Human

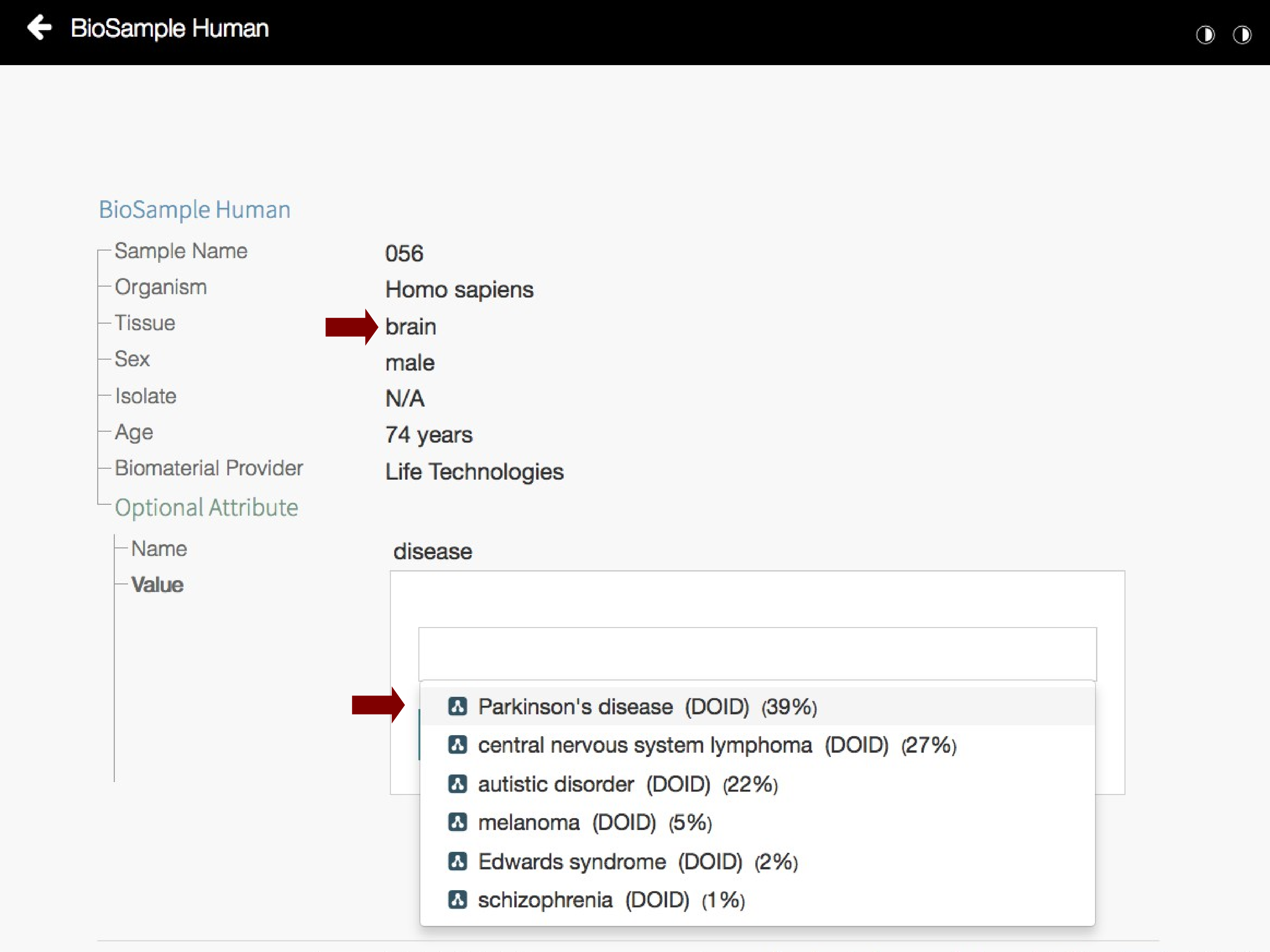
- Sample Name
- Organism
- Tissue
- Sex
- Isolate
- Age
- Biomaterial Provider
- Optional Attribute
  - Name
  - Value



BioSample Human

BioSample Human

- Sample Name: 056
- Organism: Homo sapiens
- **Tissue**
- Sex
- Isolate
- Age
- Biomaterial Provider
- Optional Attribute
  - Name
  - Value

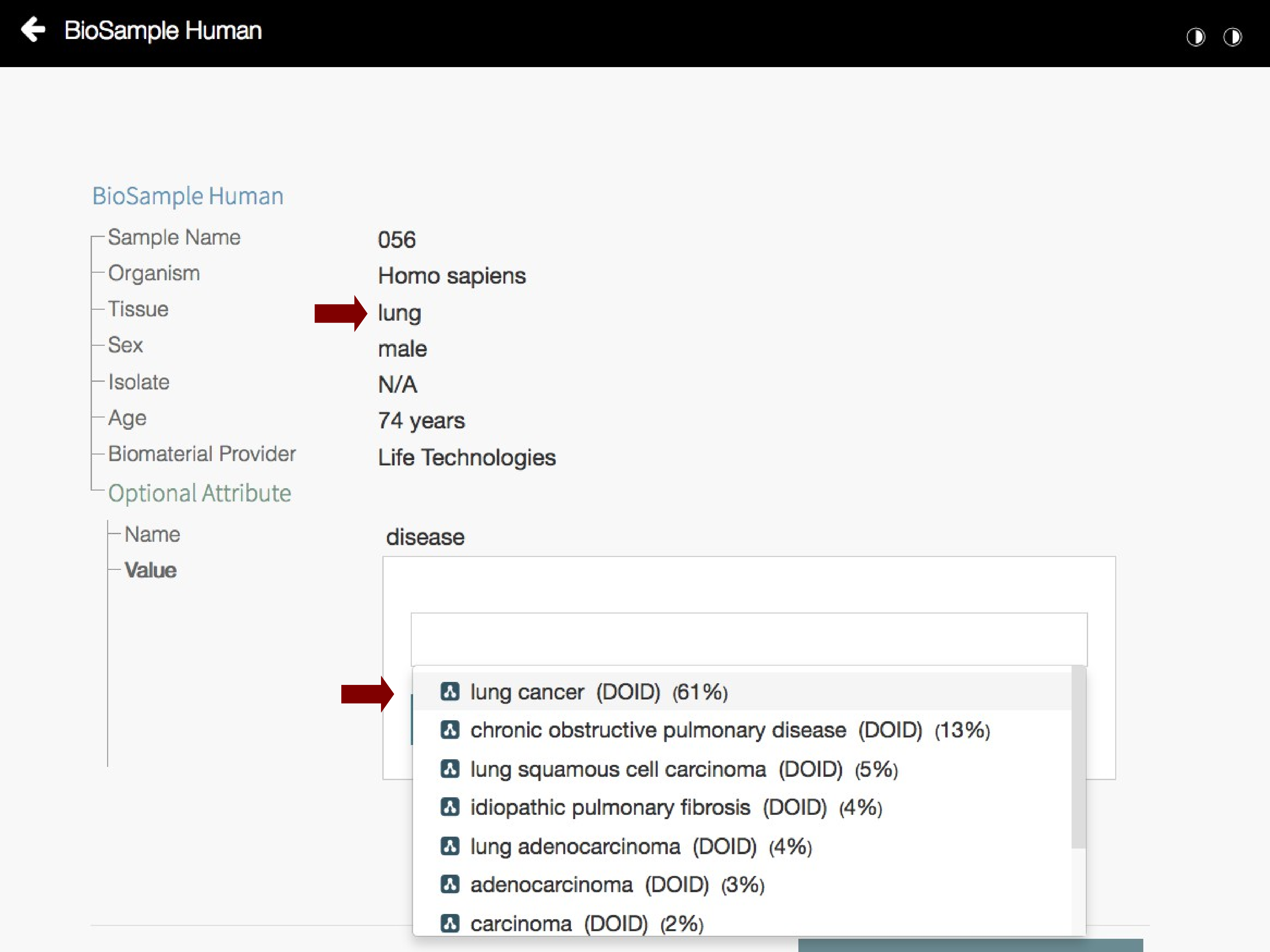


BioSample Human

BioSample Human

Sample Name	056
Organism	Homo sapiens
Tissue	brain
Sex	male
Isolate	N/A
Age	74 years
Biomaterial Provider	Life Technologies
Optional 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%)



BioSample Human

BioSample Human

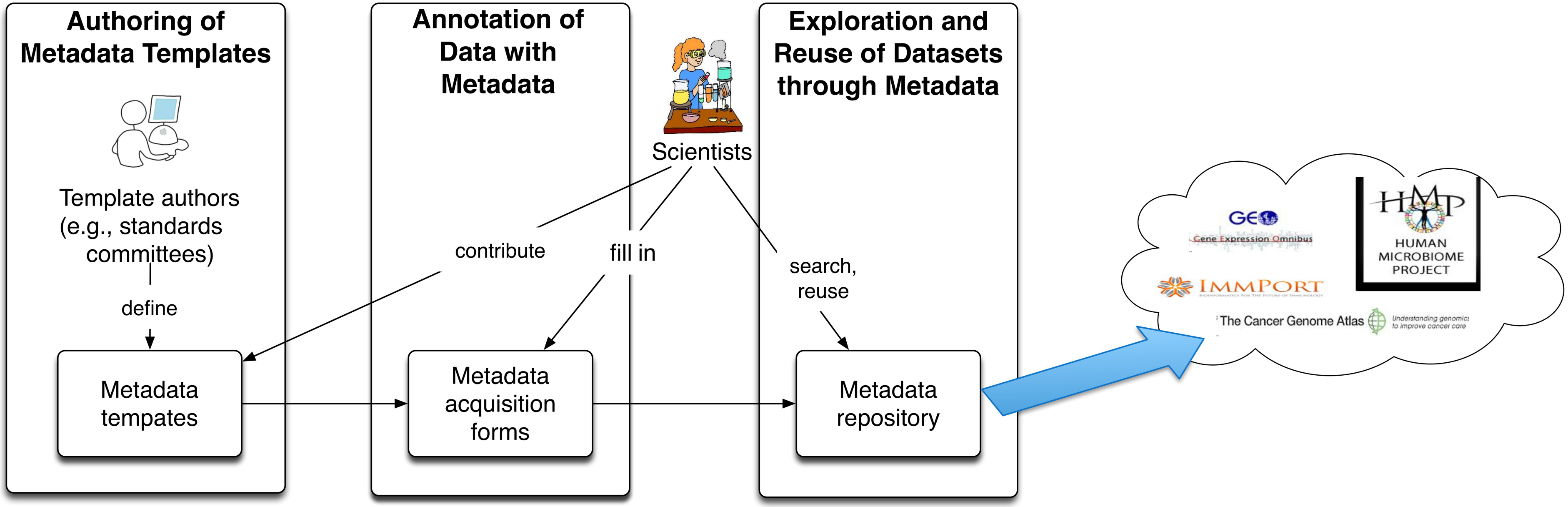
Sample Name	056
Organism	Homo sapiens
Tissue	lung
Sex	male
Isolate	N/A
Age	74 years
Biomaterial Provider	Life Technologies

Optional Attribute

Name	disease
Value	

- lung cancer (DOID) (61%)
- chronic obstructive pulmonary disease (DOID) (13%)
- lung squamous cell carcinoma (DOID) (5%)
- idiopathic pulmonary fibrosis (DOID) (4%)
- lung adenocarcinoma (DOID) (4%)
- adenocarcinoma (DOID) (3%)
- carcinoma (DOID) (2%)

The CEDAR Approach to Metadata



Authoring of Metadata Templates

Template authors (e.g., standards committees)

define

Metadata tempates

Annotation of Data with Metadata

Scientists

contribute

fill in

Metadata acquisition forms

Exploration and Reuse of Datasets through Metadata

search, reuse

Metadata repository

Authoring metadata using domain-specifc ontologies and templates …

- Aids dataset
  - Discovery
  - Exploration
  - Integration
  - Secondary use
- Aids communication of scientific results
- Aids knowledge management within research organizations
- Is making data more FAIR

http://metadatacenter.org

CEDAR

CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL