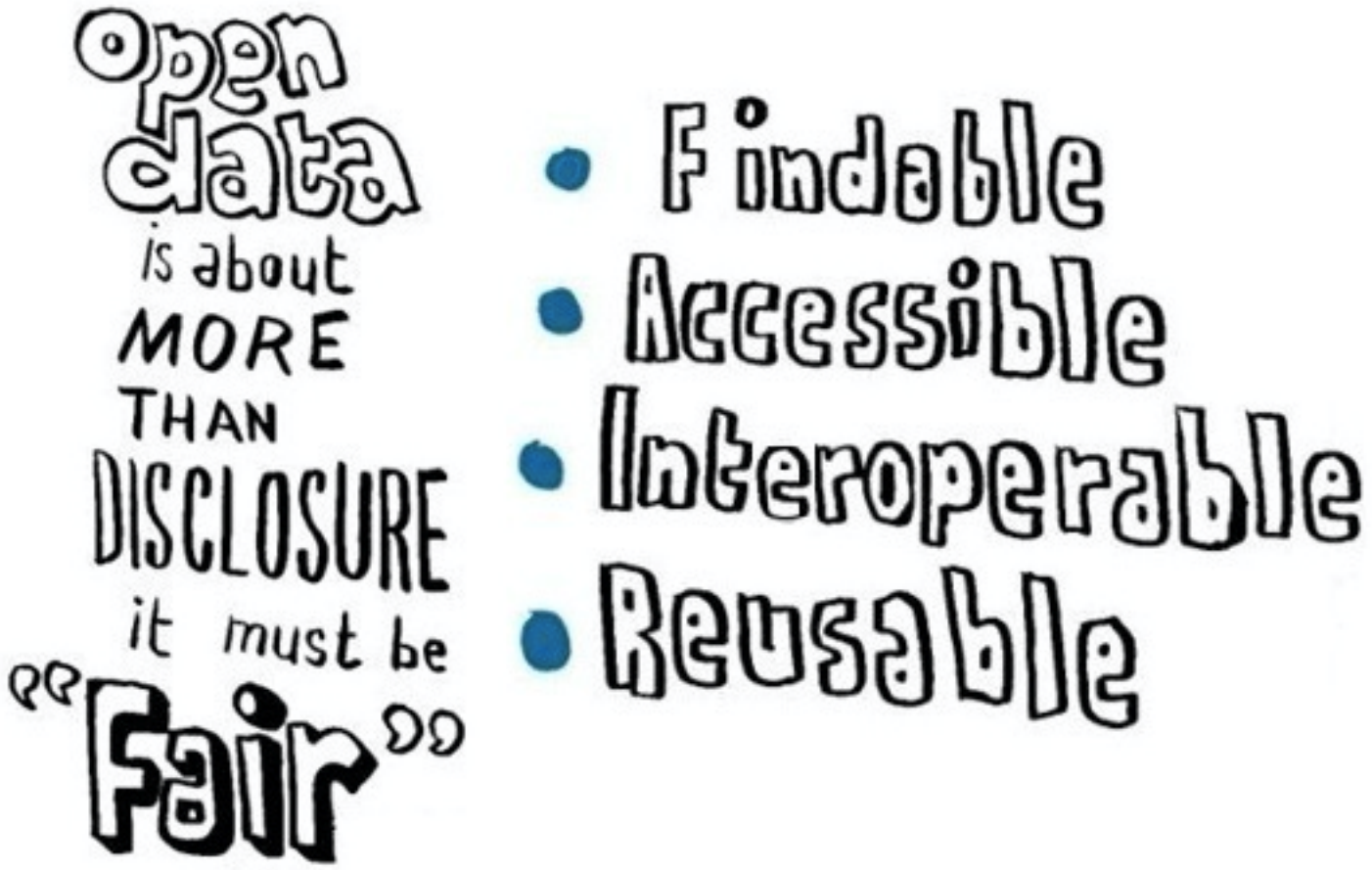
Knowledge Technology to Accelerate Open Science in Addressing the COVID-19 Pandemic

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

Failure to use standard terms makes datasets often impossible to search

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

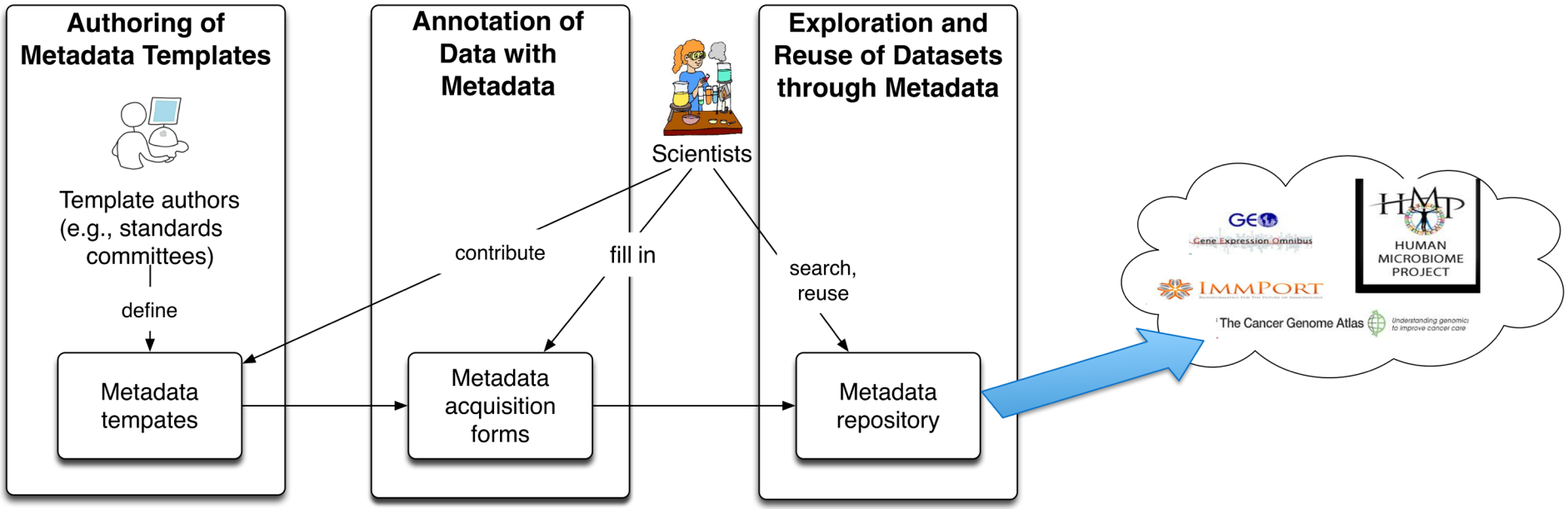
Metadata from NCBI *BioSample* are Awful!

- 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 standard vocabularies do not actually do so
  - *presumed normal, wild_type*

Why does COVID-19 research need better metadata?

- To find the data from relevant studies
- To integrate new results with those already available
- To re-explore existing data sets to make new discoveries
- To verify published claims

The CEDAR Approach to Metadata



Search

All / Users / Mark A. Musen

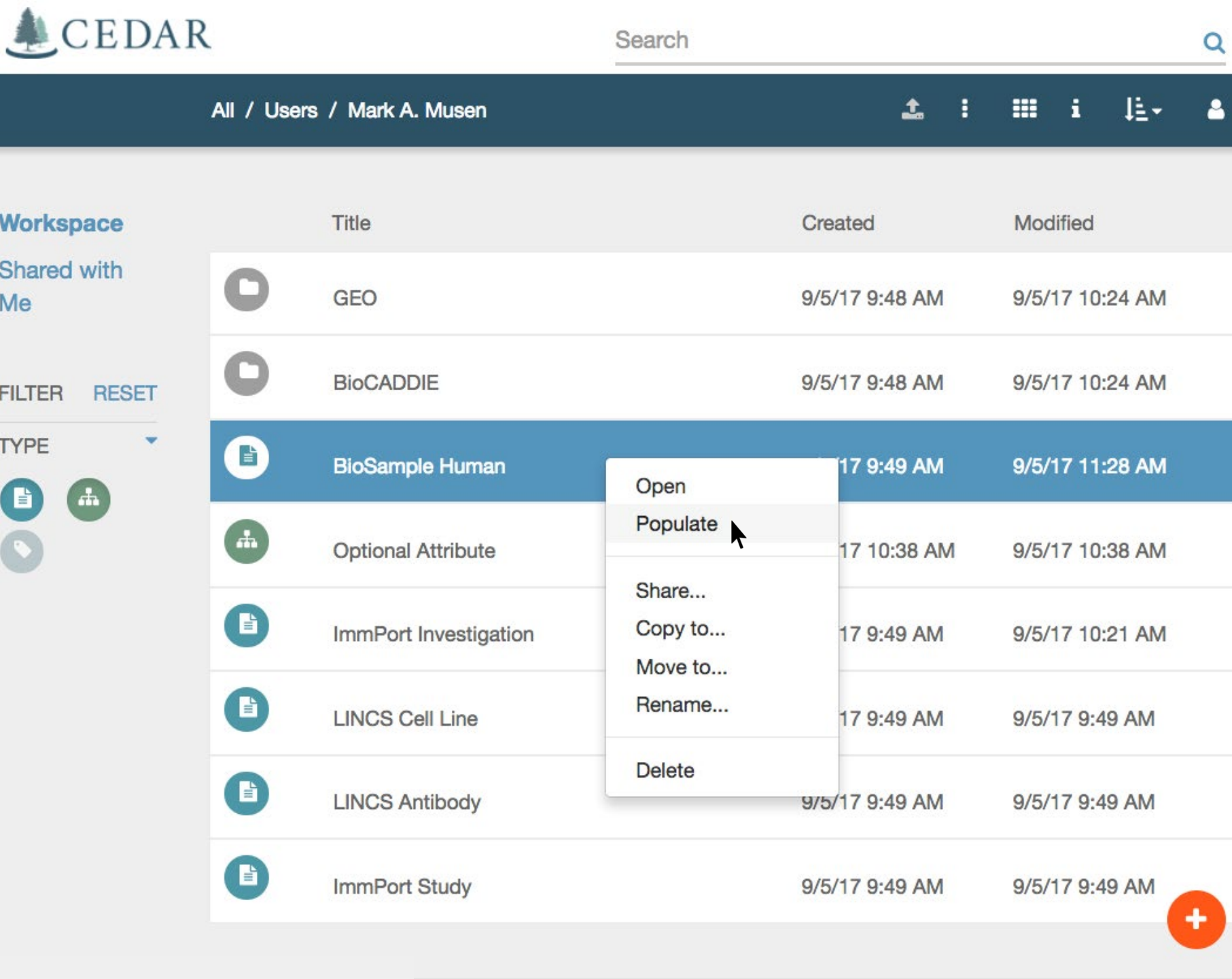
Workspace

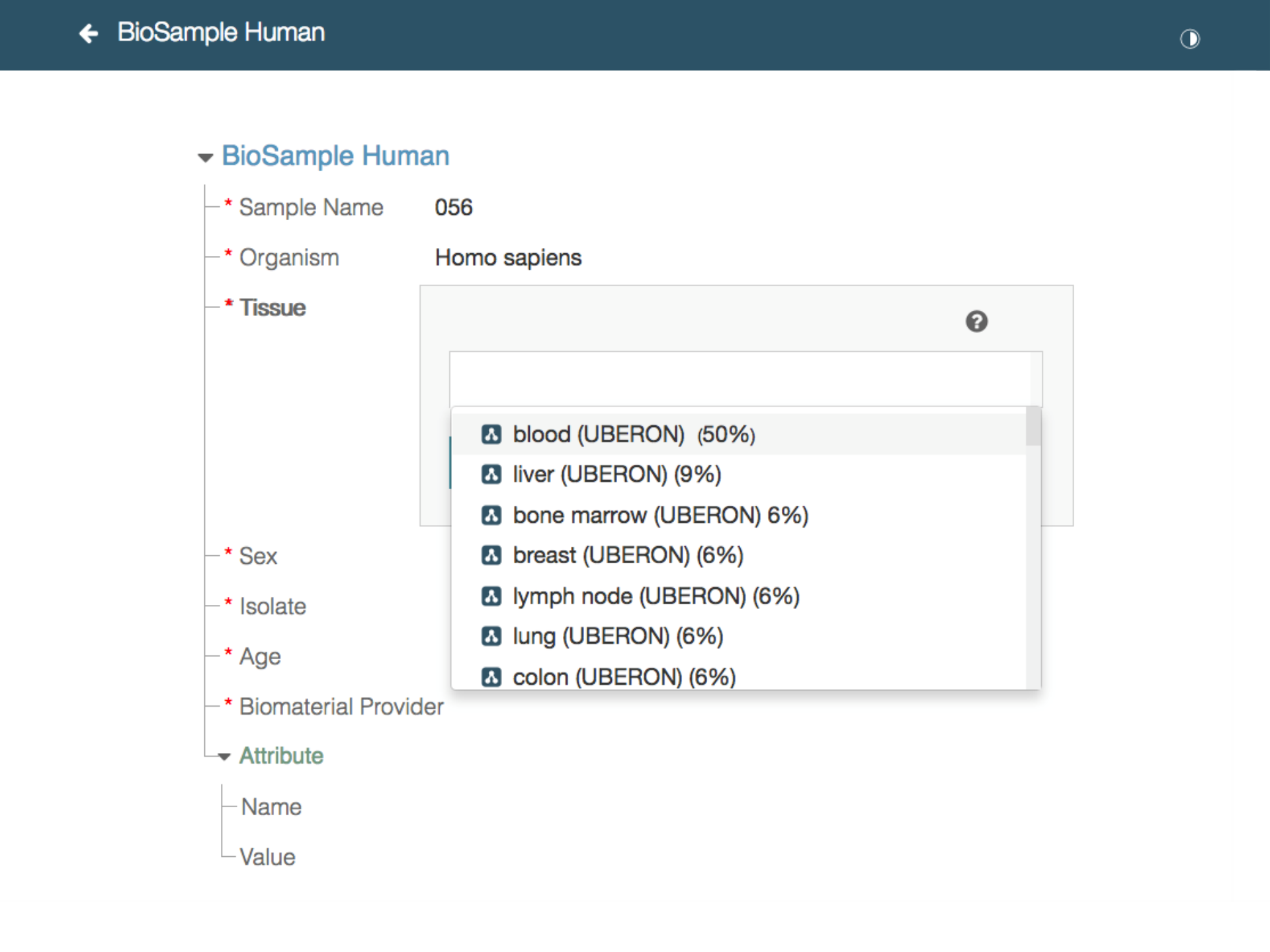
Shared with Me

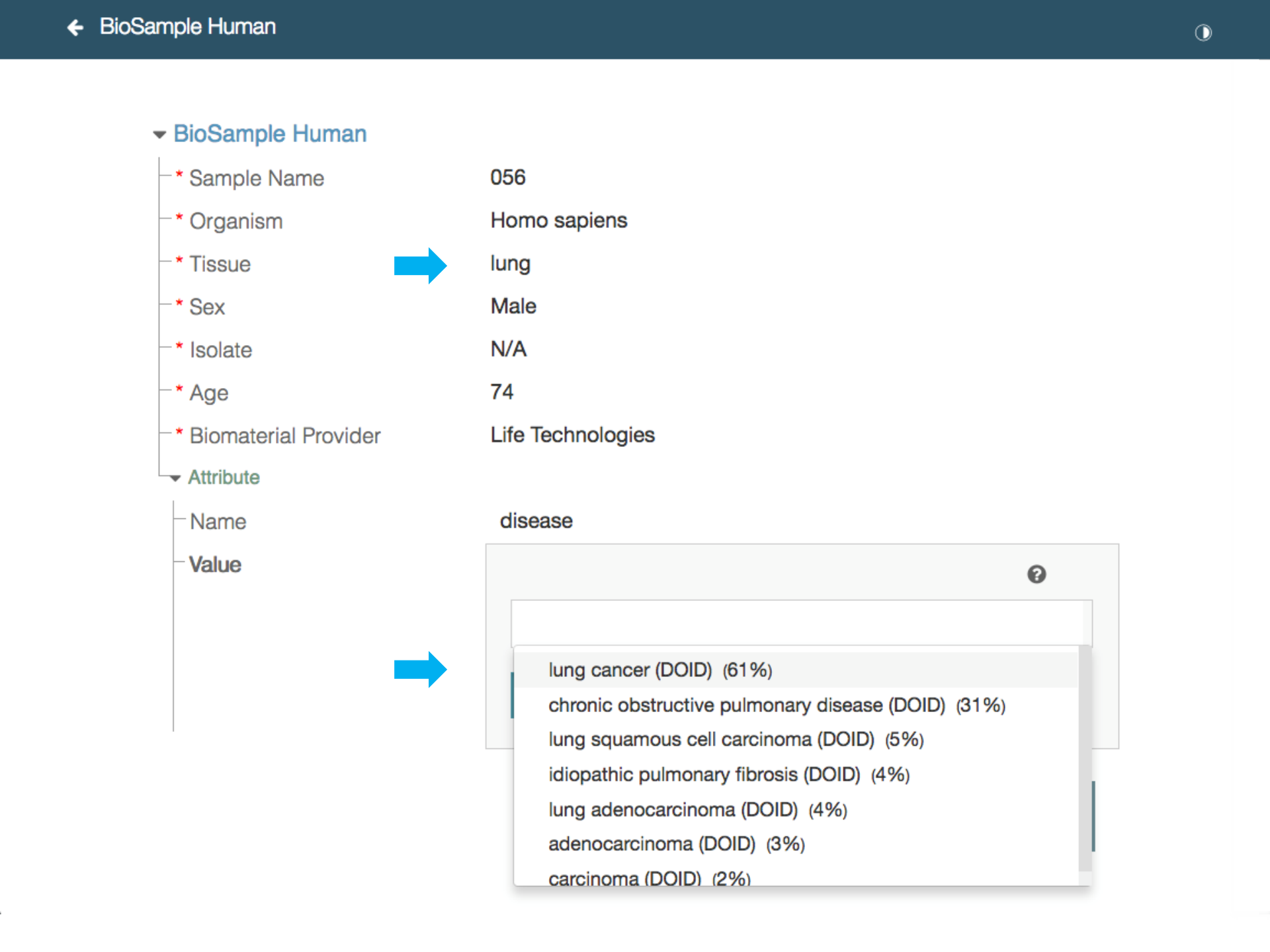
FILTER RESET

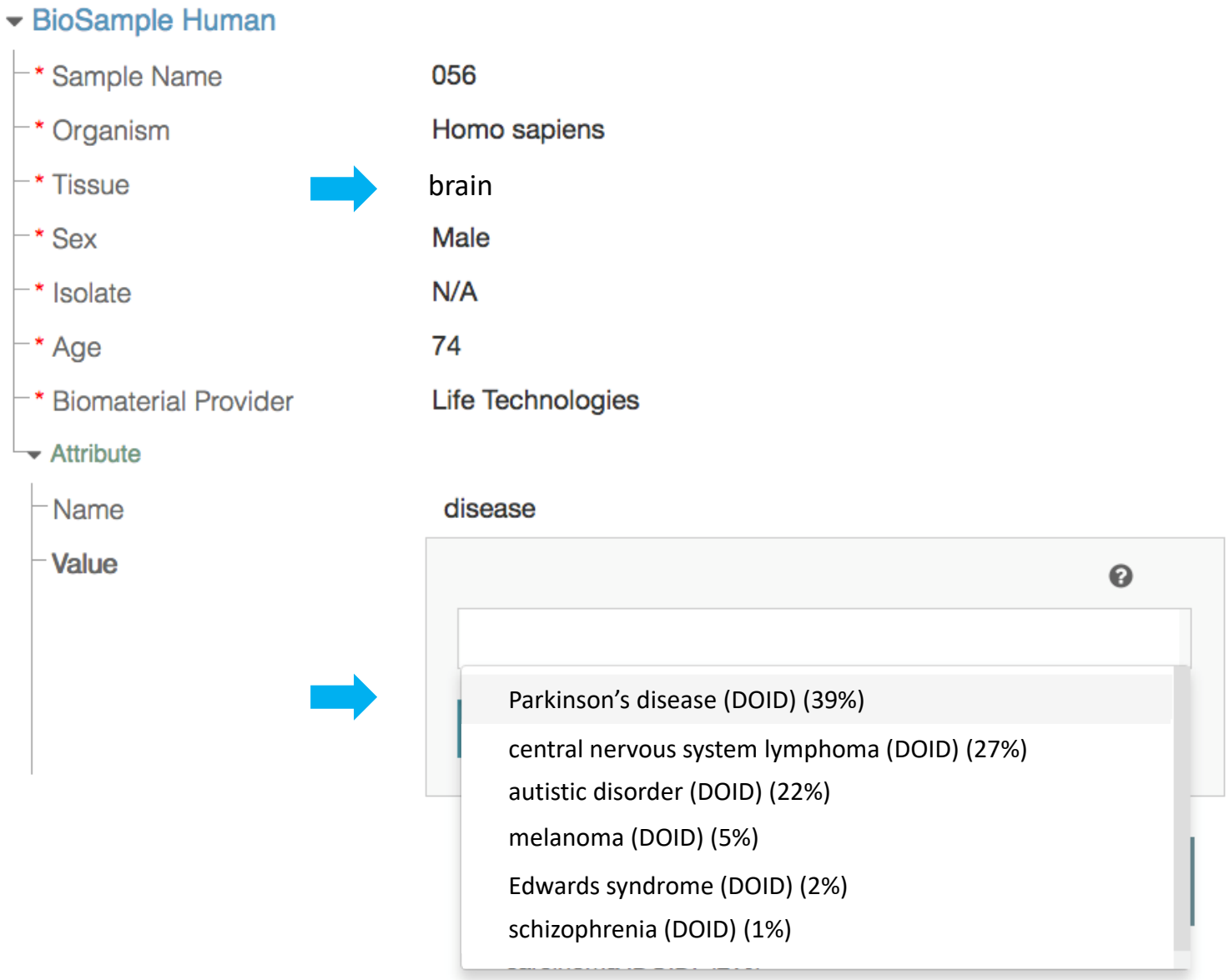
TYPE

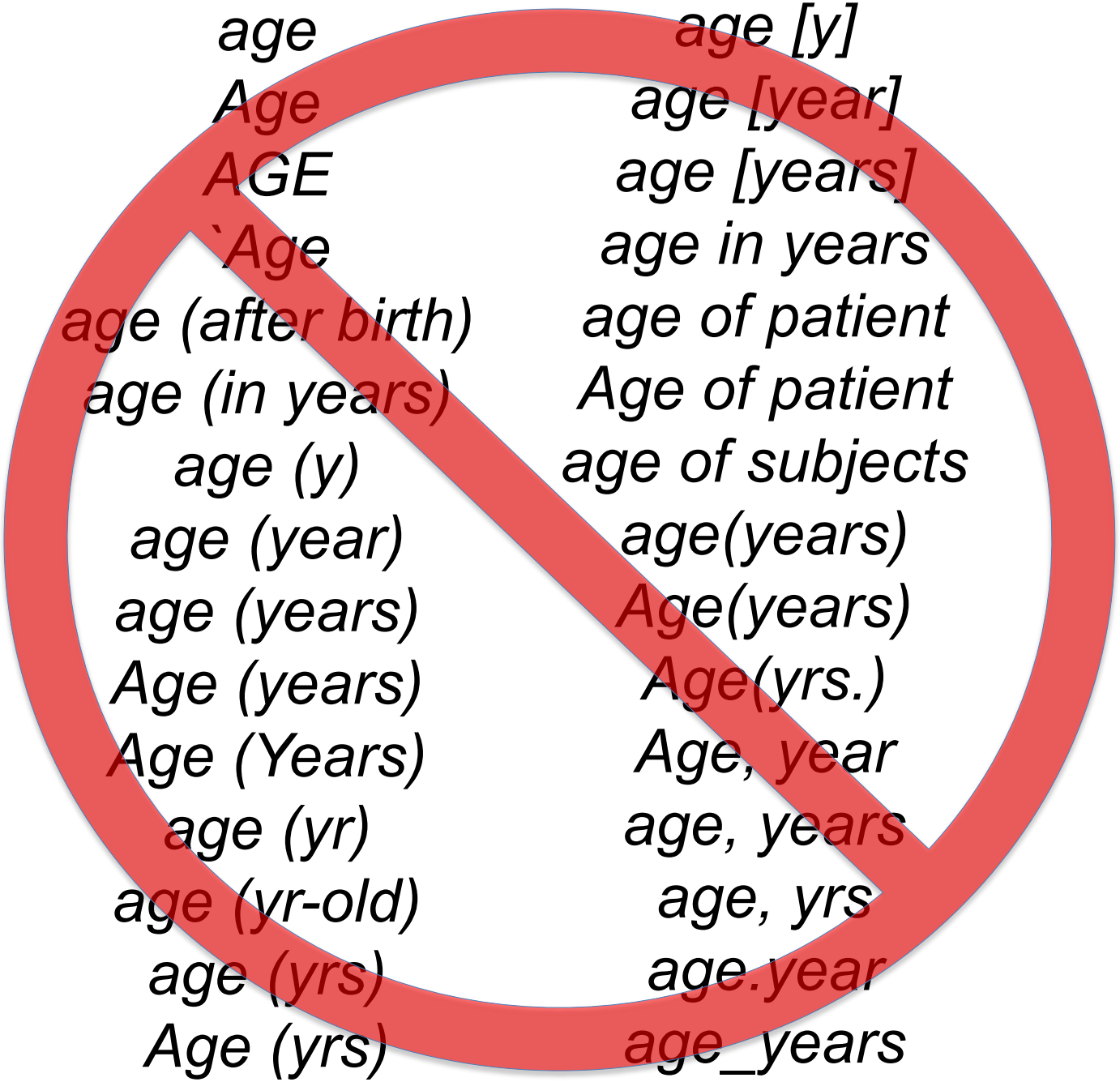
	Title	Created	Modified
	GEO	9/5/17 9:48 AM	9/5/17 10:24 AM
	BioCADDIE	9/5/17 9:48 AM	9/5/17 10:24 AM
	BioSample Human	9/5/17 9:49 AM	9/5/17 11:28 AM
	Optional Attribute	9/5/17 10:38 AM	9/5/17 10:38 AM
	ImmPort Investigation	9/5/17 9:49 AM	9/5/17 10:21 AM
	LINCS Cell Line	9/5/17 9:49 AM	9/5/17 9:49 AM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


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

Virus Outbreak Data Network (VODAN)

*The GO FAIR Office is currently experiencing significant increase in support requests due to high interest in the VODAN Implementation Network. While we are excited to see such enthusiasm in the data community, the office staff capacity is limited. If you need to contact us, please email us at **office@go-fair.org** and we will do our best to respond in a timely manner. Thank you for understanding. Additional contacts for projects and sub-teams will be announced soon.*

Active GO FAIR Implementation Network

The spread of the virus causing the COVID-19 outbreak is far from over. During this epidemic and in earlier occasions, we have seen severely suboptimal data management and data reuse. Moreover, access to the immensely valuable data of past and current epidemics is not always equally accessible for different affected populations and countries. For instance, the data from the past Ebola epidemics are very difficult to find, to access, and if accessible, they are not interoperable, *let alone reusable*. In the case of Ebola this is even more harrowing and ironic as the data are *least available* to the population that were *most affected* by the disaster. Under the urgent need to harness machine-learning and future AI approaches to discover meaningful patterns in epidemic outbreaks, we need to do better and ensure that data are FAIR (in this sense also meaning **F**ederated, **AI**–**R**eady).

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

CEDAR
CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL

http://metadatacenter.org

CEDAR

CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL