

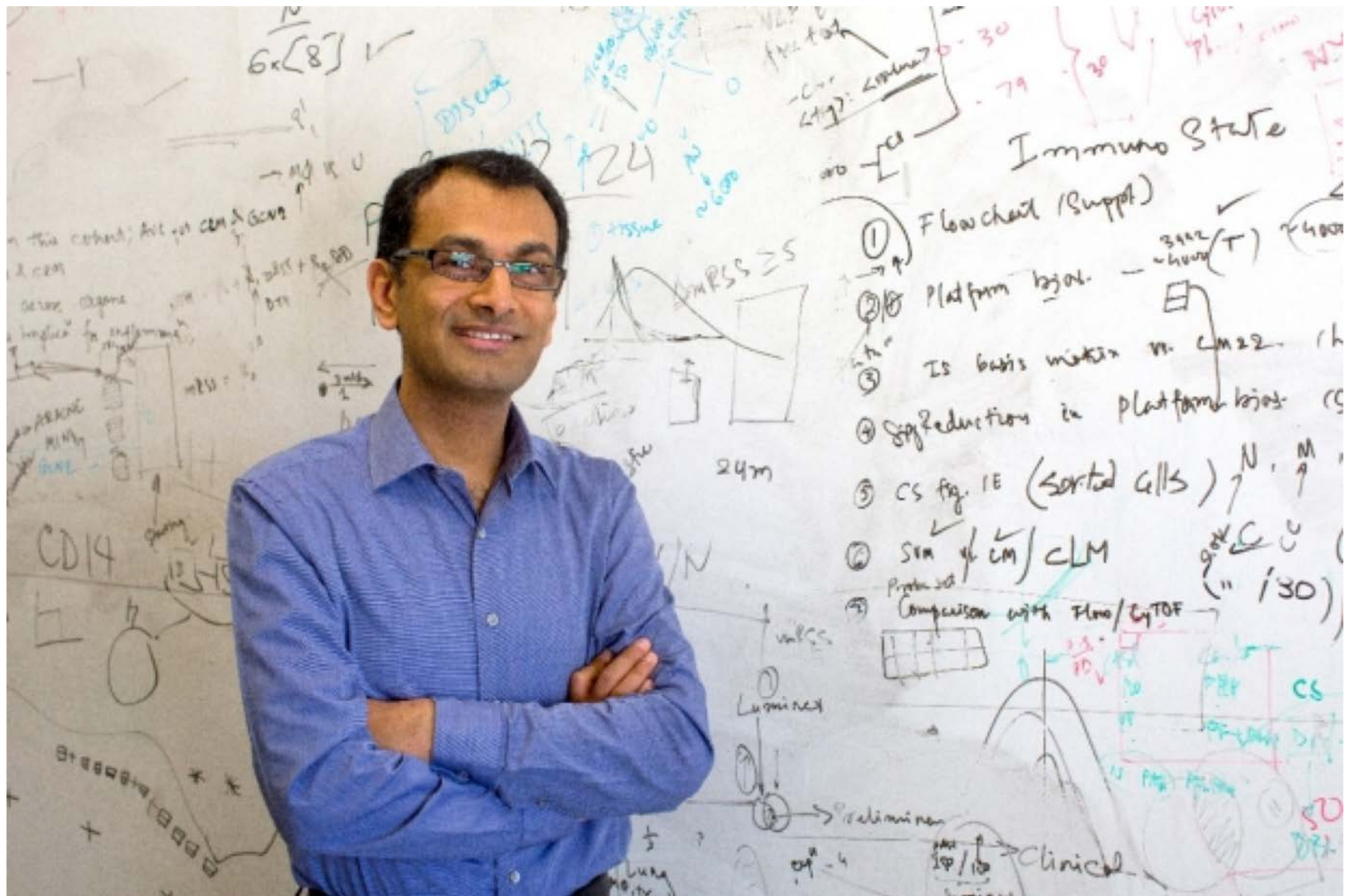
Of Course Data Stewards Benefit from Training. But What They Really Need is Better *Technology*

Mark A. Musen, M.D., Ph.D
and the CEDAR Team
musen@Stanford.EDU

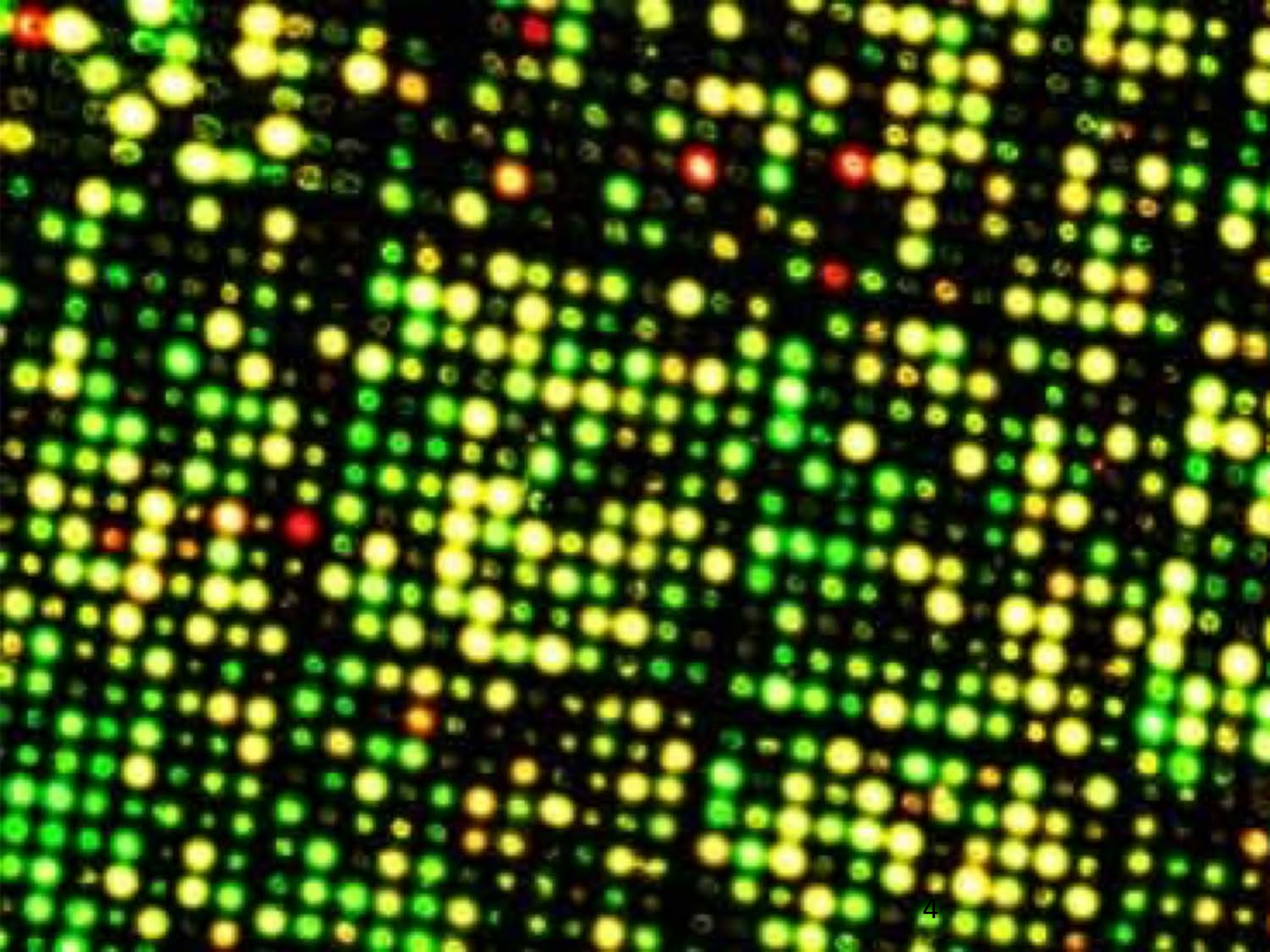


CENTER FOR EXPANDED DATA ANNOTATION
AND RETRIEVAL

A Story



Purvesh Khatri, Ph.D. A self-professed “data parasite”



Gene Expression Omnibus (GEO)

www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS6000

NCBI

CURATED
DATASET
BROWSER

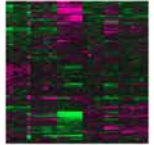
GEO
Gene Expression Omnibus

Search for

DataSet Record GDS6000: [Expression Profiles](#) [Data Analysis Tools](#) [Sample Subsets](#)

Title:	High-fat diet effect on brown adipose tissue development		
Summary:	Analysis of interscapular brown adipose tissues from mice fed a high fat diet for up to 24 weeks. Results compared to those from epididymal white adipose tissues (GDS6247) in order to provide insight into the effect of high-fat diets on the development of brown and white adipose tissues.		
Organism:	<i>Mus musculus</i>		
Platform:	GPL6887: Illumina MouseWG-6 v2.0 expression beadchip		
Citation:	Kim HS, Ryoo ZY, Choi SU, Lee S. Gene expression profiles reveal effect of a high-fat diet on the development of white and brown adipose tissues. <i>Gene</i> 2015 Jul 1;565(1):15-21. PMID: 25895476		
Reference Series:	GSE64718	Sample count:	33
Value type:	transformed count	Series published:	2015/01/08

Cluster Analysis



Download

- DataSet full SOFT file
- DataSet SOFT file
- Series family SOFT file
- Series family MINIML file
- Annotation SOFT file

Data Analysis Tools

Find genes [?](#)

Compare 2 sets of samples

Cluster heatmaps

Experiment design and value distribution

Find gene name or symbol:

Find genes that are up/down for this condition(s): ☒ time ☒ protocol

Khatri has reused public datasets in GEO to identify genomic signatures ...

- For incipient sepsis
- For active tuberculosis
- For distinguishing viral from bacterial respiratory infection
- For rejection of organ transplants

... and he has never touched a pipette!

But the online datasets that Khatri studies
are a mess!

- Investigators view their work as publishing papers, not leaving a legacy of reusable data
- Funding agencies may require data sharing, but they do not explicitly pay for it
- Creating the metadata to describe data sets is unbearably hard
- Ensuring that metadata are standardized and searchable is just about impossible

Use this template for 3' or whole Gene expression studies when summarization probe set data will be provided as **CHP files**.
 # 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**
 # Incomplete submissions will be returned. Click the **Metadata Example** tab below to view a completed worksheet
 # A complete submission will consist of: (1) a completed metadata worksheet, (2) the CHP files, and (3) the original CEL files.
 # **Field names** (in blue on this page) should not be edited. Hover over cells containing **field names** to view field content guidelines or,
 # [CLICK HERE](#) for Field Content Guidelines Web page.

SERIES

This section describes the overall

title

summary

summary

overall design

contributor

contributor

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

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

SAMPLES

The **Sample names** in the first column are arbitrary but they must match the column headers of the Matrix table (see next worksheet).

Sample name

title

CHP file

source name

organism

characteristics: tag

SAMPLE 1

SAMPLE 2

SAMPLE 3

SAMPLE 4

SAMPLE 5

SAMPLE 6

SAMPLE 7

SAMPLE 8

SAMPLE 9

SAMPLE X

**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).

PROTOCOLS

This section includes protocols and fields which are common to all Samples.

Protocols which are applicable to specific Samples or specific channels should be included in additional columns of the **SAMPLES** section instead.

growth protocol

treatment protocol

extract protocol

label protocol

hyb protocol

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

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

An Analysis of Metadata from *BioSample*

- 85% of submissions avoid using a predefined “package” for regularizing metadata
- 73% of “Boolean” metadata values are not actually *true* or *false*
- 26% of “integer” metadata values cannot be parsed into integers
- 68% of metadata entries that are supposed to represent terms from biomedical ontologies do not actually do so.

At a minimum, scientists need

- Open, online access to experimental data sets
- Annotation of online data sets with adequate metadata
- Use of controlled terms in metadata whenever possible
- Technology that can help them curate their data—not training to instill specific skills

Open
data
is about
MORE
THAN
DISCLOSURE
it must be
“Fair”

- Findable
- Accessible
- Interoperable
- Reusable

Open
data
is about
MORE
THAN
DISCLOSURE
it must be
Fair

- Findable
- Accessible
- Interoperable
- Reusable

Requirement #1: Have standard terms to describe what exists in a dataset completely and consistently

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

Search for a class

Q

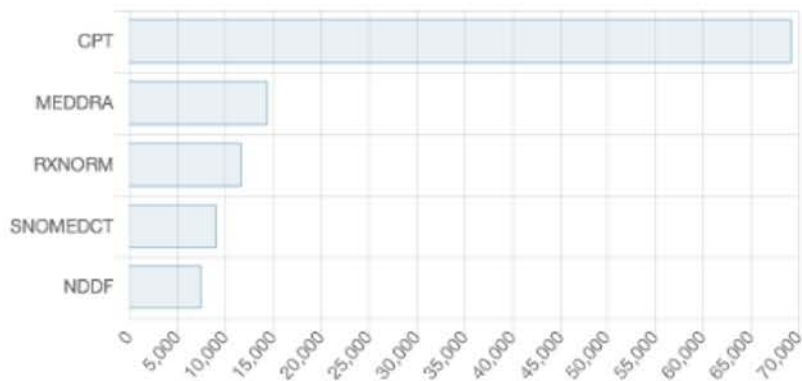
Advanced Search

Find an ontology

Q

Browse Ontologies

Ontology Visits (February 2018)



More

BioPortal Statistics

Ontologies	692
Classes	8,848,090
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

Browse

Browse the library of ontologies [?](#)

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** (21)
- ☐ **Arabidopsis** (2)
- ☐ **Biological Process** (44)
- ☐ **Biomedical Resources** (55)
- ☐ **Cell** (6)

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

115,514

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

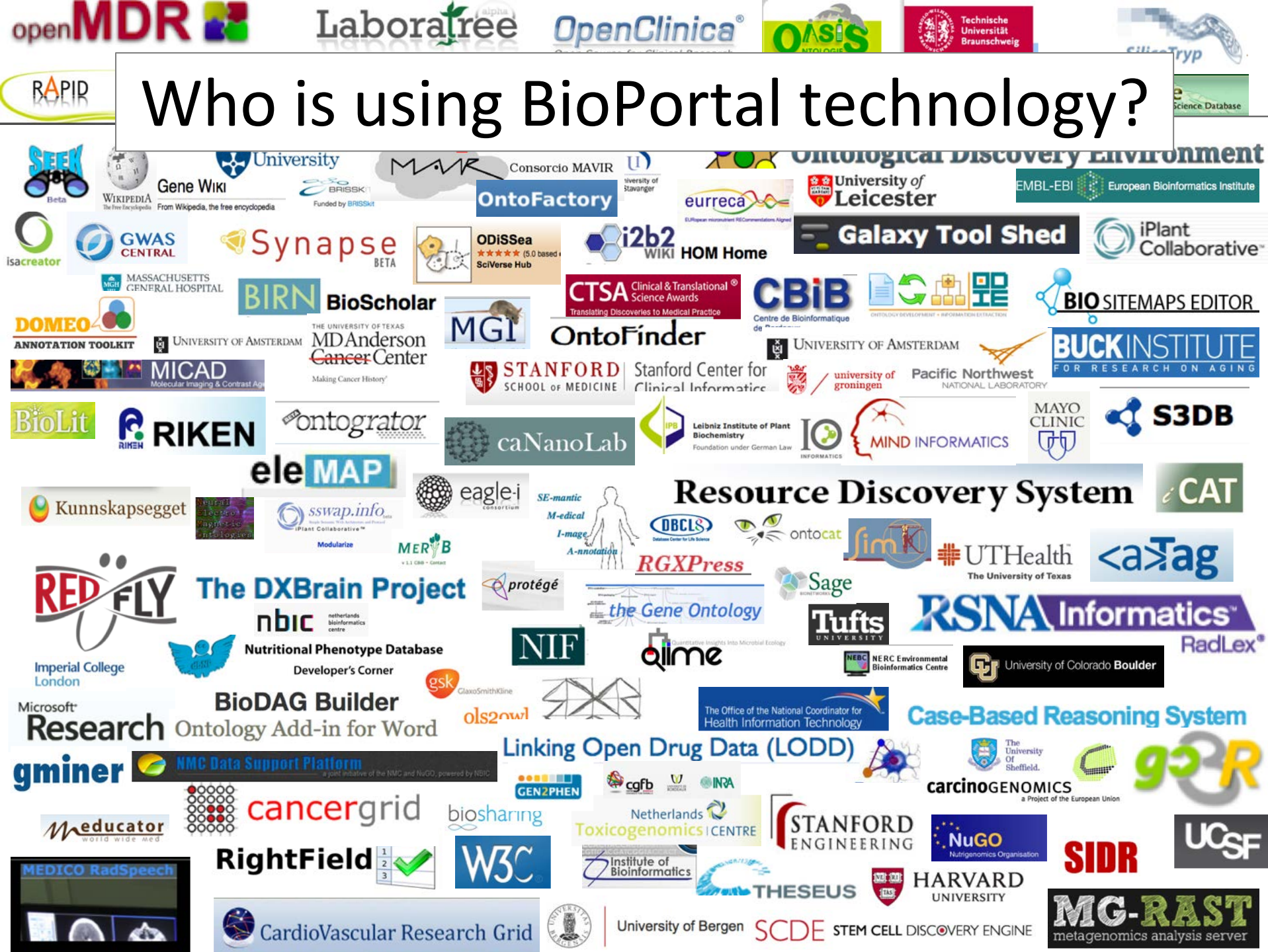
Uploaded: 2/6/17

projects

1

classes

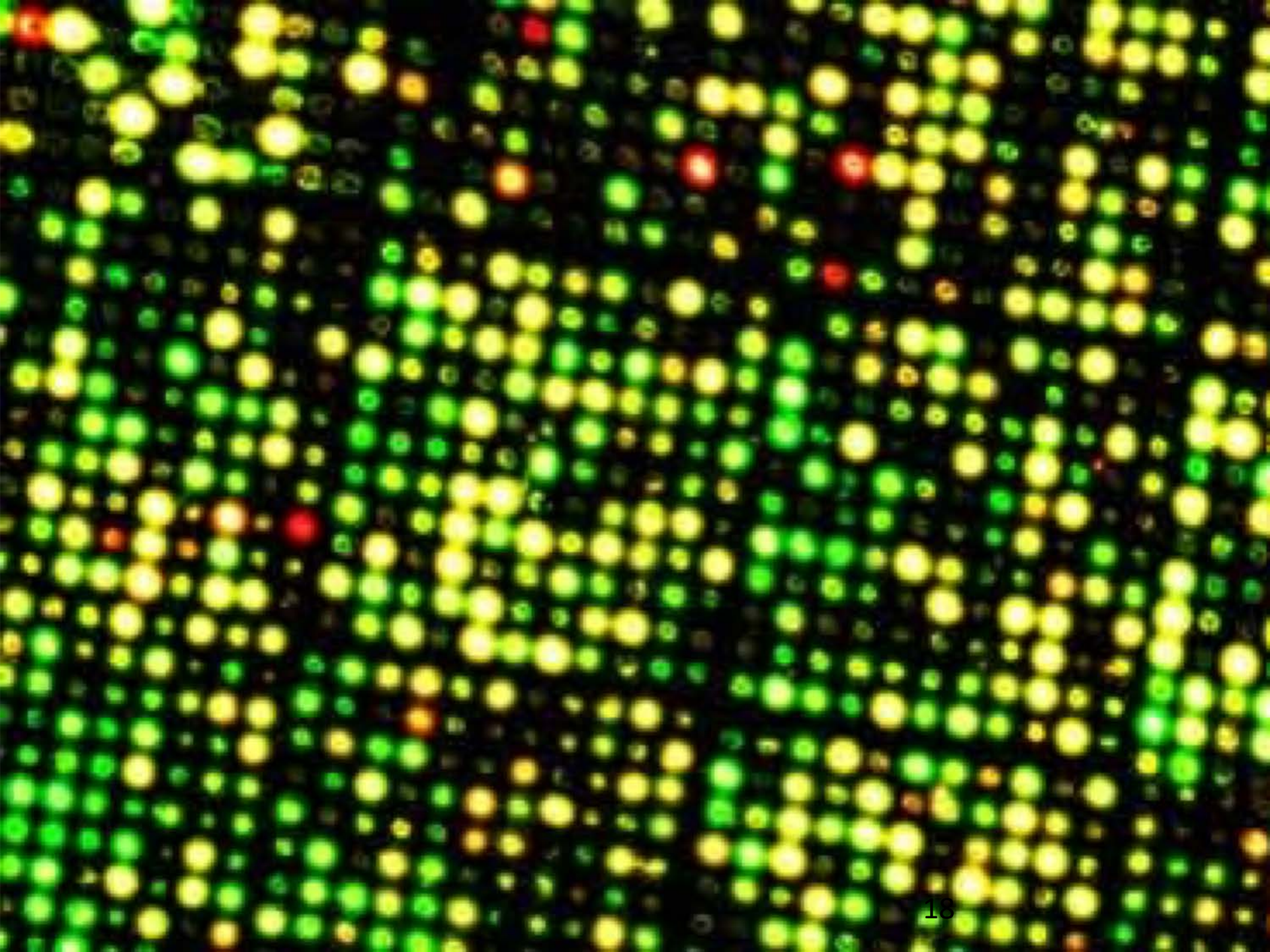
28,111



open
data
is about
MORE
THAN
DISCLOSURE
it must be
Fair

- Findable
- Accessible
- Interoperable
- Reusable

Requirement #2: Describe properties of experiments completely and consistently



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)

Minimal Information Guidelines are not Models

- MIAME and its kin specify only the “kinds of things” that investigators should include in their metadata
- They do not provide a detailed list of standard metadata elements
- They do not provide datatypes for valid metadata entries
- It takes work to convert a *prose checklist* into a computable model

open
data
is about
MORE
THAN
DISCLOSURE
it must be
Fair

- Findable
- Accessible
- Interoperable
- Reusable

Requirement #3: Make it easy to describe experiments completely and consistently

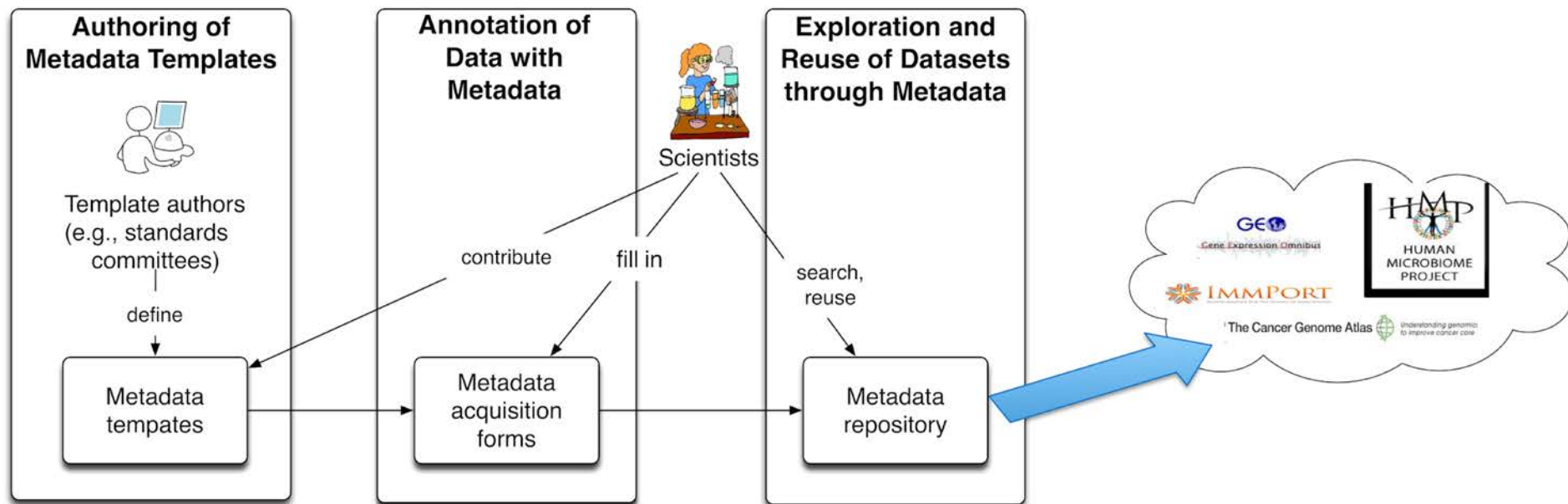
<http://metadatacenter.org>



CEDAR

CENTER FOR EXPANDED DATA ANNOTATION
AND RETRIEVAL

The CEDAR Approach to Metadata




The CEDAR Workbench provides

- Mechanisms
 - To author metadata templates that reflect community standards
 - To fill out templates to encode experimental metadata
- A repository of metadata from which we can
 - Learn metadata patterns
 - Guide predictive entry of new metadata
- Links to BioPortal to ensure that metadata are encoded using appropriate ontology terms









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











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

Open

Populate

Share...

Copy to...

Move to...

Rename...

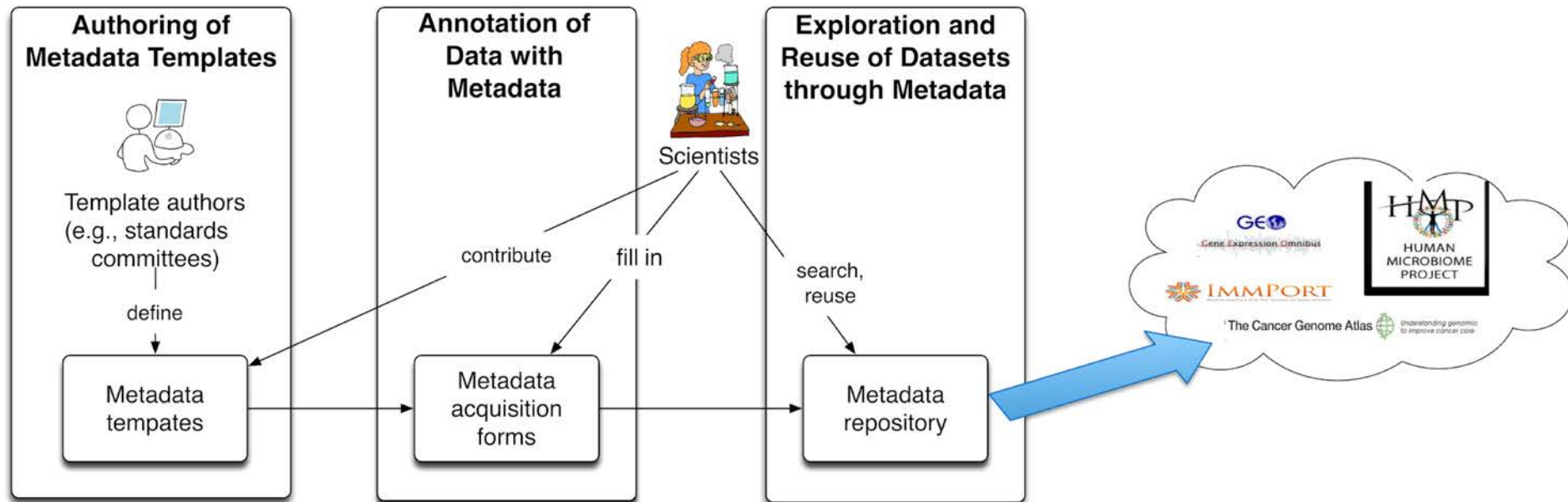
Delete



▼ 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)	
Name	disease
Value	dermatitis
▼ Attribute (2)	
Name	description
Value	Cell line was cultured until the 5th passage
▼ Attribute (3)	
Name	treatment
Value	350mg brodalumab

The CEDAR Workbench



[a](#)
[1](#)
[31](#)
<#>
[...](#)

Find terms in BioPortal or [Create New Terms](#) to constrain the values of the 'Tissue' field

[Start Over](#)

Search in BioPortal

Tissue



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 Metencephalon
 - ⊕ Macula
 - Intervillous Pockets
 - Purkinje Cell Layer Corpus
 - Mossy Fiber
 - Pars Basilaris
 - Dermis Of Feather Follicle
 - Upper Oral Valve
 - ⊕ Anlage
 - Anterior Lateral Plate Mesoderm
 - Molecular Layer Valvula Cerebri

TERM DETAILS		ONTOLOGY DETAILS
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

BRANCH

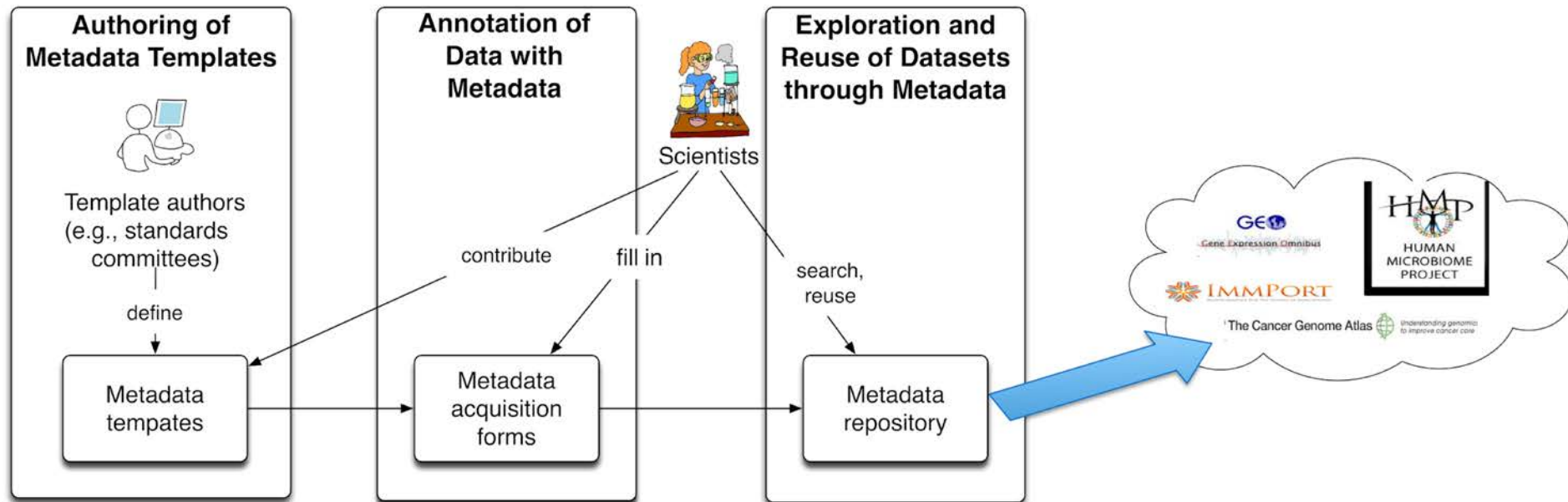
ONTOLOGY

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

Click to add all the descendants of the selected term

ADD

The CEDAR Workbench



▼ BioSample Human

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

CANCEL

VALIDATE

SAVE

* Sample Name 056

—* Organism Homo sapiens

— * Tissue

— * Sex

- ★ Isolate

★ Age

—* Biomaterial Provider

Attribute

Name	Age	Gender	Address	City	State	Zip
John Doe	35	Male	123 Main St	New York	NY	10001
Jane Smith	28	Female	456 Elm St	Los Angeles	CA	90001
Bob Johnson	42	Male	789 Oak St	Chicago	IL	60601
Alice Brown	31	Female	101 Pine St	San Francisco	CA	94101
Charlie Davis	25	Male	202 Maple St	Seattle	WA	98101
Eve White	38	Female	303 Birch St	Portland	OR	97201
Frank Green	45	Male	404 Cedar St	Denver	CO	80201
Grace Black	29	Female	505 Spruce St	Phoenix	AZ	85001
Henry Blue	33	Male	606 Willow St	San Diego	CA	92101
Ivy Red	27	Female	707 Ash St	San Jose	CA	95101
Jack Yellow	40	Male	808 Hickory St	San Antonio	TX	78201
Karen Purple	36	Female	909 Cypress St	San Luis Obispo	CA	93401
Leo Grey	22	Male	1010 Dogwood St	San Francisco	CA	94101
Mia Silver	39	Female	1111 Magnolia St	San Francisco	CA	94101
Noah Gold	41	Male	1212 Sycamore St	San Francisco	CA	94101
Olivia Bronze	34	Female	1313 Tulip St	San Francisco	CA	94101
Peter Copper	26	Male	1414 Violet St	San Francisco	CA	94101
Quinn Iron	43	Female	1515 Zinnia St	San Francisco	CA	94101
Rachel Steel	30	Male	1616 Aster St	San Francisco	CA	94101
Sam Bronze	24	Female	1717 Begonia St	San Francisco	CA	94101
Tina Silver	44	Male	1818 Camellia St	San Francisco	CA	94101
Uma Gold	21	Female	1919 Dandelion St	San Francisco	CA	94101
Victor Copper	37	Male	2020 Foxglove St	San Francisco	CA	94101
Wendy Iron	23	Female	2121 Geranium St	San Francisco	CA	94101
Xavier Steel	46	Male	2222 Hibiscus St	San Francisco	CA	94101
Yara Bronze	32	Female	2323 Iris St	San Francisco	CA	94101
Zoe Silver	29	Male	2424 Jasmine St	San Francisco	CA	94101

Value

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

▼ 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%)

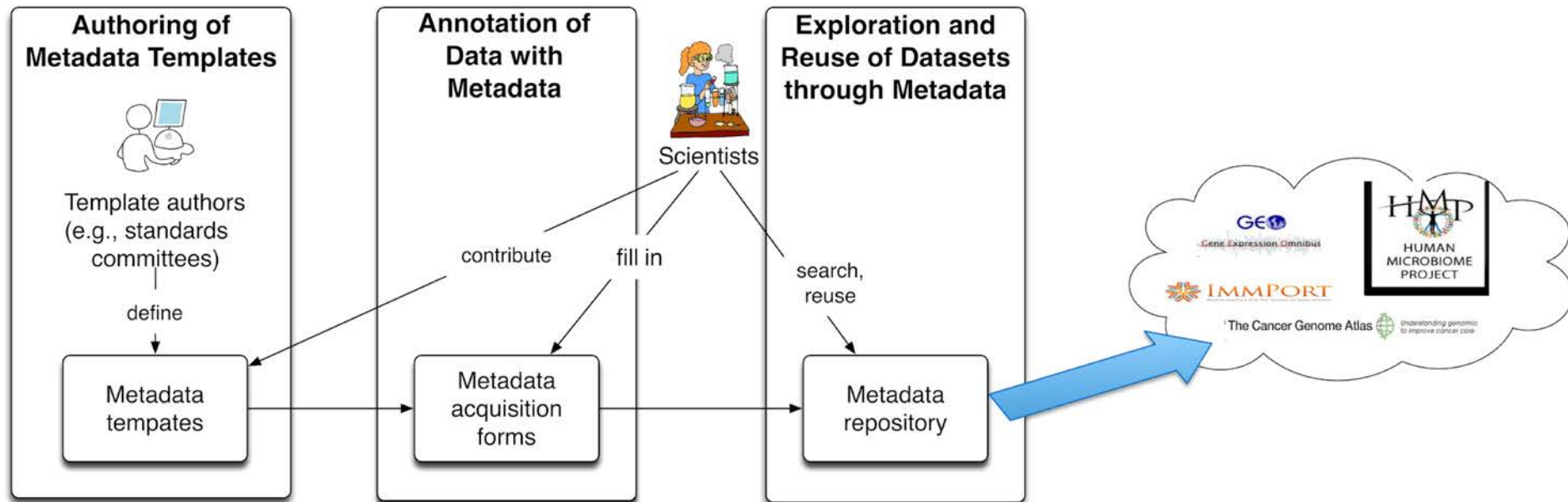
autistic disorder (DOID) (22%)

melanoma (DOID) (5%)

Edwards syndrome (DOID) (2%)

schizophrenia (DOID) (1%)

The CEDAR Workbench



Workspace

Shared with
Me

FILTER RESET

TYPE ▾



Submit to Repository



Repository

Upload

Metadata 

search workspace

Select files to upload.

SELECT

BioSample Human metadata.json



SAMN02911274.xlsx



PAUSE

RESUME

CLEAR

CLOSE

SUBMIT



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





NIH LINCS
PROGRAM



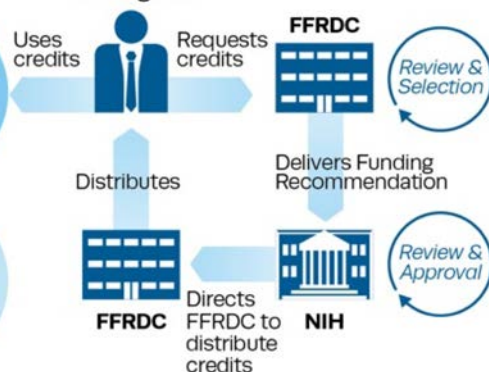
AIRR Community

Commons Credits Pilot

The Commons



Investigator



Biomedical Researchers Clearly are Ahead of the Pack

- They have been creating standard ontologies for years
- They are proposing increasing numbers of “minimal information models” that are ripe for conversion to formal metadata templates
- They are beginning to turn to technology such as CEDAR to enhance their online datasets



The CEDAR Approach is Generalizable to Other Areas of Science

- The building blocks needed for developing high-quality metadata are clear:
 - Standard ontologies
 - Stanford templates
- Nothing in CEDAR is hardwired to the life-sciences domain
- Most important: Operators are standing by



<http://metadatacenter.org>



CEDAR

CENTER FOR EXPANDED DATA ANNOTATION
AND RETRIEVAL