

There's No Easy Way Out

Making Data FAIR Requires Better Metadata. Information Technology Can Help.

Mark A. Musen, M.D., Ph.D.
Stanford Center for Biomedical Informatics Research
Stanford University
musen@stanford.edu



The
Economist

OCTOBER 19TH-25TH 2013

economist.com

Washington's lawyer surplus
How to do a nuclear deal with Iran
Investment tips from Nobel economists
Junk bonds are back
The meaning of Sachin Tendulkar

HOW SCIENCE GOES WRONG.

99
Einsteinium

THE PARASITE AWARDS

Celebrating rigorous secondary data analysis



THE "PARASITES"

PSB Awards for rigorous secondary data analysis. A companion to the [Research Symbiont Awards](#).

Amended: Addendum

OPEN

SUBJECT CATEGORIES

- » Research data
- » Publication characteristics

Received: 10 December 2015

Accepted: 12 February 2016

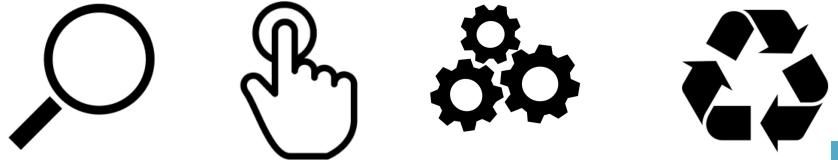
Published: 15 March 2016

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 measurable 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.

Findable Accessible Interoperable Reusable



Why we chose Fair Data



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

The FAIR Evaluator - automated testing of Web resources for their compliance

FAIR Metrics Evaluation

HIDE DETAILS

Registry URL

<http://smart-api.info/registry?q=ad830426bed193d36838091ef5d14407>

SmartAPI ID

ad830426bed193d36838091ef5d14407

[Copy](#)

Metadata URL

https://w3id.org/FAIR_Evalu...

Version

V 0.3.0

Contact

 Mark Wilkinson



FAIRness of LINCS Datasets and Tools project

FAIR evaluation of the LINCS NIH Program tools and datasets

Tags: 📈 DCPPC

URL(s):

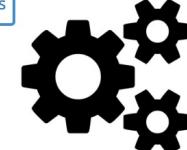
<http://lincsproject.org/>



[View Analytics](#)

[View Assessments](#)

Associated Digital Objects (81)



L1000CDS2

tool

An ultra-fast LINCS L1000 Characteristic Direction signature search engine

✖ test

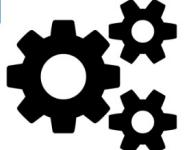


iLINCS

tool

An integrative web platform for analysis of LINCS data and signatures

✖ test

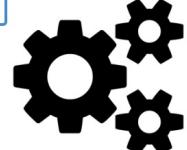


Query App

tool

Connections to user-defined signatures

✖ test



Drug-Pathway Browser

tool

Interactive map of key signal transduction pathways and drug-target data

✖ test

F-UJI 1.0.0 OAS3

/fuji/api/v1/openapi.json

A Service for Evaluating Research Data Objects Based on [FAIRsFAIR Metrics](#).

This work was supported by the [FAIRsFAIR](#) project (H2020-INFRAEOSC-2018-2020 Grant Agreement 831558).

[Contact the developer](#)

[MIT License](#)

[Find out more about F-UJI](#)

Servers

/fuji/api/v1 ▾

[Authorize](#) 

FAIR object FAIRness assessment of a data object

POST /evaluate



FAIR metric FAIRsFAIR assessment metrics

GET /metrics Return all metrics and their definitions



Systems to evaluate the FAIRness of existing datasets have struggled to find 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

Many FAIR principles depend on community standards for metadata and 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

Data are not FAIR because metadata 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 good 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 title

11 summary

12 summary

13 overall design

14 contributor

15 contributor

16

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

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

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).

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

[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-

GO_molecular_function cytoplasm;extracellular exosome;tertiary granule lumen;ficolin-1-rich granule

amyloid-beta binding;protease binding;endopeptidase inhibitor activity;cystein

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

IndividualID sle1

Description

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

Full ▾

Send to: ▾

JHH-2, human cell line STR and SNP profiles from GNE, Genentech

Identifiers BioSample: SAMN03473249; GNE: GNE Tracking ID: 586138

Organism [Homo sapiens](#) (human)

Attributes	cell line	JHH-2
	culture collection	GNE:586138
	repository	Genentech (GNE)
	tissue	liver
	disease	carcinoma hepatocellular
	sex	male
	ethnicity	japanese
	age	57 year
	development stage	adult
	canonical name	JHH-2
	human cell line STR profile	yes
	human cell line STR profile status	repository authenticated
	human cell line SNP profile	yes

BioProject

[PRJNA271020](#) Homo sapiens
Retrieve [all samples](#) from this project

Related information

[BioProject](#)

[BioCollections](#)

[Taxonomy](#)

Recent activity

[Turn Off](#) [Clear](#)

JHH-2, human cell line STR and SNP profiles from GNE, sul biosample

Human sample from Homo sapiens biosample

"disease=carcinoma hepatocellular"[attr] (28) [BioSample](#)

Con rep1 biosample

carcinoma hepatocellular (11749) [BioSample](#)

[See more...](#)

LinkOut to external resources

[JHH-2 \(CVCL_2786\)](#)

Full ▾

Send to: ▾

JHH-2, human cell line STR and SNP profiles from GNE, Genentech

Identifiers BioSample: SAMN03473249; GNE: GNE Tracking ID: 586138

Organism [Homo sapiens](#) (human)

Attributes	cell line	JHH-2
	culture collection	GNE:586138
	repository	Genentech (GNE)
	tissue	liver
	disease	carcinoma hepatocellular
	sex	male
	ethnicity	japanese
	age	57 year
	development stage	adult
	canonical name	JHH-2
	human cell line STR profile	yes
	human cell line STR profile status	repository authenticated
	human cell line SNP profile	yes

BioProject

[PRJNA271020](#) Homo sapiens
Retrieve [all samples](#) from this project

Related information

[BioProject](#)

[BioCollections](#)

[Taxonomy](#)

Recent activity

[Turn Off](#) [Clear](#)

- [JHH-2, human cell line STR and SNP profiles from GNE, sut biosample](#)
- [Human sample from Homo sapiens biosample](#)
- ["disease=carcinoma hepatocellular"\[attr\] \(28\) BioSample](#)
- [Con rep1 biosample](#)
- [carcinoma hepatocellular \(11749\) BioSample](#)

[See more...](#)

LinkOut to external resources

[JHH-2 \(CVCL_2786\)](#)

NCBI *BioSample* metadata are ... disappointing

- 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*

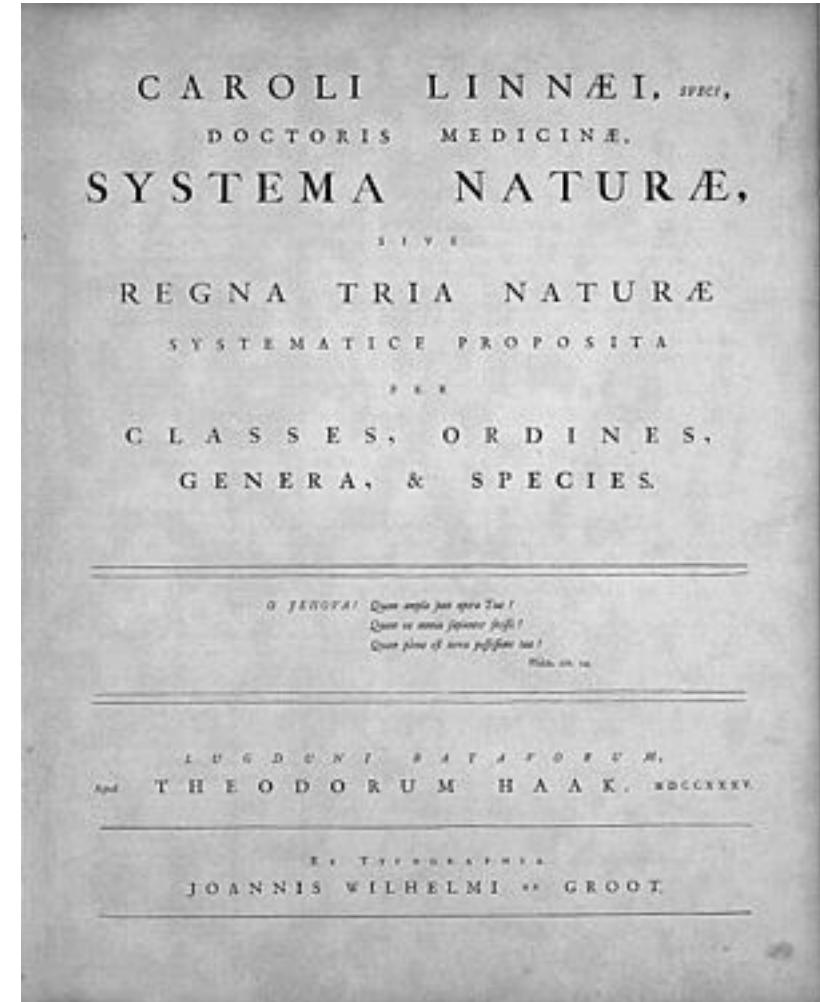
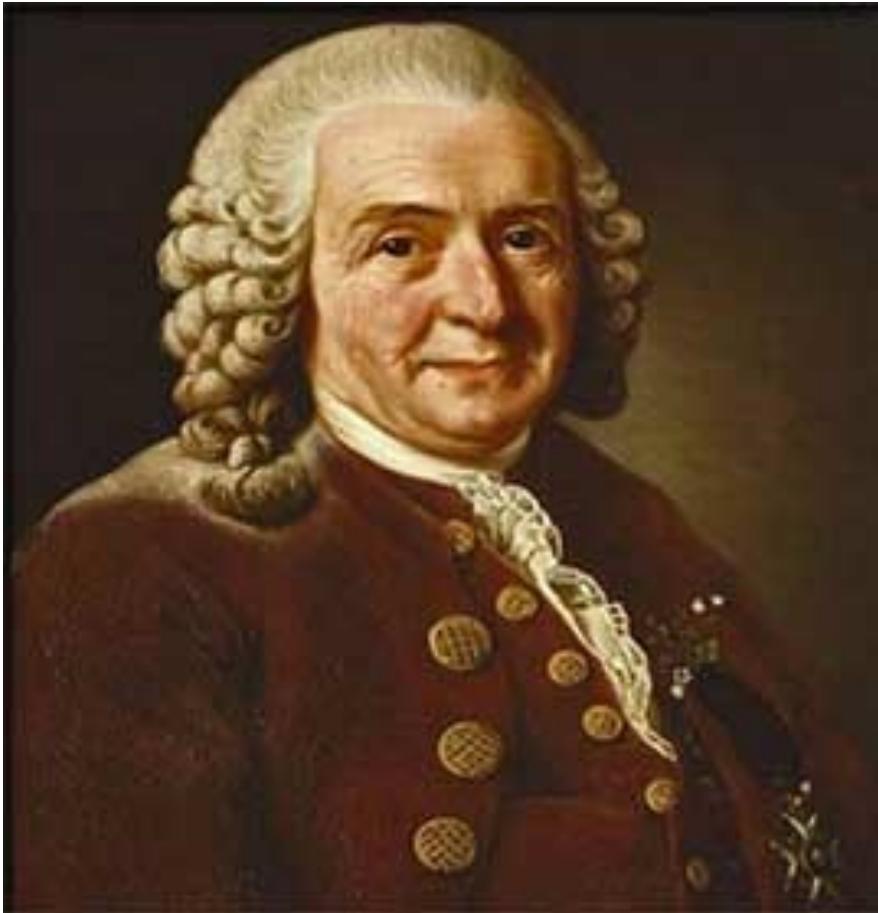
Metadata authors need to use controlled terms!

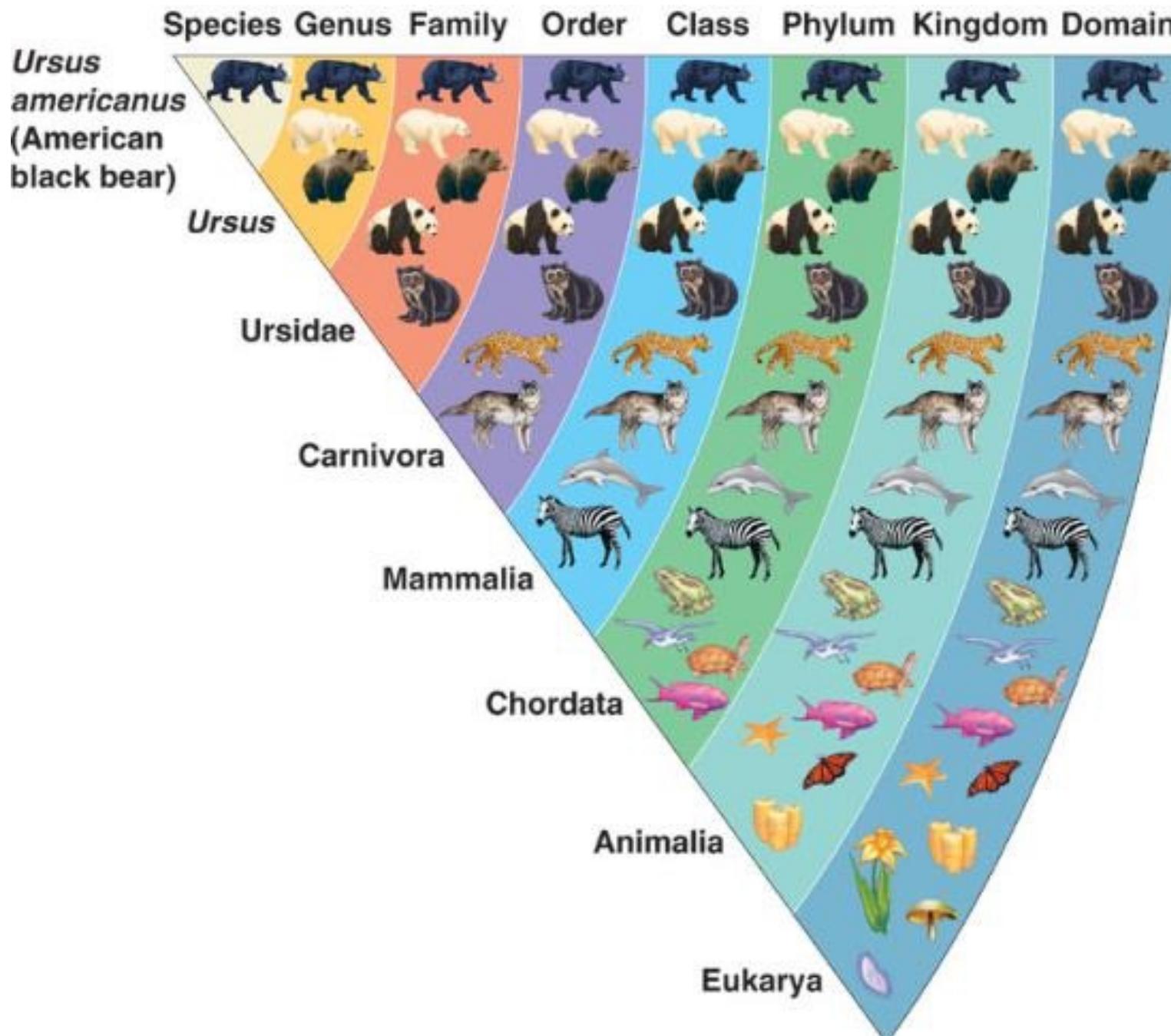
<i>age</i>	<i>age [y]</i>
<i>Age</i>	<i>age [year]</i>
<i>AGE</i>	<i>age [years]</i>
<i>`Age</i>	<i>age in years</i>
<i>age (after birth)</i>	<i>age of patient</i>
<i>age (in years)</i>	<i>Age of patient</i>
<i>age (y)</i>	<i>age of subjects</i>
<i>age (year)</i>	<i>age(years)</i>
<i>age (years)</i>	<i>Age(years)</i>
<i>Age (years)</i>	<i>Age(yrs.)</i>
<i>Age (Years)</i>	<i>Age, year</i>
<i>age (yr)</i>	<i>age, years</i>
<i>age (yr-old)</i>	<i>age, yrs</i>
<i>age (yrs)</i>	<i>age.year</i>
<i>Age (yrs)</i>	<i>age_years</i>

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

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

We've known for 300 years that science needs controlled terms







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

Search for a class

Enter a class, e.g. Melanoma



[Advanced Search](#)

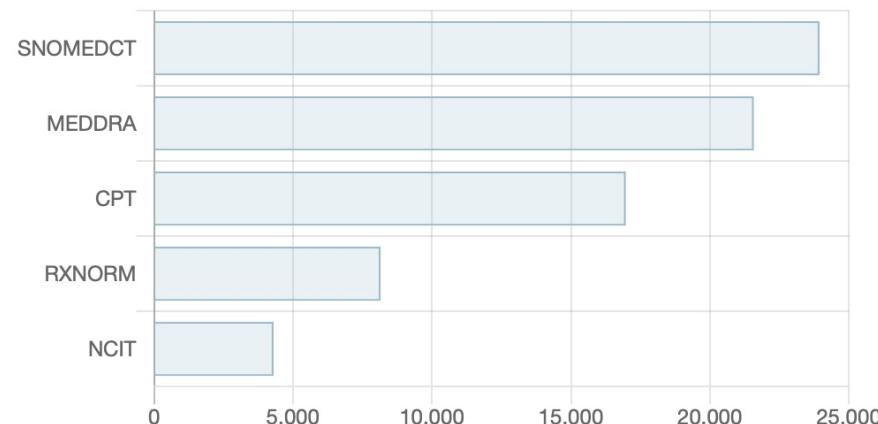
Find an ontology

Start typing ontology name, then choose from list



[Browse Ontologies ▾](#)

Ontology Visits (November 2020)



[More](#)

BioPortal Statistics

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

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](#) (2)
- [Arabidopsis](#) (2)
- [Biological Process](#) (44)
- [Biomedical Resources](#) (58)
- [Cell](#) (69)

Group

- [BIBLIO](#) (9)
- [BIS](#) (3)
- [CGIAR](#) (1)
- [CTSA](#) (6)
- [OBO_Foundry](#) (9)

Current Procedural Terminology (CPT)

Current Procedural Terminology

Uploaded: 2/6/17

projects
1classes
13,289

Medical Dictionary for Regulatory Activities (MEDDRA)

Medical Dictionary for Regulatory Activities Terminology (MedDRA)

Uploaded: 2/6/17

notes
1projects
10classes
69,107

RxNORM (RXNORM)

RxNorm Vocabulary

Uploaded: 2/6/17

projects
7classes
115,514

SNOMED CT (SNOMEDCT)

SNOMED Clinical Terms

Uploaded: 2/6/17

notes
2projects
22classes
327,128

National Drug Data File (NDDF)

National Drug Data File Plus Source Vocabulary

Uploaded: 2/6/17

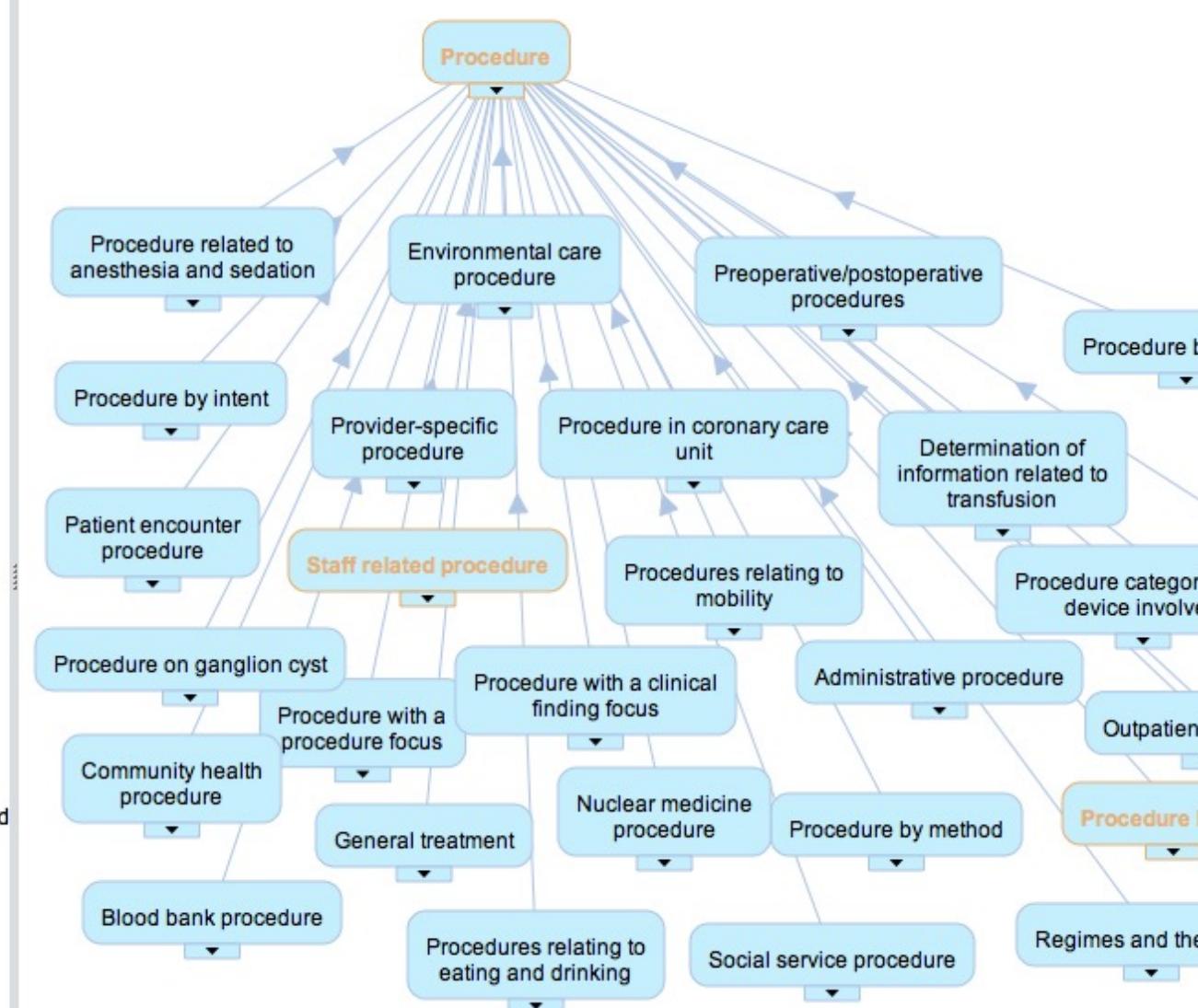
projects
1classes
28,111

SNOMED Clinical Terms

Terms ▾

Jump To: [Details](#) [Visualization](#) [Notes \(0\)](#) [Term Mappings \(31\)](#) [Term Resources](#)

path to root ▾

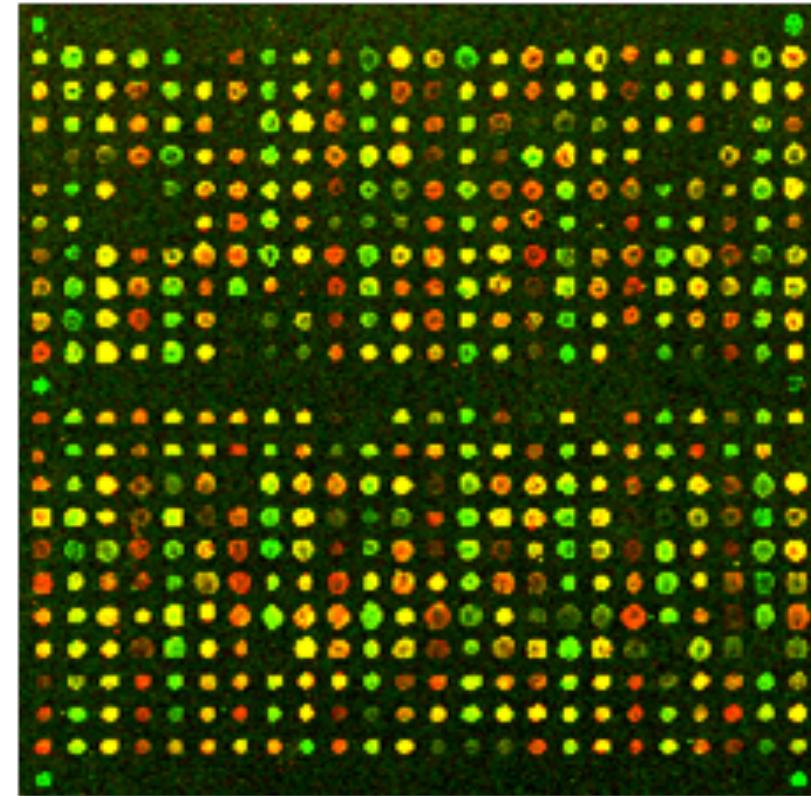


If we want to have FAIR data, we need good metadata. Good metadata need:

- **Reporting guidelines** to provide a standardized structure for the metadata components
- **Ontologies** to provide controlled terms
- **Technology** to make it easy to author good metadata 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 MEtagenome Sequence analysis Standard (MINIMESS)
- Minimal Information Specification For In Situ Hybridization and Immunohistochemistry Experiments (MISFISHIE)
- Et cetera

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

- **Reporting guidelines**—like MIAME—to provide a standardized structure for the metadata components
- **Ontologies** to provide controlled terms
- **Technology** to make it easy to author good metadata in the first place

Our approach in CEDAR

- Encode standard, community-endorsed *reporting guidelines* as **metadata templates** that offer fill-in-the-blank authoring opportunities
- Use selections from *ontologies* whenever possible to provide **standardized values** for the template fields



Some FAIR principles can be addressed only by domain experts themselves using custom metadata templates

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

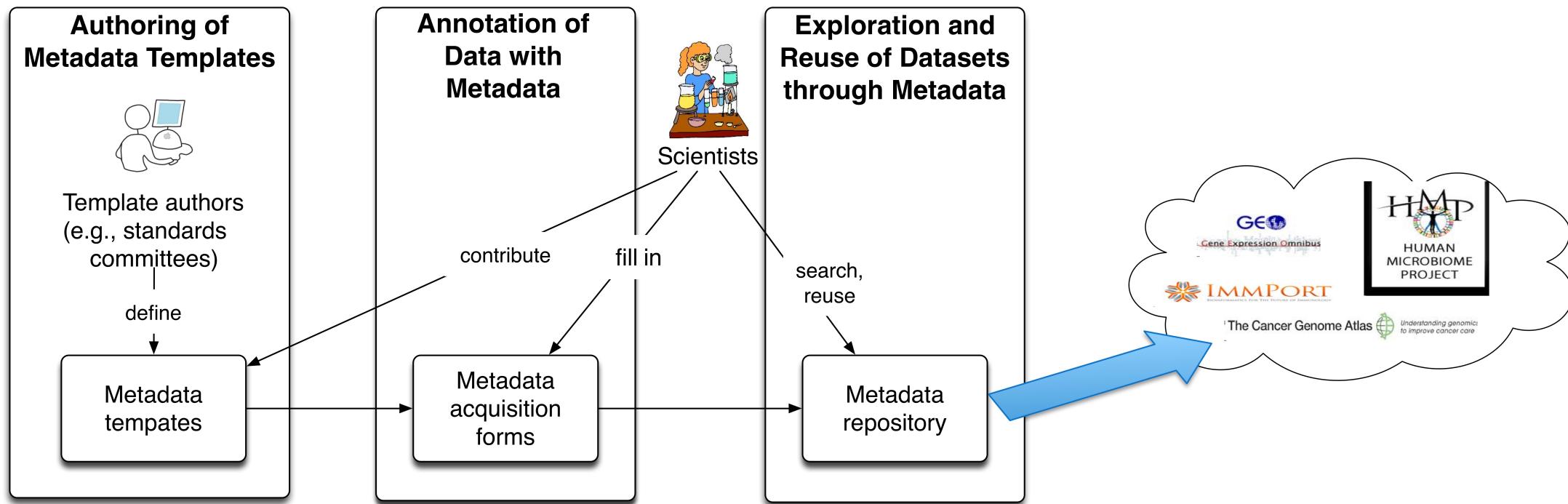
R1.3: (Meta)data meet domain-relevant community standards

A **metadata template** can ensure compliance with all investigator-controlled FAIR principles, including:

- Making metadata “rich”
- Using metadata vocabularies that follow the FAIR principles
- Meeting domain-relevant community metadata standards

BioSample Human	
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



WorkspaceShared with
Me[FILTER](#) [RESET](#)

TYPE



Workspace	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



WorkspaceShared 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	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

- Open
- Populate
- Share...
- Copy to...
- Move to...
- Rename...
- Delete





▼ BioSample Human

* Sample Name

* Organism

* Tissue

* Sex

* Isolate

* Age

* Biomaterial Provider

▼ Attribute

Name

Value

CANCEL

VALIDATE

SAVE

▼ 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%)

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

- **Reporting guidelines**—like MIAME—to provide a uniform structure
- **Ontologies** to provide controlled terms
- **Technology** to make it easy to author good metadata in the first place

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 title

11 summary

12 summary

13 overall design

14 contributor

15 contributor

16
17 **SAMPLES**

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

19 Sample name

title

CHP file

source name

organism

characteristics: tag

20 SAMPLE 1

21 SAMPLE 2

22 SAMPLE 3

23 SAMPLE 4

24 SAMPLE 5

25 SAMPLE 6

26 SAMPLE 7

27 SAMPLE 8

28 SAMPLE 9

29 SAMPLE X

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

31 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).

32
33 **PROTOCOLS**

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

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

36 growth protocol

37 treatment protocol

38 extract protocol

39 label protocol

40 hyb protocol

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

The diagram consists of two concentric circles. The inner circle is light blue and contains the following text:

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)

The outer ring is red and contains the following text:

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

The FAIR Evaluator - automated testing of Web resources for their compliance

FAIR Metrics Evaluation

HIDE DETAILS

Registry URL

<http://smart-api.info/registry?q=ad830426bed193d36838091ef5d14407>

SmartAPI ID

ad830426bed193d36838091ef5d14407 ↗

Copy

Metadata URL

https://w3id.org/FAIR_Evalu... ↗

Version

V 0.3.0

Contact

Mark Wilkinson

A **metadata template** can ensure compliance with all investigator-controlled FAIR principles, including:

- Making metadata “rich”
- Using metadata vocabularies that follow the FAIR principles
- Meeting domain-relevant community metadata standards

BioSample Human	
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

Metadata for Machines Workshops

- Are intensive 2–3 day invited, highly participatory sessions
- 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



M4M for the Danish e-infrastructure Cooperation

POSTED ON 8 JULY 2020

Making it easy for humans to make metadata for machines

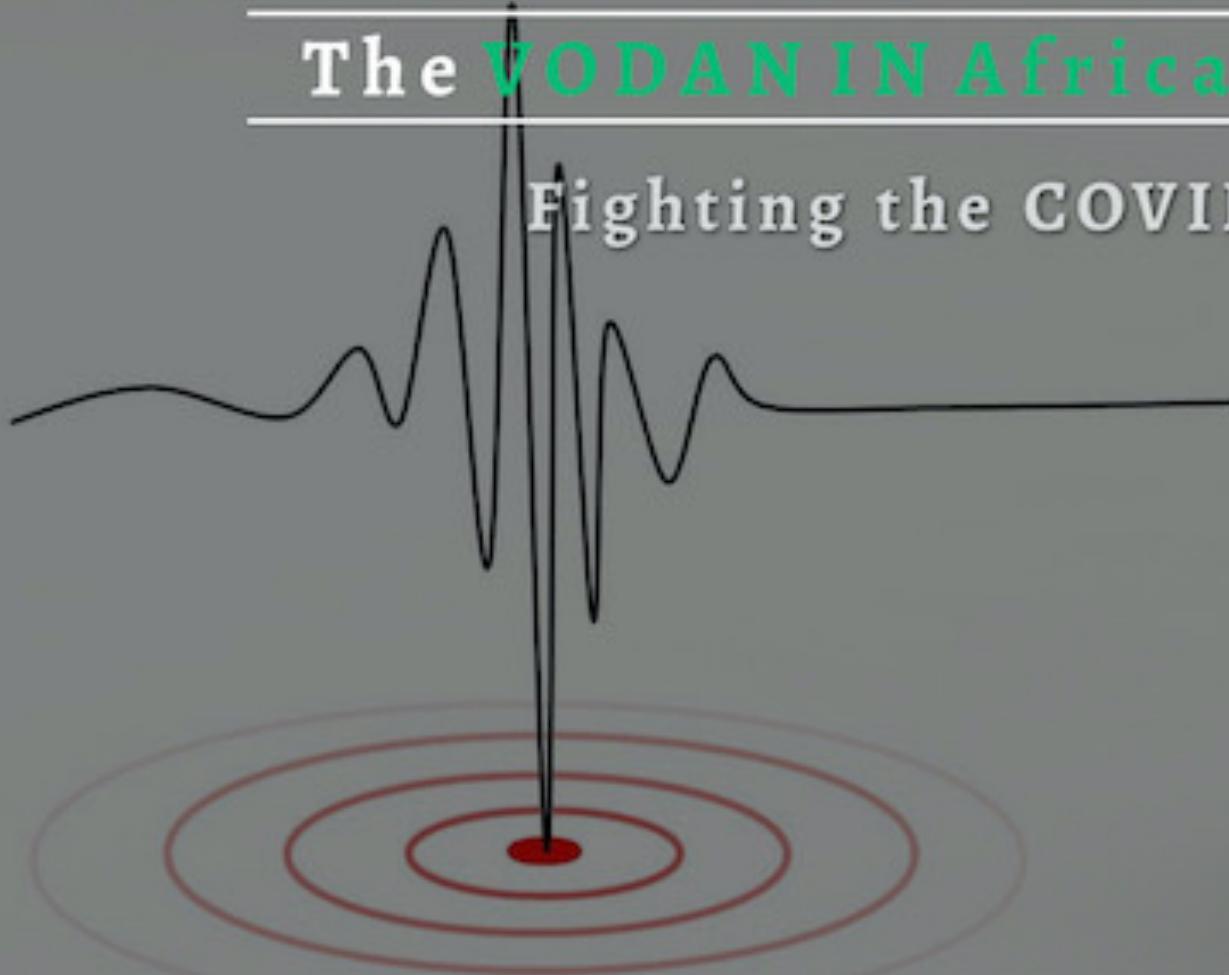
On June 26, the [Danish e-Infrastructure Cooperation](#) (DeiC), in cooperation with the [GO FAIR Foundation](#), launched two Metadata for Machine workshops on behalf of two research communities seeking to upgrade the FAIRness of research data. The workshops are conducted via teleconference in a series of five modules to be completed in mid-September.



Participants include: members from the [AnaEE research infrastructure](#) (M4M.5), the National Energy System Transition Facilities project represented by the [Wind Energy department](#) at Danmarks Tekniske Universitet (M4M.6), and John Graybeal from Stanford University's [Center for Expanded Data Annotation and Retrieval](#) (CEDAR). The workshop is co-directed by Erik Schultes from the GO FAIR International Support and Coordination Office, and Diba Terese Markus, [Aalborg University Library](#).

The VODAN IN Africa Training the Trainers

Fighting the COVID-19 with FAIR Data



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



But what about all the *legacy* metadata out there?

- The goal: A **metadata powerwash** to transform extant metadata into standards-compliant, FAIR data annotations, where ...
- CEDAR metadata templates offer insight into how authors of existing metadata might have intended to structure their metadata
- Algorithms akin to spell checkers can align:
 - Metadata field *names* with those used in metadata templates
 - Metadata field *values* with terms from indicated ontologies



Online Data Will Never Be FAIR

- Until we standardize metadata structure using common **metadata 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 **investigators** appreciate the importance of creating good metadata from the start



But what are the challenges?

- Most areas of science don't have adequate ontologies
- Most areas of science don't have adequate reporting guidelines
- Many people like the idea of *FAIR data* as a slogan, but there are few people motivated to do the real work needed to make data FAIR
- The world needs a lot more people who have expertise in data stewardship
- Investigators *hate* authoring metadata!

