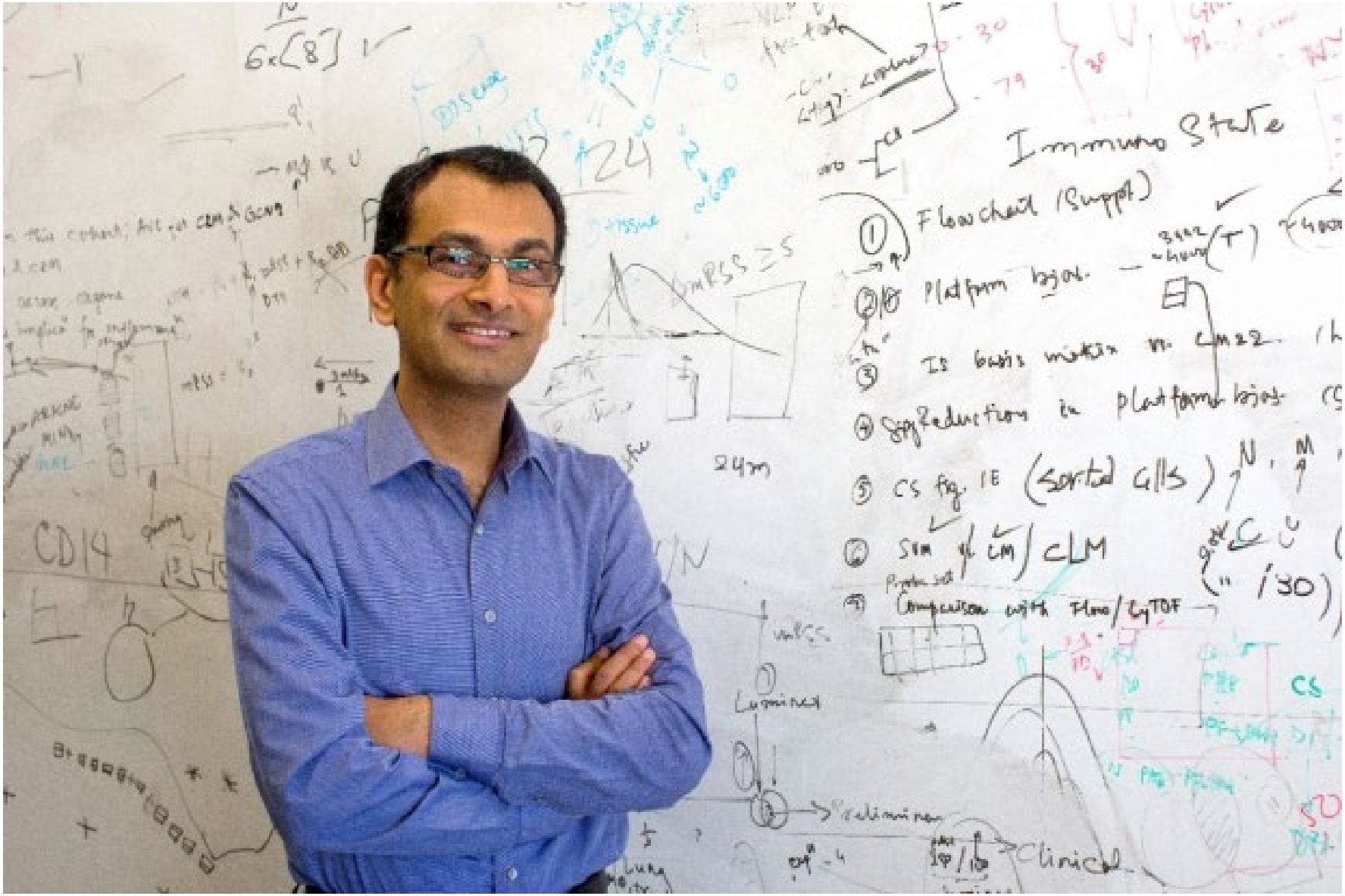
CEDAR: Semantic Technology in Support of Open Science and Improved Knowledge Management

Mark A. Musen, M.D., Ph.D.

Stanford University
musen@Stanford.EDU

CEDAR

CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL



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

Khatri has reused public data sets 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!

Getting access to other people's data is hard!

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

	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						
11	**title**					
12	**summary**					
13	**summary**					
14	**overall design**					
15	**contributor**					
16	**contributor**					
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**					
32						
33						
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**					

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

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

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).

[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

Gene Expression Omnibus

An Analysis of Metadata from *BioSample*

- 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.

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

BioPortal Browse Search Mappings Recommender Annotator Resource Index Projects

Gene Ontology

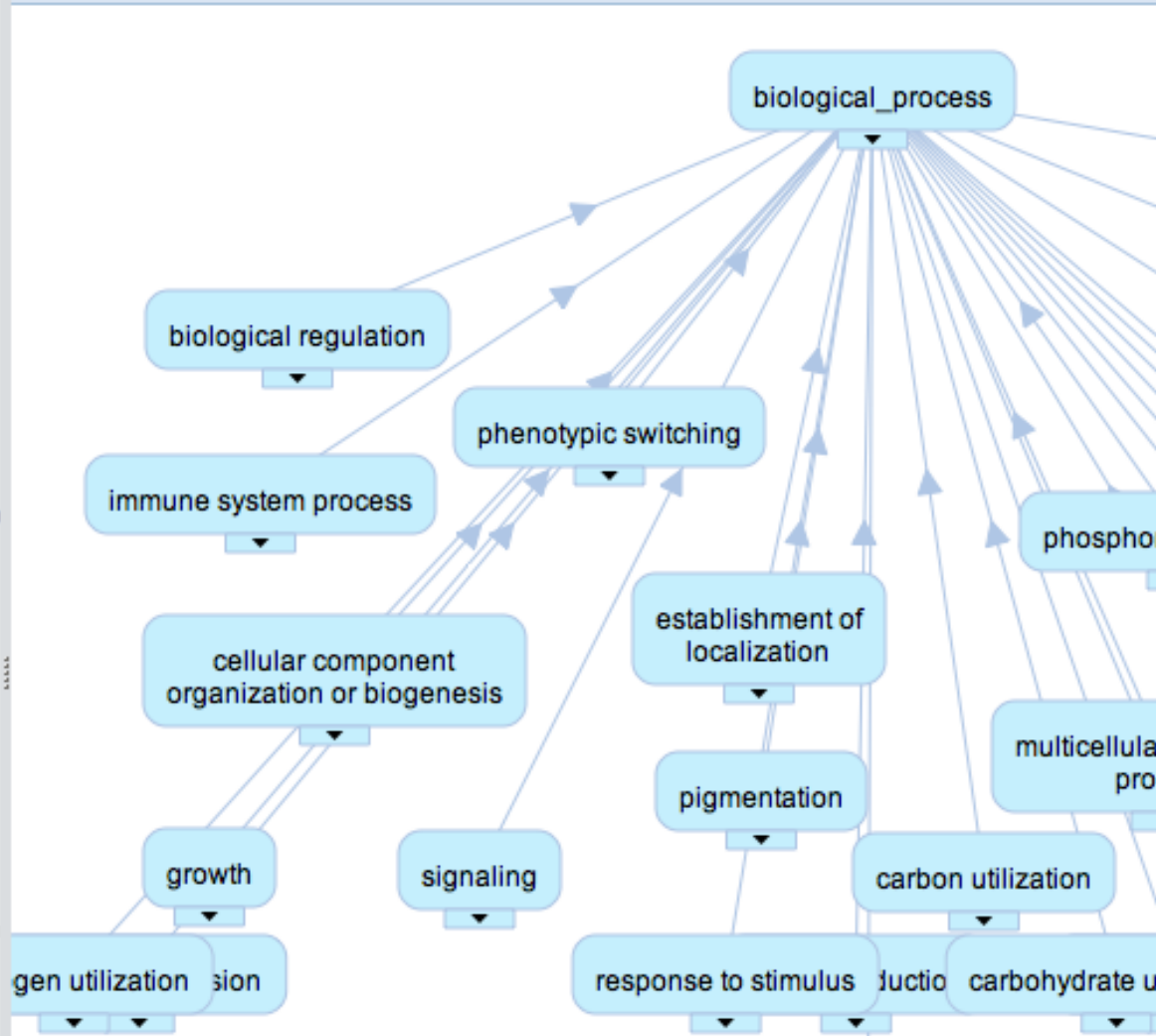
Terms

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

Details | Visualization | Notes (0) | Term Mappings (31) | Term Resources

path to root



BioPortal Browse Search Mappings Recommender Annotator Resource Index Projects

SNOMED Clinical Terms

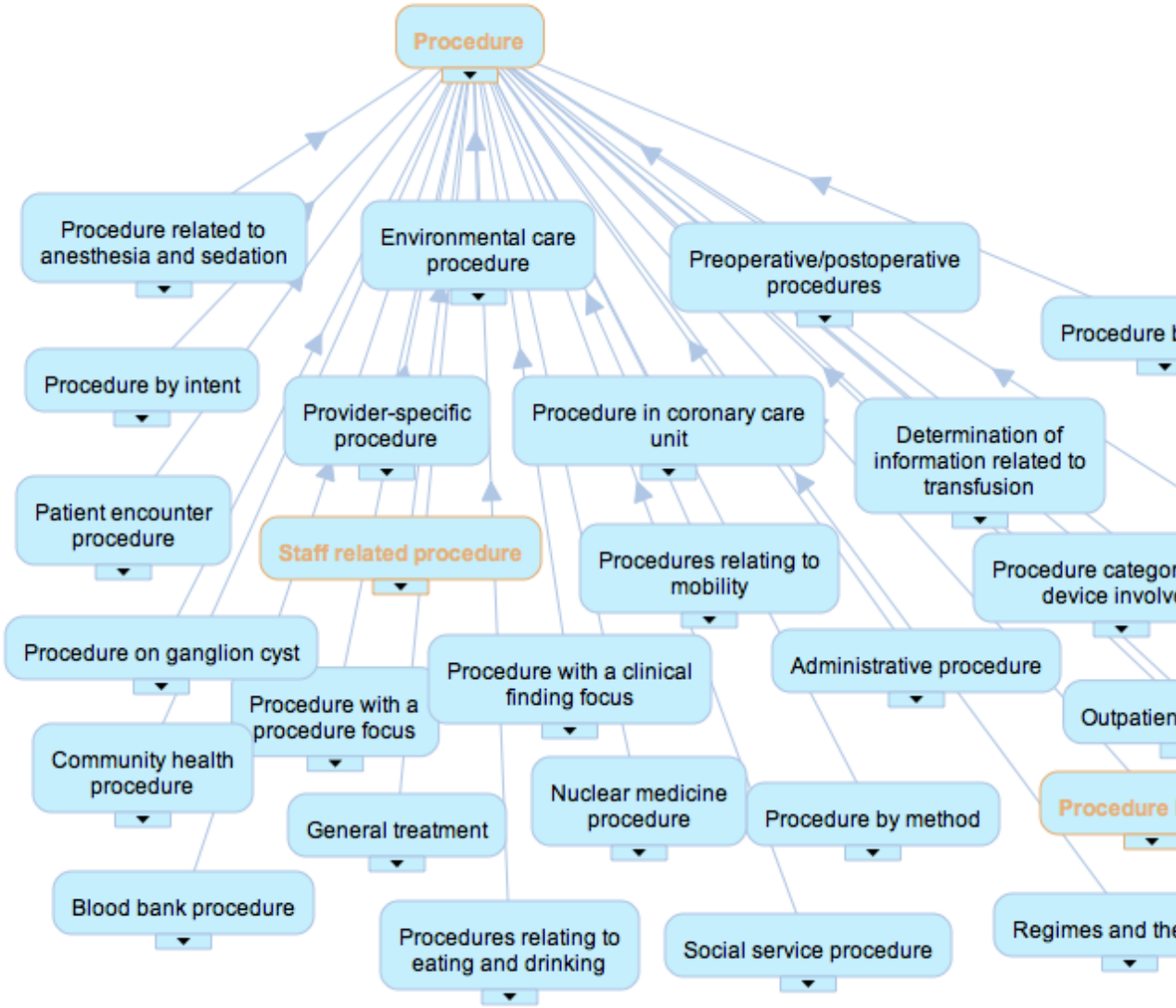
Terms

Jump To:

- 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

Details | Visualization | Notes (0) | Term Mappings (31) | Term Resources

path to root



BioPortal Ontologies Search Annotator Recommender Mappings Resource Index

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

Search for a class

Enter a class, e.g. Melanoma

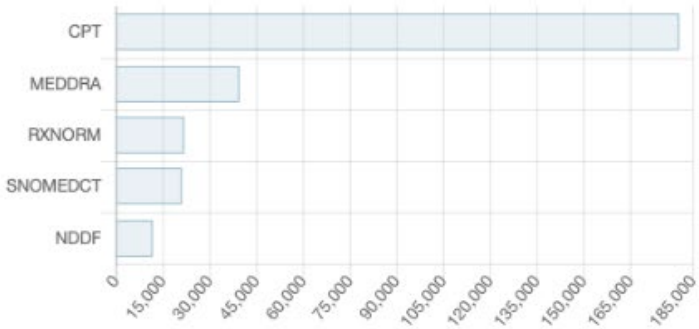
Advanced Search

Find an ontology

Start typing ontology name, then choose from list

Browse Ontologies

Ontology Visits (October 2018)



More

Screenshot

BioPortal Statistics

Ontologies	737
Classes	9,605,019
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

Requirement #2: Describe properties of experiments completely and consistently

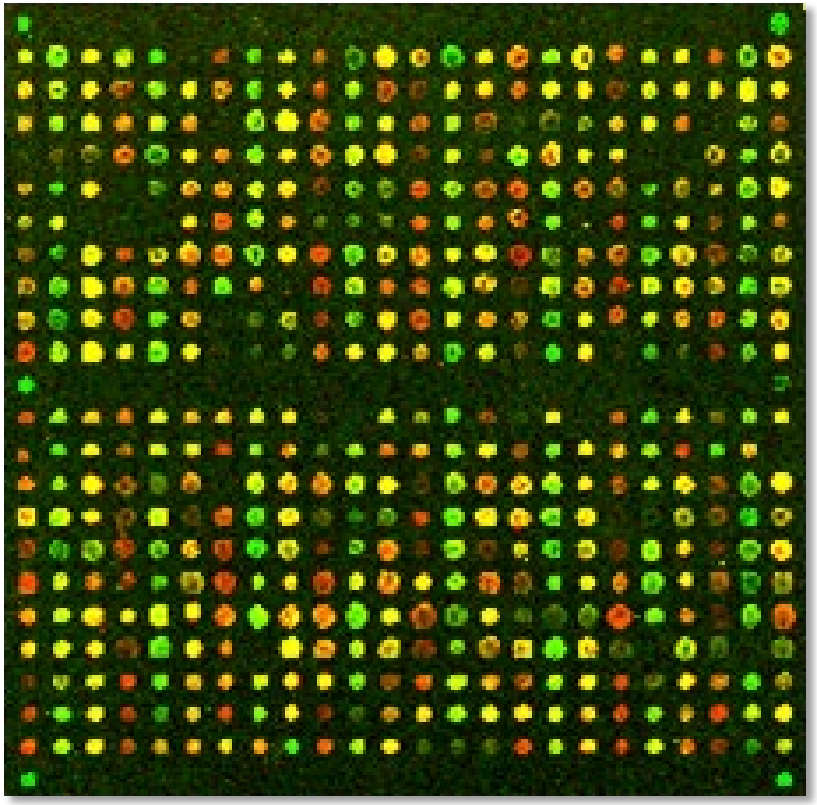
We need metadata to describe

- The digital context (properties of the file)
- The investigators and stakeholders
- The scientific context
  - The motivation for the experiment
  - The data that were collected
  - The methods of the experiment
  - The instruments that were used
  - When and where the data were collected
- The parameters of the data

— DataOne Primer on Data Management

The microarray community took the lead in standardizing biological metadata

- 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)

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

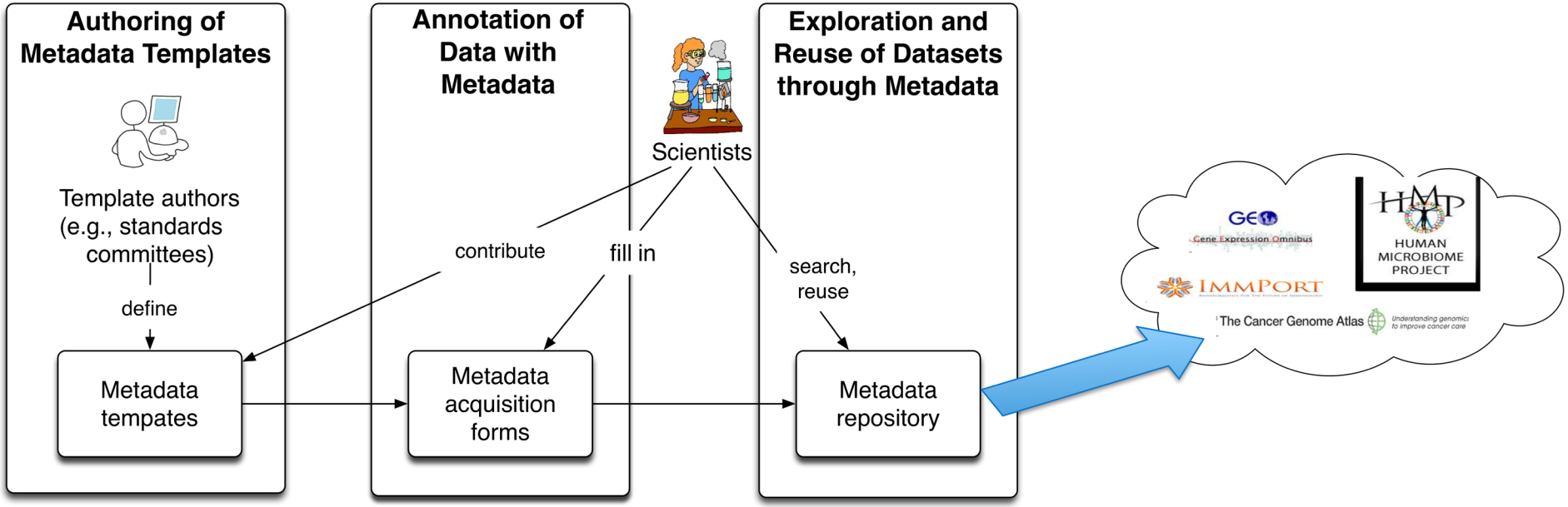
Requirement #3: Make it palatable to describe experiments completely and consistently

http://metadatacenter.org

CEDAR

CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL

The CEDAR Approach to Metadata



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 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	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	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

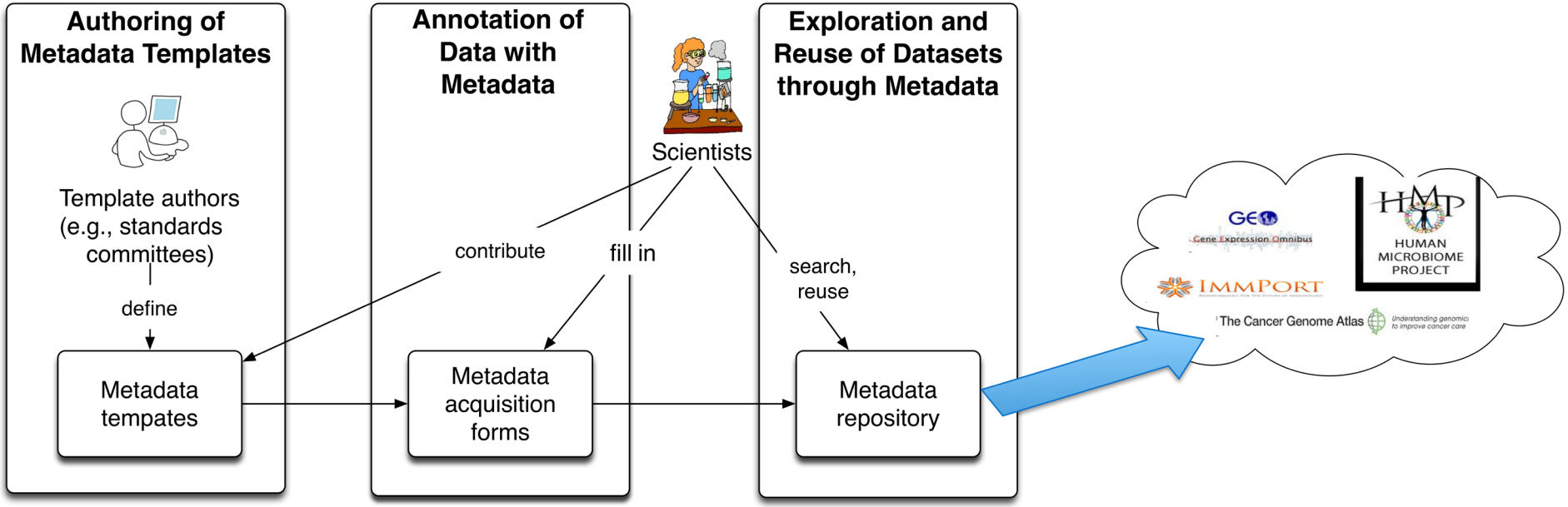
Attribute (2)

Name	description
Value	Cell line was cultured until the 5th passage

Attribute (3)

Name	treatment
Value	350mg brodalumab

The CEDAR Workbench



BioSample Human

a *

Sample Name*

a *

Organism*

a *

Enter Field Title
Tissue

Enter Field Description (Help Text)
Enter the type of tissue.

Enter Default Value

VALUES | MULTIPLE | REQUIRED | SUGGESTIONS | HIDDEN | INSTANCE TYPE

Name	Type	Source	Identifier	No. Values

SEARCH

a *

Sex*

a *

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 Metencepha
    - 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

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.

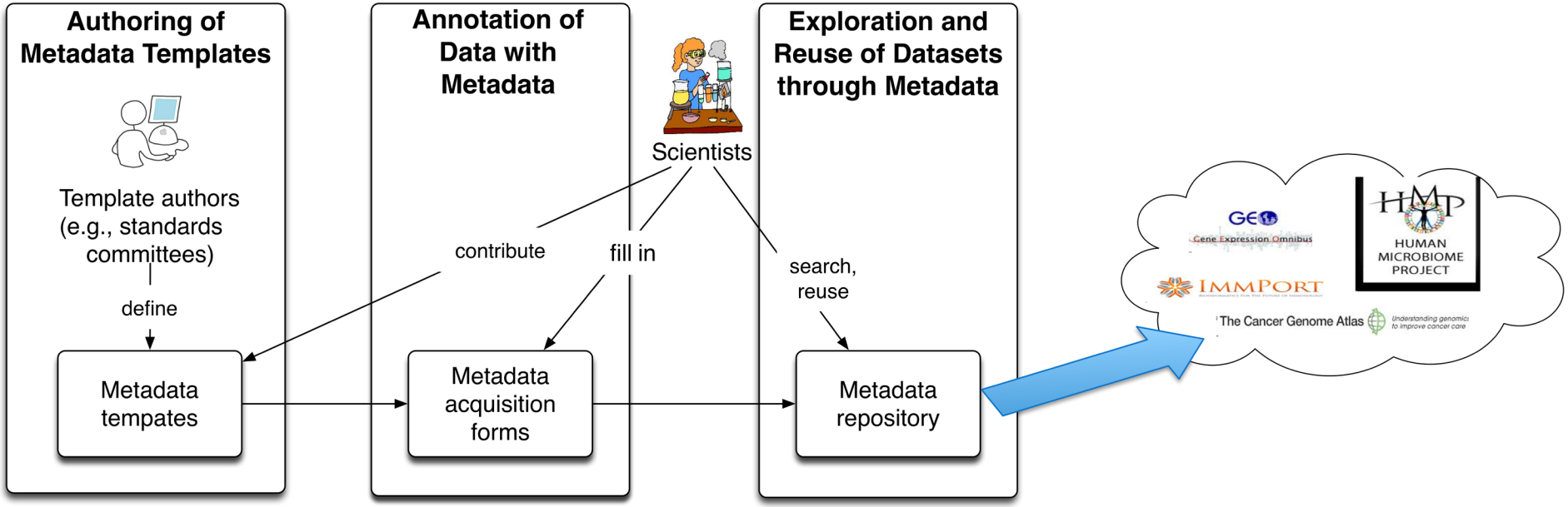
TERM | **BRANCH** | ONTOLOGY

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

- 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

BioSample Human

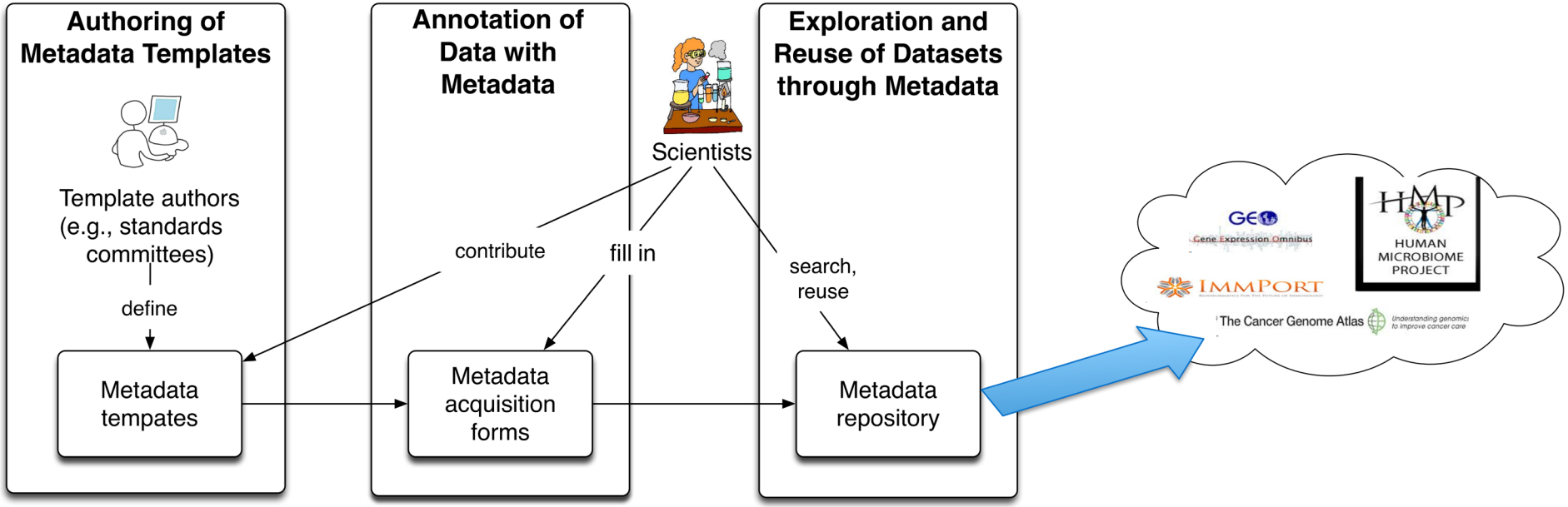
* Sample Name	056
* Organism	Homo sapiens
* Tissue	brain
* Sex	Male
* Isolate	N/A
* Age	74
* Biomaterial Provider	Life Technologies

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



Some key features of CEDAR

- All semantic components—template elements, templates, ontologies, and value sets—are managed as first-class entities
- User interfaces and drop-down menus are not hardcoded, but are generated on the fly from CEDAR's semantic content
- All software components have well defined APIs, facilitating reuse of software by a variety of clients
- CEDAR generates all metadata in JSON-LD, a widely adopted Web standard that can be translated into other representations

AIRR Community

Home | News | Meetings | Working Groups

About the AIRR Community

The Adaptive Immune Receptor Repertoire (AIRR) Community is a community-driven organization that is organizing and coordinating stakeholders in the use of NGS technologies to study antibody (Ab)/B-cell and T-cell receptor (TcR) repertoires. Recent advances in sequencing technology have made it possible to sample the immune repertoire in exquisite detail. AIRR sequencing has enormous promise for understanding the dynamics of the immune repertoire in vaccinology, infectious disease, autoimmunity, and cancer biology, but also poses substantial challenges. To meet these challenges, we have established the AIRR Community.

AIRR is providing our first experience uploading CEDAR-authored metadata directly to NCBI

NIH LINCS
PROGRAM

LIBRARY

HOME CENTERS DATA COMMUNITY PUBLICATIONS NEWS

LINCS aims to create a network-based understanding of biology by cataloging changes in gene expression and other cellular processes that occur when cells are exposed to a variety of perturbing agents

GO FAIR: a bottom-up international approach

for the practical implementation of the European Open Science Cloud (EOSC) as part of a global Internet of FAIR Data & Services

Context of GO FAIR

Watch videos

Vision

Fostering the coherent development of the global Internet of FAIR Data & Services (IFDS), with the main focus on early developments in the European Open Science Cloud (EOSC).

LEARN MORE

Strategy

GO FAIR follows a bottom-up open implementation strategy for the technical governance and funding needed to establish the first phase of the European Open Science Cloud (EOSC) as part of a broader global Internet of FAIR Data & Services. The approach is largely based on the EOSC communication and the recommendations of the High Level Expert Group.

LEARN MORE

Open data is about MORE THAN DISCLOSURE it must be "Fair"

- Findable
- Accessible
- Interoperable
- Reusable

Where is this ecosystem leading?

- Technology such as CEDAR will assist in the automated "publication" of scientific results online
- Computer-based, intelligent agents will
  - Search and "read" the "literature"
  - Integrate information
  - Track scientific advances
  - Re-explore existing scientific datasets
  - Suggest the next set of experiments to perform
  - And maybe even do them!

CEDAR
CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL

http://metadatacenter.org

CEDAR

CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL