

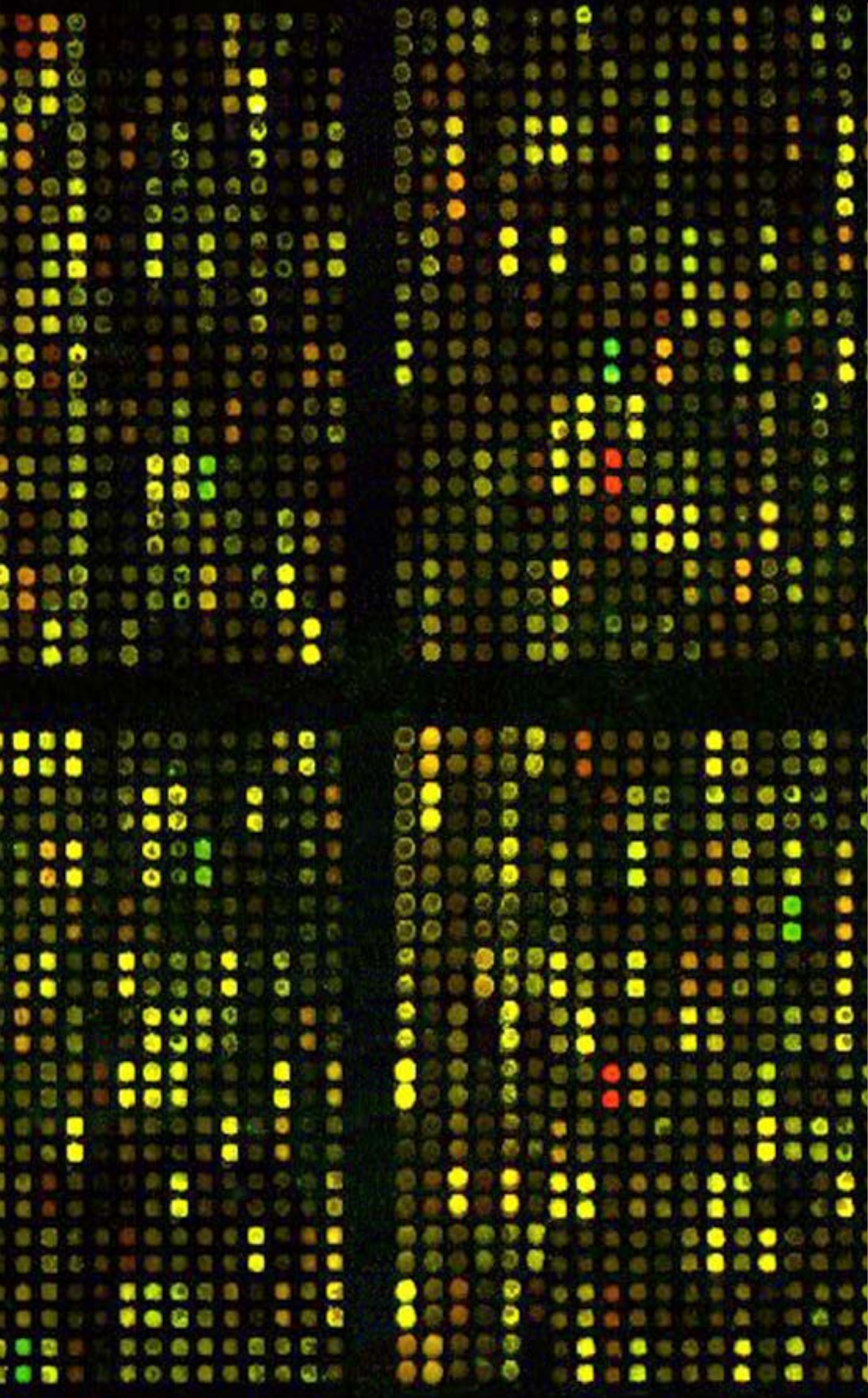
CEDAR

CENTER FOR EXPANDED DATA ANNOTATION  
AND RETRIEVAL

*Making it Easier, Possibly Even Pleasant,  
to Author Rich Experimental Metadata*

**High Quality Metadata are  
Essential  
for Large-Scale Reuse  
and Biomedical Discovery**

**What is this colored picture about?**



## 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](#).

MIAME does not specify a particular format, however, obviously the data are more usable, if it is encoded in a way that the essential information specified by MIAME can be accessed easily. [FGED](#) recommends the use of [MAGE-TAB](#) format, which is based on spreadsheets, or [MAGE-ML](#).

MIAME also does not specify any particular terminology, however for automated data exchange the use of standard controlled vocabularies and ontologies are desirable. [FGED](#) recommends the use of [MGED](#) [Ontology](#) for the description of the key experimental concepts, and where possible ontologies developed by the respective community for describing terms such as anatomy, disease, chemical compounds etc (see [OBO](#) page for more detail).

## MIBBI portal

- Registration form [for](#) the MIBBI Portal (please return to [christfaylor@gmail.com](mailto:christfaylor@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

Core Information for **Metabolomics Reporting**

#### GIATE

Guidelines for Information About **Therapy Experiments**

#### MIABE

Minimal Information About a **Bioactive Entity**

#### MIABIE

Minimum Information About a **Biofilm Experiment**

#### MIACA

Minimal Information About a **Cellular Assay**

#### MIAME

Minimum Information About a **Microarray Experiment**

#### MIAPA

Minimum Information About a **Phylogenetic Analysis**

#### MIAPAR

Minimum Information About a **Protein Affinity Reagent**

#### MIAPE

Minimum Information About a **Proteomics Experiment**

#### MIAPepAE

Minimum Information About a **Peptide Array Experiment**

#### MIARE

Minimum Information About a **RNAi Experiment**

#### MIASE

Minimum Information About a **Simulation Experiment**

#### MIASPPe

Minimum Information About **Sample Preparation** for a **Phosphoproteomics Experiment**

#### MIATA

Minimum Information About **T Cell Assays**

#### MICEE

Minimum Information about a **Cardiac Electrophysiology Experiment**

# Dataset Descriptions: HCLS Community Profile

<http://www.w3.org/TR/hcls-dataset/>

## Editors' working draft.

### Editors:

Alasdair J.G. Gray, Heriott-Watt University, UK <[A.J.G.Gray@hw.ac.uk](mailto:A.J.G.Gray@hw.ac.uk)>  
 Joachim Baran, Stanford University, USA <[joachim.baran@stanford.edu](mailto:joachim.baran@stanford.edu)>  
 M. Scott Marshall, MAASTRO Clinic, The Netherlands <[m.scott.marshall@maastro.nl](mailto:m.scott.marshall@maastro.nl)>  
 Michel Dumontier, Stanford University, USA <[michel.dumontier@stanford.edu](mailto:michel.dumontier@stanford.edu)>

### Contributors:

Vladimir Alexiev, Ontotext Corp, Bulgaria <[vladimir.alexiev@ontotext.com](mailto:vladimir.alexiev@ontotext.com)>  
 Peter Ansell, CSIRO, Australia <[peter.ansell@csiro.au](mailto:peter.ansell@csiro.au)>  
 Gary D. Bader, The Donnelly Centre, University of Toronto, Canada <[gary.bader@utoronto.ca](mailto:gary.bader@utoronto.ca)>  
 Asuka Bando, NIBDC, Japan <[bando@biosciencedbc.jp](mailto:bando@biosciencedbc.jp)>  
 Jerven Bolleman, SIB Swiss Institute of Bioinformatics, Switzerland <[jerven.bolleman@isb-sib.ch](mailto:jerven.bolleman@isb-sib.ch)>  
 Alison Callahan, Carleton University, Canada <[alison.callahan@carleton.ca](mailto:alison.callahan@carleton.ca)>  
 José Cruz-Toledo, Carleton University, Canada <[josecruztoledo@gmail.com](mailto:josecruztoledo@gmail.com)>  
 Pascale Gaudet, SIB Swiss Institute of Bioinformatics, Switzerland <[pascal.gaudet@isb-sib.ch](mailto:pascal.gaudet@isb-sib.ch)>  
 Erich Gombocz, IO Informatics, USA <[egombocz@io-informatics.com](mailto:egombocz@io-informatics.com)>  
 Alejandra Gonzalez-Beltran, University of Oxford, UK <[alejandra.gonzalez.beltran@gmail.com](mailto:alejandra.gonzalez.beltran@gmail.com)>  
 Paul Groth, VU University Amsterdam, The Netherlands <[p.t.groth@vu.nl](mailto:p.t.groth@vu.nl)>  
 Melissa Haendel, Oregon Health and Science University, USA <[haendel@ohsu.edu](mailto:haendel@ohsu.edu)>  
 Maori Ito, NIBIO, Japan <[maori@nibio.go.jp](mailto:maori@nibio.go.jp)>  
 Simon Jupp, EMBL-EBI, UK <[sjupp@ebi.ac.uk](mailto:sjupp@ebi.ac.uk)>  
 Nick Juty, EMBL-EBI, UK <[njuty@ebi.ac.uk](mailto:njuty@ebi.ac.uk)>  
 Toshiaki Katayama, Database Center for Life Sciences, Japan <[ktvnm@dbcls.jp](mailto:ktvnm@dbcls.jp)>  
 Norio Kobayashi, RIKEN, Japan <[norio.kobayashi@riken.jp](mailto:norio.kobayashi@riken.jp)>  
 Kalpana Krishnaswami, Metaome, USA <[kalpana@metaome.com](mailto:kalpana@metaome.com)>  
 Camille Laibe, EMBL-EBI, UK <[laibe@ebi.ac.uk](mailto:laibe@ebi.ac.uk)>  
 Nicolas Le Novère, Babraham Institute, UK <[n.lenovere@email.com](mailto:n.lenovere@email.com)>  
 Simon Lin, Marshfield Clinic Research Foundation, USA <[sim.lin@mcrcf.mfldclin.edu](mailto:sim.lin@mcrcf.mfldclin.edu)>  
 James Malone, EMBL-EBI, UK <[malone@ebi.ac.uk](mailto:malone@ebi.ac.uk)>  
 Michael Miller, Institute for Systems Biology, USA <[mmiller@systemsbiology.org](mailto:mmiller@systemsbiology.org)>  
 Chris Mungall, Lawrence Berkeley National Laboratory, USA <[cjm@berkeleybop.org](mailto:cjm@berkeleybop.org)>  
 Laurens Rietveld, VU University Amsterdam, The Netherlands <[laurens.rietveld@vu.nl](mailto:laurens.rietveld@vu.nl)>  
 Sara M. Wimalaratne, EMBL-EBI, UK <[sarala@ebi.ac.uk](mailto:sarala@ebi.ac.uk)>  
 Atsuko Yamaguchi, Database Center for Life Sciences, Japan <[atsuko@dbcls.jp](mailto:atsuko@dbcls.jp)>



# FAIR – Findable, Accessible, Interoperable, Reuseable

## → Data Sharing Plans

### To be Findable:

- F1. (meta)data are assigned a globally unique and eternally persistent identifier
- F2. data are described with rich metadata
- F3. (meta)data are registered or indexed in a searchable resource
- F4. metadata specify the data identifier

### To be Accessible:

- A1 (meta)data are retrievable by their identifier using a standardized communications protocol
- A1.1 the protocol is open, free, and universally implementable
- A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2 metadata are eternally accessible, even when the data are no longer available

### To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

### To be Re-usable:

- R1. meta(data) have 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 their provenance
- R1.3. (meta)data meet domain-relevant community standards

# biosharing.org

## STANDARDS

BioSharing standards have been partly compiled by linking to BioPortal, MIBBI and the Equator Network. Or you can filter on [MIBBI Foundry reporting guidelines](#) or [OBO Foundry terminology artifacts](#).

View as Grid | View as Table

Standard Type	Count
REPORTING GUIDELINE	40
EXCHANGE FORMAT	0
TERMINOLOGY ARTIFACT	0
<b>Domains</b>	
ASSAY	14
DNA	12
RNA	8
PROTEIN	7
TRANSCRIPTOME	5
BIOCHEMISTRY	4
BRAIN	4
CELL	1

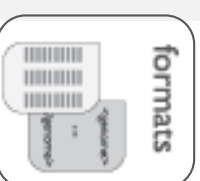
40 records in view

<p><b>BiDDBCore</b></p> <p>Core Attributes of Biological Databases REPORTING GUIDELINE</p> <p>EB Systems: 1 EG Publications: 1</p> <p>1 Taxa types, including: <input type="checkbox"/> ALL</p> <p>1 Data types, including: <input type="checkbox"/> PACKAGE</p>	<p><b>CIMR</b></p> <p>Core Information for Metabolomics Reporting</p> <p>EB Systems: 2 EG Publications: 2</p> <p>No taxa defined.</p> <p>5 Data types, including: <input type="checkbox"/> METABOLITE</p>	<p><b>GIATE</b></p> <p>Guidelines for Information About Therapy Experiments</p> <p>EB Systems: 0 EG Publications: 5</p> <p>No taxa defined.</p> <p>2 Data types, including: <input type="checkbox"/> TREATMENT <input type="checkbox"/> ANTIBODY</p>	<p><b>MIABE</b></p> <p>Minimum Information About a Bioactive Entity</p> <p>EB Systems: 1 EG Publications: 1</p> <p>No taxa defined.</p> <p>2 Data types, including: <input type="checkbox"/> SOLVENT <input type="checkbox"/> MOLECULAR ENTITY</p>
--	---	--	--

68



168





[GEO help](#): Mouse over screen elements for information.

Scope:  Format:  Amount:

**Series GSE35240**

Status Public on Aug 20, 2012

Title Gene expression in mitotic tissues of *Drosophila* too many centrosomes

Organism [Drosophila melanogaster](#)

Experiment type Expression profiling by array

Summary

Centrosome defects are a common feature can proceed through the majority of development amplified centrosomes in most of their cells. centrosome defects do not cause many problems they can adapt to cope with any problems and centrosome amplification predispose fly to assess how centrosome loss or centrosome a by profiling the global transcriptome of *Drosophila* that either lack centrosomes or have too many

Overall design

Mitotic tissues (brains and imaginal discs *Drosophila* larvae of mutants lacking centrosomes with too many centrosomes (SakOE) and *Drosophila* and OregonR). We extracted RNA from three used it for hybridisation to Affymetrix *Drosophila* biological sample, material dissected from expression of the mutant strains was compared

Contributor(s)

[Baumbach J](#), [Levesque MP](#), [Raft JW](#)

Citation(s)

[Baumbach J](#), [Levesque MP](#), [Raft JW](#). Centrosome defects dramatically perturb global gene expression in *Drosophila*. *PLoS One* 15;1(10):983-93. PMID: [23213376](#)

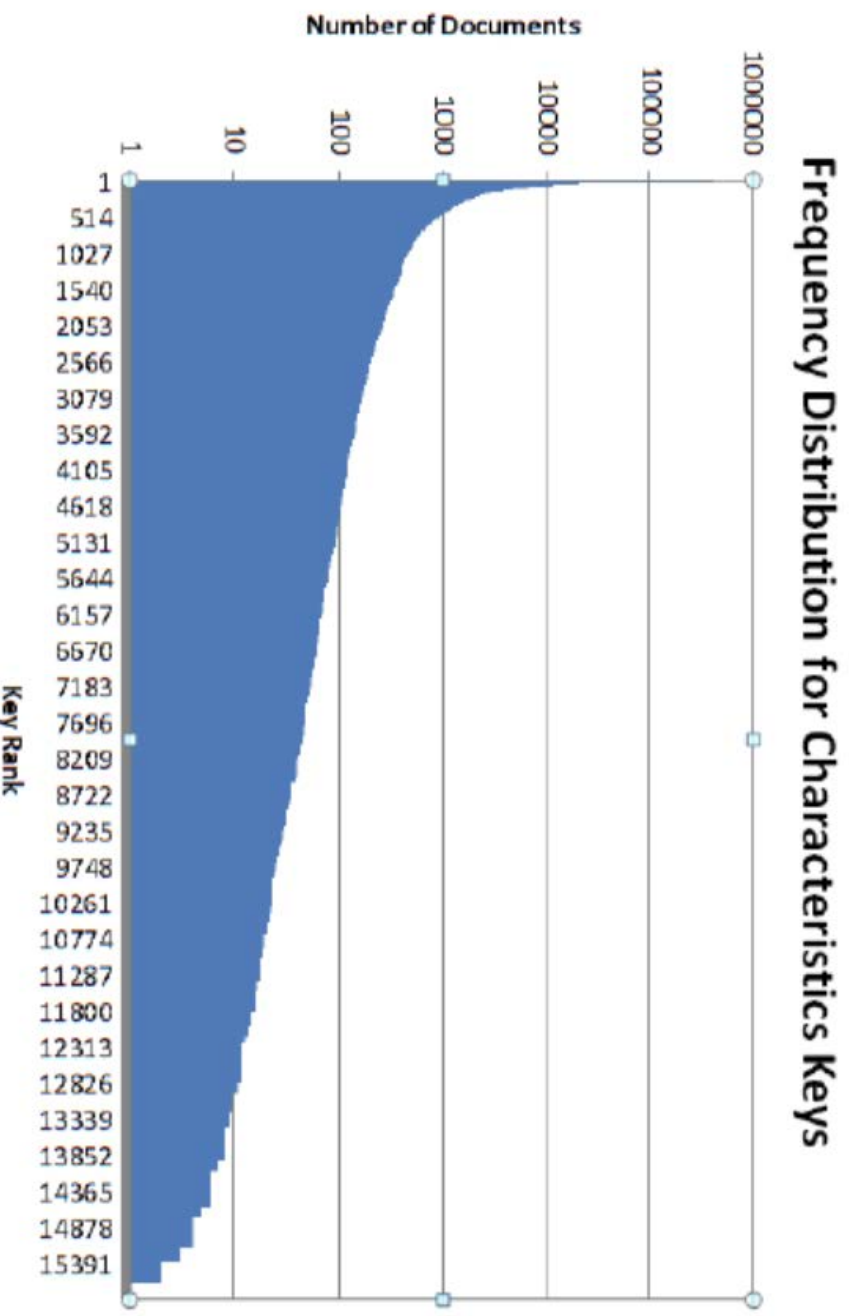
**The Good News:** Minimal information checklists, such as MIAME, are being advanced from all sectors of the biomedical community

**The Bad News:** Investigators view requests for even “minimal” information as burdensome

# Unguided Data Entry is Problematic



age	207147
Age	18089
age (yrs)	9891
age (years)	9272
age (Y)	6226
age in years	1387
age_years	607
AGE	588
age(years)	558
age (year)	433
age (yr)	373
Age (years)	318
age (in years)	310
Age(years)	267
age [year]	97
age [y]	84
age [years]	83
Age(yrs.)	81
age:year	70
age (yr-old)	64
age(yrs)	59
age of patient	40
Age, year	39
Age (yrs)	36
Age of patient	33
age, years	24
'Age	21
Age (Years)	20
age (after birth)	18
age, yrs	12
age of subjects	4



# Research Questions

- What are the **most effective approaches to maximize reuse** of pre-defined metadata elements?
- How can we get metadata authors to **generate more metadata in shorter time periods**
  - To what extent can we **successfully predict metadata** values from existing metadata?
  - Can we use manuscripts, published work, user history, social media to **improve prediction**?
- To what extent can experts and crowds **find, verify, and fix metadata**?
- To what degree does **ontology-based metadata improve the precision and recall of dataset search**?

# CEDAR technology will give us

- **Tools**
  - To **author metadata templates** that define a community standard while reusing previously defined elements
  - To **facilitate the capture of experimental metadata** that conform to one or more community standards
  - To **discover datasets** that meet a plurality of dataset and experimental requirements
- **Methods**
  - To **learn metadata patterns** from unstructured, semi-structured, and structured metadata
  - To efficiently **guide predictive entry** of new metadata
  - To **index, query, and reason** about **ontology-based** metadata
  - To **explore, verify, and collaboratively augment experimental metadata** even when the data are located elsewhere

# CEDAR will empower users to meet and exceed the minimal metadata standards

- Emphasis traditionally has been on development of simple checklists of metadata elements
- Little practical consideration for
  - Using **shared value sets** (search, browse, query)
  - **Common knowledge representations** (interoperable ecosystems of tools and data)
  - **Metadata validation** against community standards (richness and quality)
- We need a more expressive—and *computable*—framework for describing metadata

# CEDAR Team

## Stanford University

- Mark A. Musen (PI)
- Michel Dumontier (Co-I)
- Olivier Gevaert (Co-I)
- Purvesh Khatri (Co-I)
- John Graybeal
- Martin O'Connor
- Marcos Martinez Romero
- Mary Panahiazar
- Attila Egyedi
- Ravi Shankar

## Stanford Library

- Kim Durante

## Oxford University

- Susanna Sansone
- Phillippe Rocca-Serra
- Alejandra Gonzalez-Beltran

## Yale University

- Steven H. Kleinstein
- Kei Cheung

## Northrop Grumman (HIPC/Immport)

- Jeff Wiser

### METADATA TEMPLATES

CREATE NEW



Search



ImmPort:  
Experiment



ImmPort:  
Protocol

### TEMPLATE ELEMENTS

CREATE NEW



Search



Objectives



Official title



Sponsoring  
organization

SEE ALL

Element Name

Study type

Element Description

Describes the nature of the study

Add Item



TEXT

PARAGRAPH

MULTIPLE CHOICE

CHECKBOX

DATE

ADD AN ELEMENT

PICK FROM LIST

INTERVENTION LONGITUDINAL

INTERVENTIONAL

LONGITUDINAL

OBSERVATIONAL

MORE ITEMS

PICK FROM A LIST

CONTROLLED TERM

AUDIO VISUAL

ADD ANOTHER



Required

Advanced

Controlled term ID

[http://bioportal.bioontology.org/ontologies/CTO?p=classes&conceptid=http%3A%2F%2Fwww.co-ode.org%2Fontologies%2Font.owl%23Study\\_type](http://bioportal.bioontology.org/ontologies/CTO?p=classes&conceptid=http%3A%2F%2Fwww.co-ode.org%2Fontologies%2Font.owl%23Study_type)







 ImmPort: Basic study design

Describes a study in terms of title, goals, endpoints, criteria for study participation, subject grouping (arms or cohorts), personnel, planned visits or encounters and protocols

Add Element

-  BRIEF TITLE
-  OBJECTIVES
-  OFFICIAL TITLE
-  SPONSORING ORGANIZATION
-  STUDY TYPE
-  ADD AN ELEMENT

 MORE ELEMENTS 





Add item

ALL ITEMS 

\* Brief title  

\* Description  

Study type

-   Intervention longitudinal
-   Interventional
-   Longitudinal
-   Observational

*Click choice to set as default*

 FAVORITE

CLEAR

SAVE TEMPLATE

## Choose a Template

### Template

-  **IMPORT: BASIC STUDY DESIGN**
-  IMPORT: EXPERIMENT
-  IMPORT: PROTOCOL
-  NEW TEMPLATE

**Brief title**  
**Susceptibility and Resistance to Common Encapsulated Bacteria Infections**



**Description**  
**To map and isolate human host supergenes that confer general susceptibility and resistance to common encapsulated bacteria infections such as pneumococcus, meningococcus, and H. influenza**



### Study type



- Intervention longitudinal
- Interventional
- Longitudinal
- Observational

**Condition studied**  
**Genetic factors conferring susceptibility or resistance to common encapsulated bacteria infections**



**Detailed description**

We also want to *help* users compose  
templates and auto-magically suggest  
metadata values

# Can we use free text to predict structured and semi-structured values?

Sample GSM1230698

Query DataSets for GSM1230698

Status Public on Oct 02, 2014

Title SNG-M\_PTX\_1

Sample type RNA

Source name SNG-M Paclitaxel 24h

Organism [Homo sapiens](#)

Characteristics cell line: SNG-M

~~cell type: endometrial cancer~~

Treatment protocol Treated with eribulin and paclitaxel at 10XIC50 conc. for 24 hours. IC50 determined in growth inhibition assay for cell line separately.

Growth protocol Cell lines were growing in growth media recommended by ATCC.

Extracted molecule total RNA

Extraction protocol Total RNA was extracted using RNeasy Mini kit (Qiagen).

Label biotin

Label protocol Biotinylated fragmenetd cRNA was used.

Hybridization protocol We used manufacture recommended protocol (Affymetrix). Arrays were washed and stained using Affymetrix Fluidics Station 450

Scan protocol Arrays were scanned using Affymetrix GeneChip Scanner 3000

Description drug

Data processing Gene chips were analyzed using Affymetrix Microarray Analysis Suite (MAS) version 5. RMA normalization was performed using Affymetrix Power Tools version 1.12.0

# Predicting structured metadata (accuracy)

Classifier	GPL	Type	Organism	Molecule	Label
LDA => SVM	32.69%	88.98%	<b>87.43%</b>	94.87%	87.64%
LDA => DecisionTree	<b>73.30%</b>	<b>95.45%</b>	86.80%	<b>95.01%</b>	<b>88.00%</b>

*Accuracy = % correctly classified samples*

## Predicting semi-structured metadata

Using multi label tree

Accuracy of 72% to predict 39 most-occurring keys

Average precision 79% (for all keys)

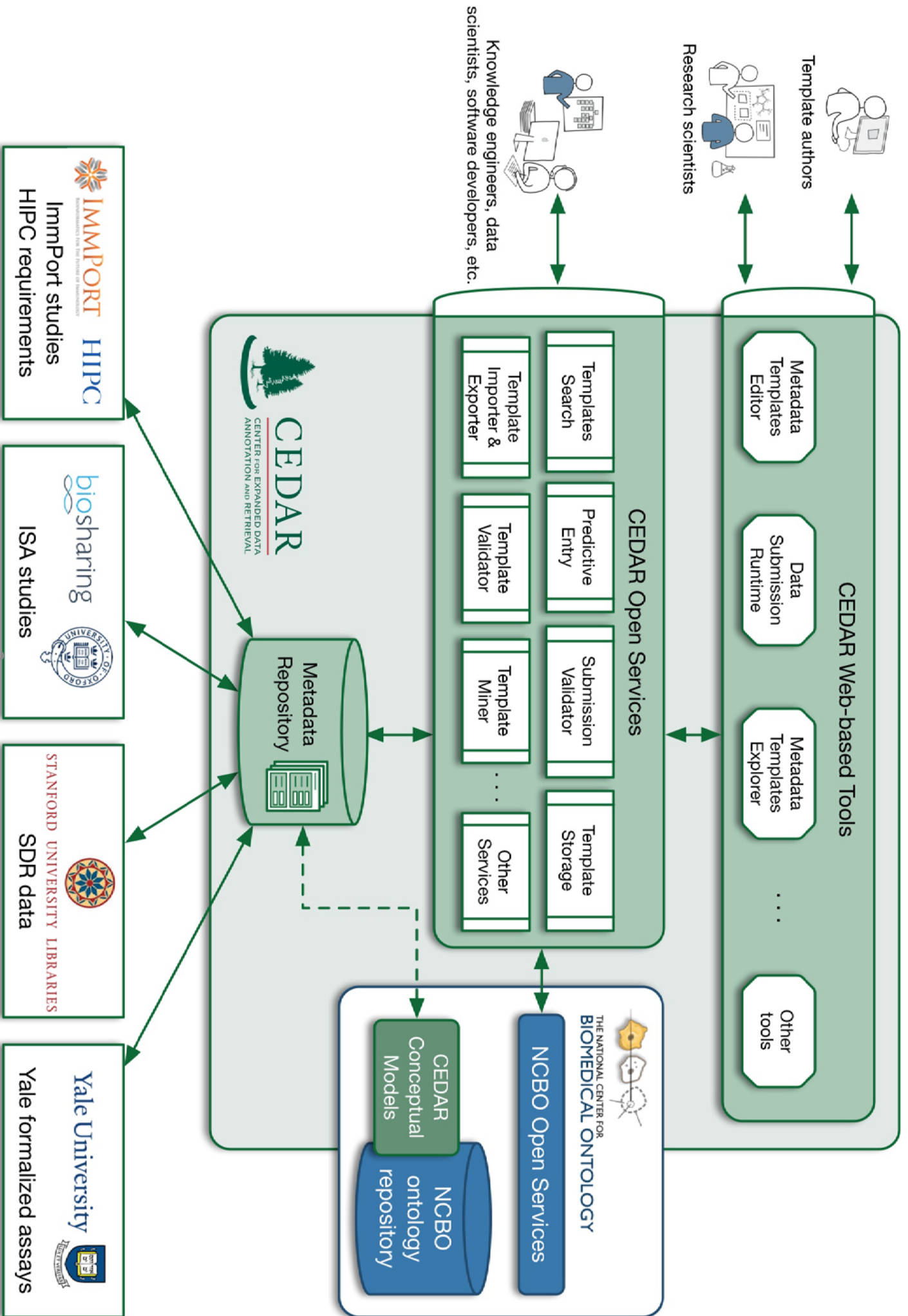
Average recall 82% (for all keys)

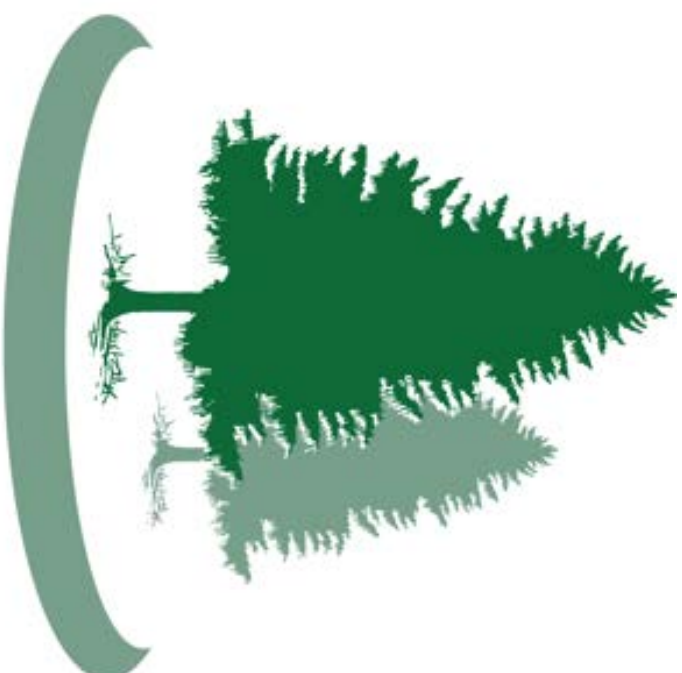
# Metadata: The Next Frontier

We welcome new partners to

- Learn about **your** metadata authoring workflow
- Evaluate CEDAR technology
- Incorporate this technology into your curation workflows

# [metadatacenter.org](http://metadatacenter.org)





# CEDAR

CENTER FOR EXPANDED DATA ANNOTATION  
AND RETRIEVAL

<http://metadatatcenter.org>