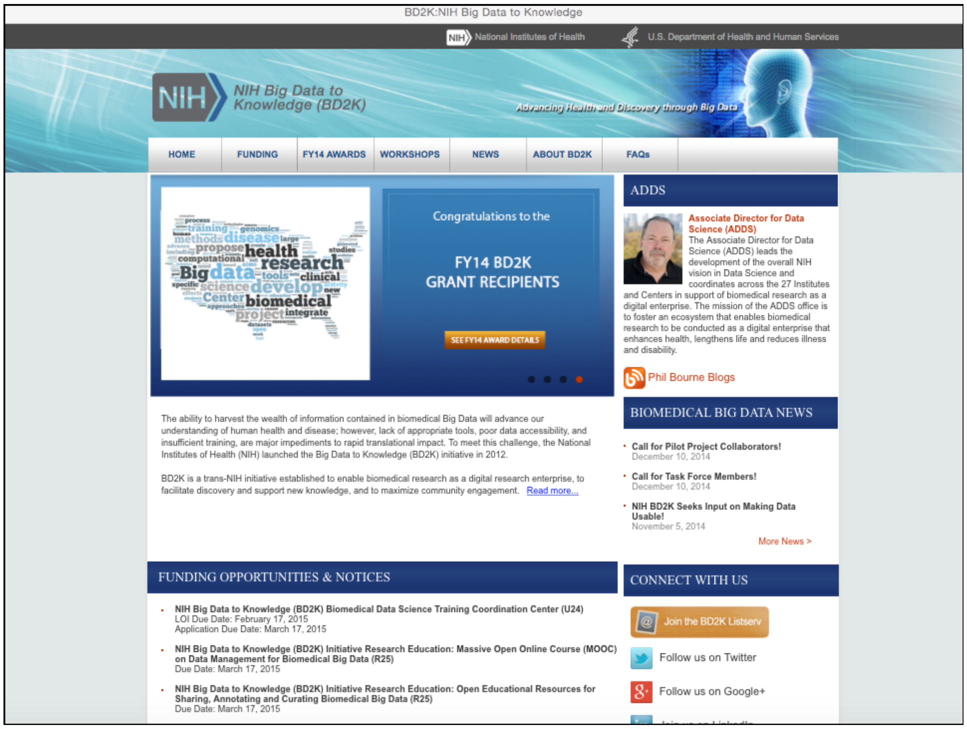
CEDAR

Making it Easier to Use Ontologies to Author Experimental Metadata

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

CEDAR
CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL


BD2K:NIH Big Data to Knowledge
NIH National Institutes of Health
U.S. Department of Health and Human Services
NIH
NIH Big Data to Knowledge (BD2K)
Advancing Health and Discovery through Big Data
HOME
FUNDING
FY14 AWARDS
WORKSHOPS
NEWS
ABOUT BD2K
FAQs
Congratulations to the
FY14 BD2K GRANT RECIPIENTS
SEE FY14 AWARD DETAILS
Phil Bourne Blogs
More News >
Join the BD2K Listserv
Follow us on Twitter
Follow us on Google+

The ability to harvest the wealth of information contained in biomedical Big Data will advance our understanding of human health and disease; however, lack of appropriate tools, poor data accessibility, and insufficient training, are major impediments to rapid translational impact. To meet this challenge, the National Institutes of Health (NIH) launched the Big Data to Knowledge (BD2K) initiative in 2012.

BD2K is a trans-NIH initiative established to enable biomedical research as a digital research enterprise, to facilitate discovery and support new knowledge, and to maximize community engagement. Read more...

ADDS

Associate Director for Data Science (ADDS)
The Associate Director for Data Science (ADDS) leads the development of the overall NIH vision in Data Science and coordinates across the 27 Institutes and Centers in support of biomedical research as a digital enterprise. The mission of the ADDS office is to foster an ecosystem that enables biomedical research to be conducted as a digital enterprise that enhances health, lengthens life and reduces illness and disability.

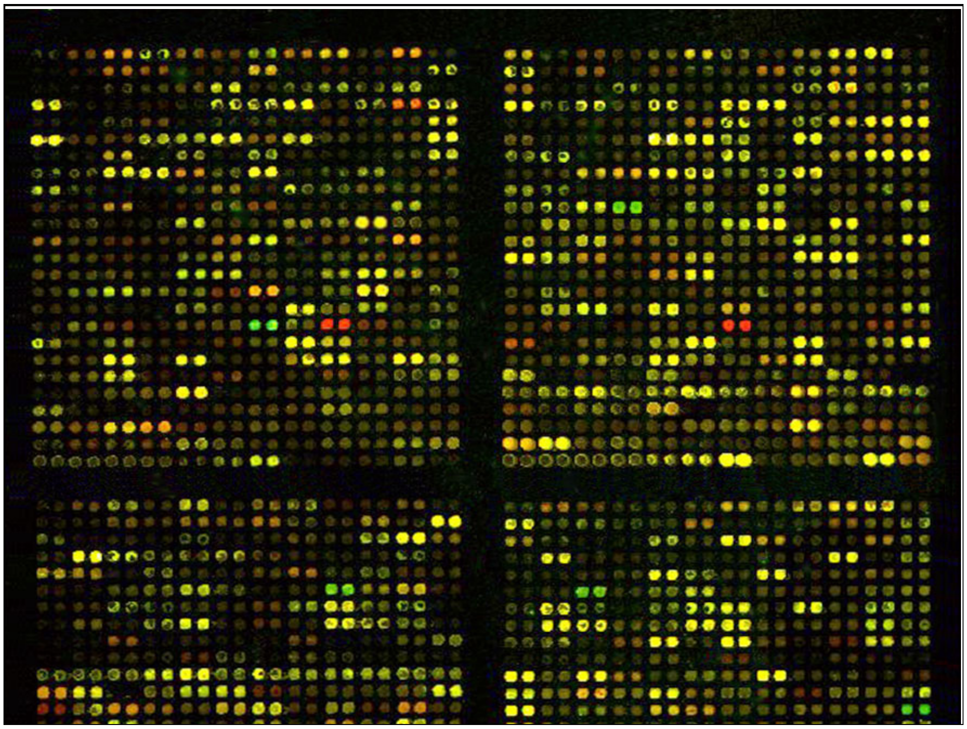
BIOMEDICAL BIG DATA NEWS

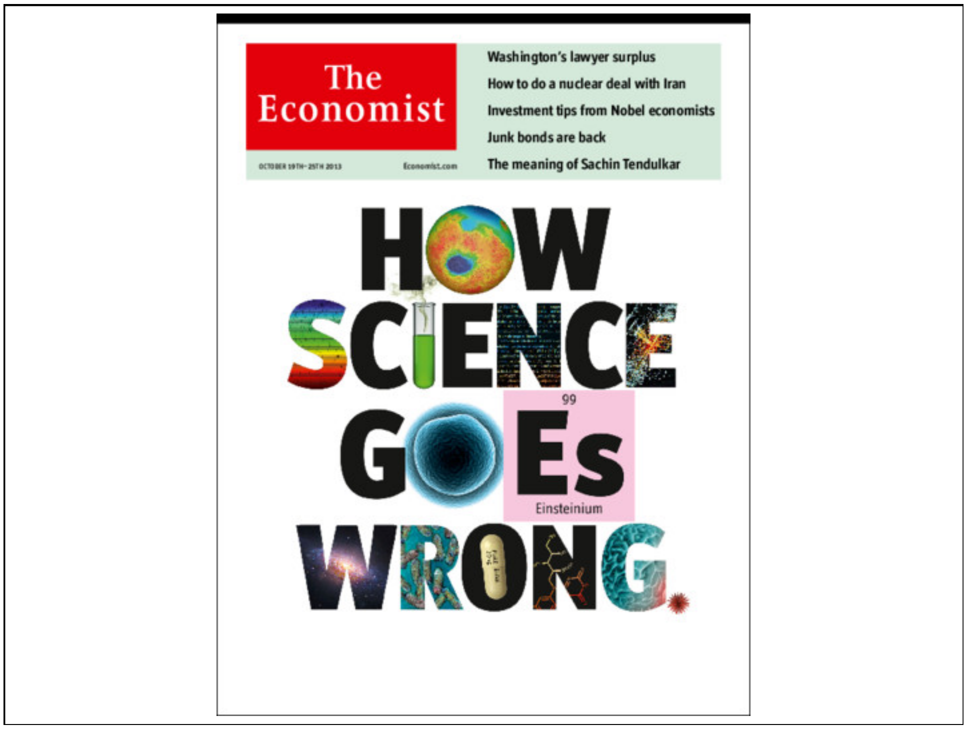
- **Call for Pilot Project Collaborators!**
  December 10, 2014
- **Call for Task Force Members!**
  December 10, 2014
- **NIH BD2K Seeks Input on Making Data Usable!**
  November 5, 2014

FUNDING OPPORTUNITIES & NOTICES

- **NIH Big Data to Knowledge (BD2K) Biomedical Data Science Training Coordination Center (U24)**
  LOI Due Date: February 17, 2015
  Application Due Date: March 17, 2015
- **NIH Big Data to Knowledge (BD2K) Initiative Research Education: Massive Open Online Course (MOOC) on Data Management for Biomedical Big Data (R25)**
  Due Date: March 17, 2015
- **NIH Big Data to Knowledge (BD2K) Initiative Research Education: Open Educational Resources for Sharing, Annotating and Curating Biomedical Big Data (R25)**
  Due Date: March 17, 2015

CONNECT WITH US





The Economist

OCTOBER 19TH–25TH 2013

Economist.com

Washington's lawyer surplus

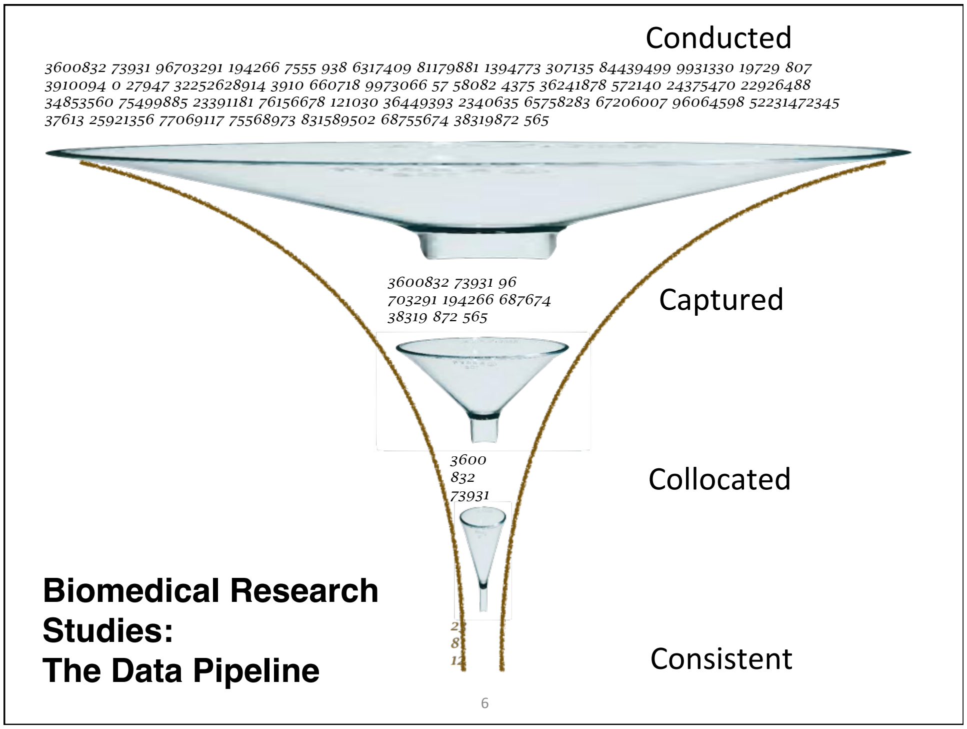
How to do a nuclear deal with Iran

Investment tips from Nobel economists

Junk bonds are back

The meaning of Sachin Tendulkar

HOW SCIENCE GOES WRONG



Conducted

3600832 73931 96703291 194266 7555 938 6317409 81179881 1394773 307135 84439499 9931330 19729 807 3910094 0 27947 32252628914 3910 660718 9973066 57 58082 4375 36241878 572140 24375470 22926488 34853560 75499885 23391181 76156678 121030 36449393 2340635 65758283 67206007 96064598 52231472345 37613 25921356 77069117 75568973 831589502 68755674 38319872 565

Captured

3600832 73931 96 703291 194266 687674 38319 872 565

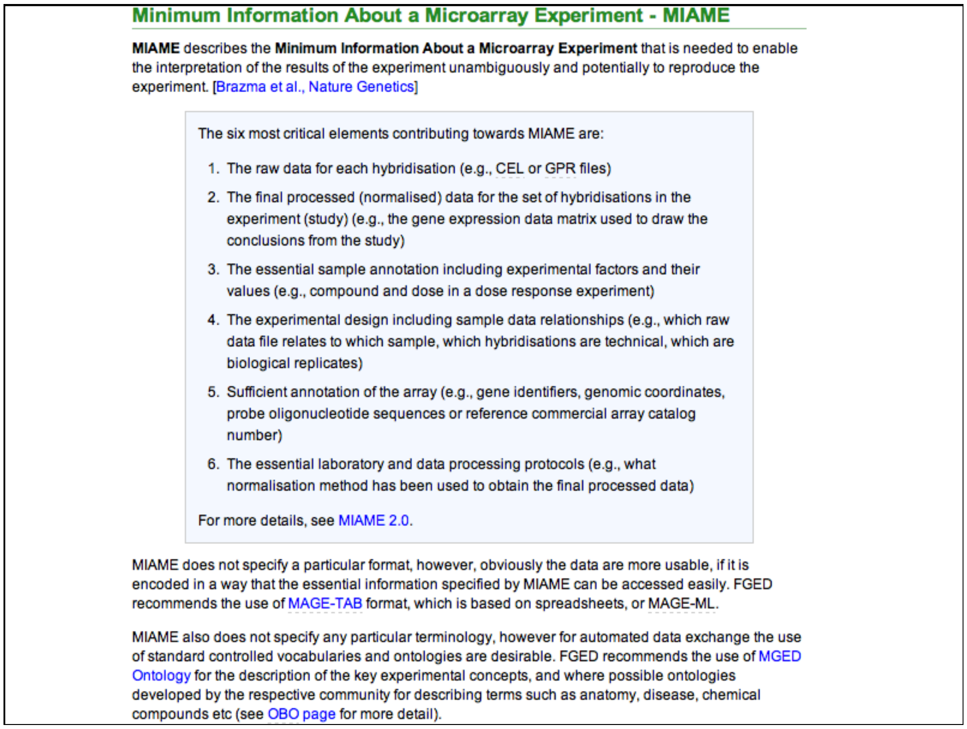
Collocated

3600 832 73931

Consistent

2 8 12

Biomedical Research Studies: The Data Pipeline



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)

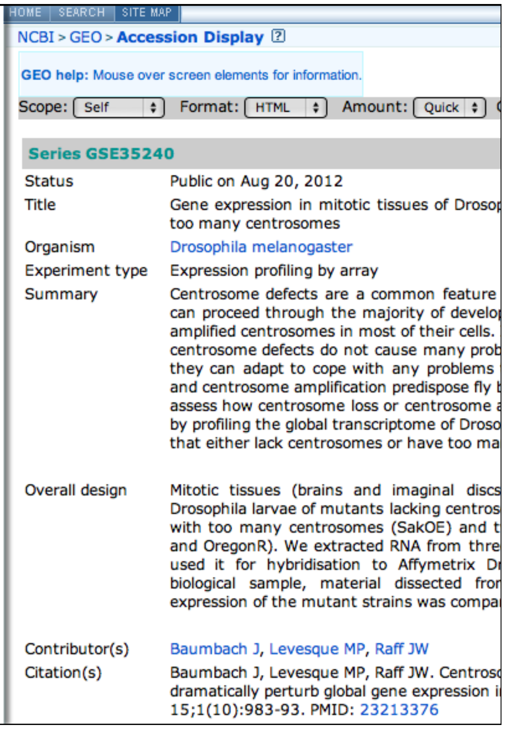
For more details, see MIAME 2.0.

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).

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



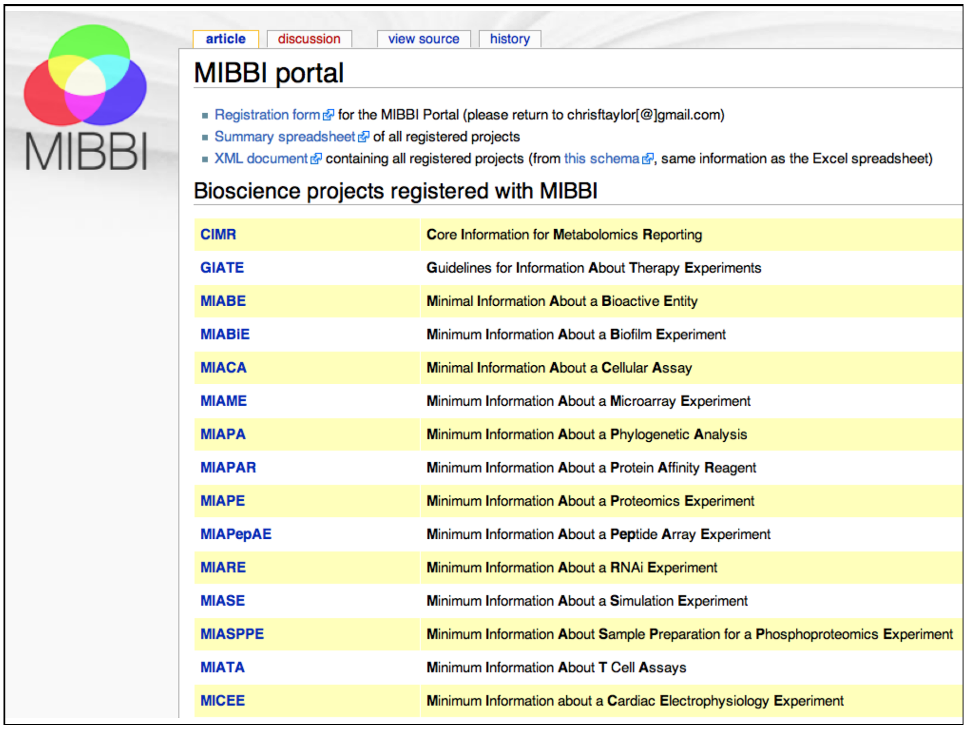
NCBI > GEO > **Accession Display**

GEO help: Mouse over screen elements for information.

Scope: Self Format: HTML Amount: Quick

Series GSE35240

Status	Public on Aug 20, 2012
Title	Gene expression in mitotic tissues of Drosop 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 develo amplified centrosomes in most of their cells. centrosome defects do not cause many prob they can adapt to cope with any problems and centrosome amplification predispose fly assess how centrosome loss or centrosome by profiling the global transcriptome of Droso that either lack centrosomes or have too ma
Overall design	Mitotic tissues (brains and imaginal discs Drosophila larvae of mutants lacking centros with too many centrosomes (SakOE) and t and OregonR). We extracted RNA from thre used it for hybridisation to Affymetrix D biological sample, material dissected fro expression of the mutant strains was compa
Contributor(s)	Baumbach J, Levesque MP, Raff JW
Citation(s)	Baumbach J, Levesque MP, Raff JW. Centros dramatically perturb global gene expression i 15;1(10):983-93. PMID: 23213376

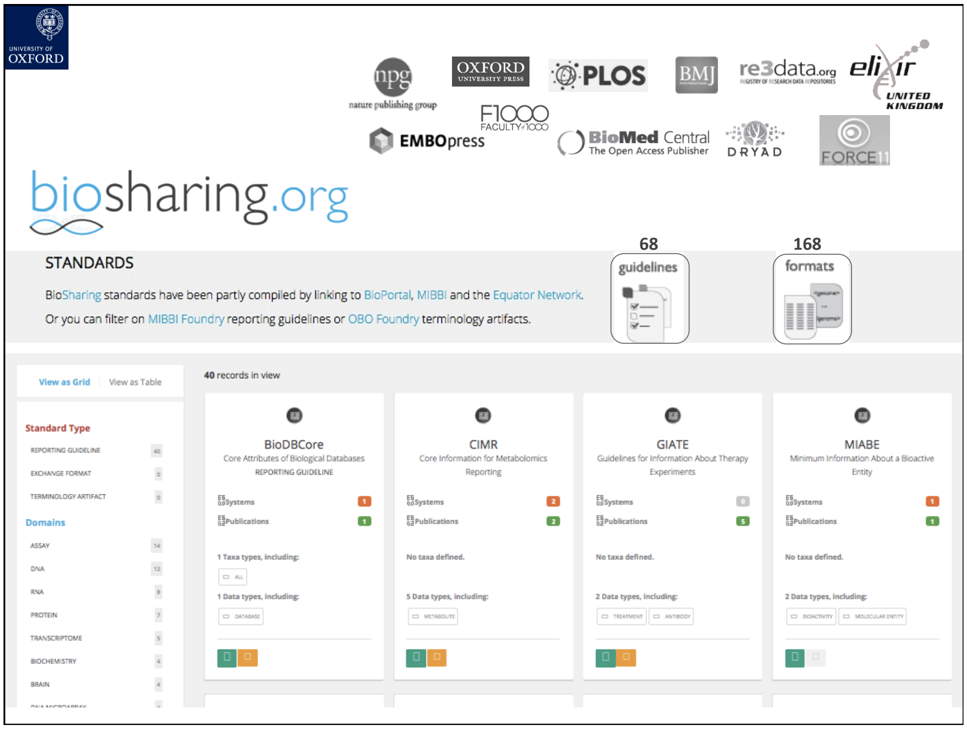


MIBBI portal

- Registration form for the MIBBI Portal (please return to chrisftaylor[@]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



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

40 records in view

Standard Type	
REPORTING GUIDELINE	40
EXCHANGE FORMAT	0
TERMINOLOGY ARTIFACT	0
Domains	
ASSAY	14
DNA	12
RNA	8
PROTEIN	7
TRANSCRIPTOME	5
BIOCHEMISTRY	4
BRAIN	4
DNA MICROARRAY	4

BioDBCore	CIMR	GIATE	MIABE
Core Attributes of Biological Databases REPORTING GUIDELINE	Core Information for Metabolomics Reporting	Guidelines for Information About Therapy Experiments	Minimum Information About a Bioactive Entity
Systems 1	Systems 2	Systems 0	Systems 1
Publications 1	Publications 2	Publications 5	Publications 1
1 Taxa types, including: ALL	No taxa defined.	No taxa defined.	No taxa defined.
1 Data types, including: DATABASE	5 Data types, including: METABOLITE	2 Data types, including: TREATMENT ANTIBODY	2 Data types, including: BIOACTIVITY MOLECULAR ENTITY

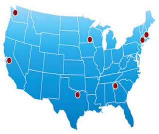
The Existing BioSharing Approach is Limited

- Emphasis traditionally has been on development of simple checklists of metadata elements
- Little practical consideration of
  - How to supply values for the metadata elements
  - Standard ontologies that might be used
- We need a more expressive—and *computable*—framework for describing metadata

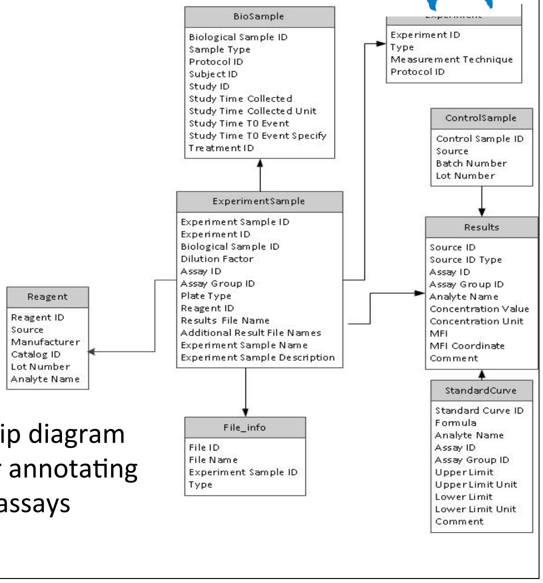
The ISA model

- Developed by BioSharing group and supported by a suite of tools
- Provides structure for metadata related to
  - Investigation
  - Study
  - Assay
- Is not easily extended within existing tool set
- Forms the foundation for the modeling of metadata in the CEDAR project

Human Immunology Project Consortium



HIPC centers submit data directly to the NIAID ImmPort data repository using detailed, experiment-specific templates



BioSample
Biological Sample ID
Sample Type
Protocol ID
Subject ID
Study ID
Study Time Collected
Study Time Collected Unit
Study Time T0 Event
Study Time T0 Event Specify
Treatment ID

Experiment ID
Type
Measurement Technique
Protocol ID

ControlSample
Control Sample ID
Source
Batch Number
Lot Number

ExperimentSample
Experiment Sample ID
Experiment ID
Biological Sample ID
Dilution Factor
Assay ID
Assay Group ID
Plate Type
Reagent ID
Results File Name
Additional Result File Names
Experiment Sample Name
Experiment Sample Description

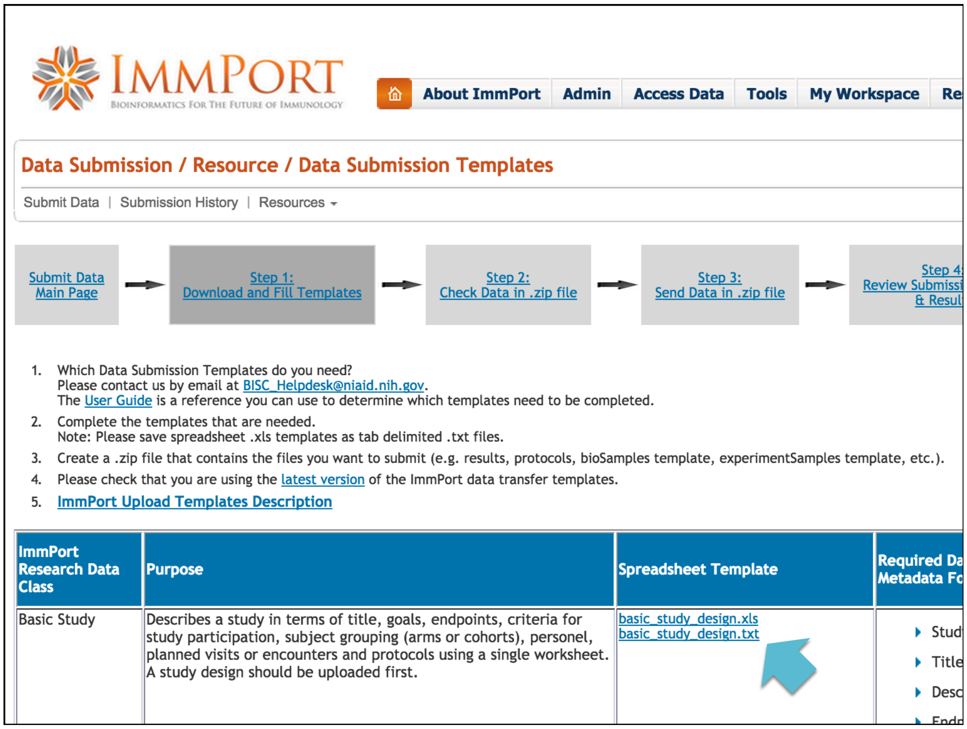
Results
Source ID
Source ID Type
Assay ID
Assay Group ID
Analyte Name
Concentration Value
Concentration Unit
MFI
MFI Coordinate
Comment

Reagent
Reagent ID
Source
Manufacturer
Catalog ID
Lot Number
Analyte Name

File_info
File ID
File Name
Experiment Sample ID
Type

StandardCurve
Standard Curve ID
Formula
Analyte Name
Assay ID
Assay Group ID
Upper Limit
Upper Limit Unit
Lower Limit
Lower Limit Unit
Comment

Example entity–relationship diagram for describing metadata for annotating multiplex bead array assays (e.g., Luminex)



IMMPORT
BIOINFORMATICS FOR THE FUTURE OF IMMUNOLOGY
About ImmPort
Admin
Access Data
Tools
My Workspace
Re

Data Submission / Resource / Data Submission Templates

Submit Data | Submission History | Resources
Submit Data Main Page
Step 1: Download and Fill Templates
Step 2: Check Data in .zip file
Step 3: Send Data in .zip file
Step 4: Review Submissi & Resul

1. Which Data Submission Templates do you need?
   Please contact us by email at BISC_Helpdesk@niaid.nih.gov.
   The User Guide is a reference you can use to determine which templates need to be completed.
2. Complete the templates that are needed.
   Note: Please save spreadsheet .xls templates as tab delimited .txt files.
3. Create a .zip file that contains the files you want to submit (e.g. results, protocols, bioSamples template, experimentSamples template, etc.).
4. Please check that you are using the latest version of the ImmPort data transfer templates.
5. ImmPort Upload Templates Description

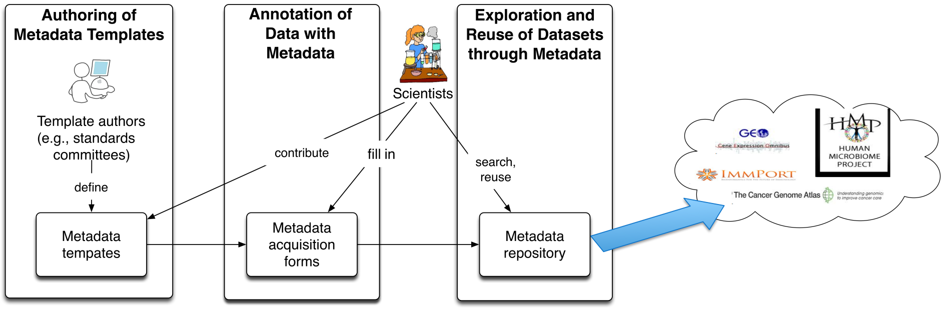
ImmPort Research Data Class	Purpose	Spreadsheet Template	Required Da Metadata Fo
Basic Study	Describes a study in terms of title, goals, endpoints, criteria for study participation, subject grouping (arms or cohorts), personel, planned visits or encounters and protocols using a single worksheet. A study design should be uploaded first.	basic_study_design.xls basic_study_design.txt	▸ Stud ▸ Title ▸ Desc ▸ Endp

A Metadata Ecosystem

CEDAR
CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL

- **HIPC investigators** perform experiments in human immunology
- **HIPC Standards Working Group** creates metadata templates to annotate experimental data in a uniform manner
- **ImmPort** stores HIPC data (and metadata) in its public repository
- CEDAR will ease
  - Template creation and management
  - The use of templates to author metadata for ImmPort
  - Analysis of existing metadata to inform the authoring of new metadata

The CEDAR Approach to Metadata



Authoring of Metadata Templates

Template authors (e.g., standards committees)

define

Metadata tempates

Annotation of Data with Metadata

Metadata acquisition forms

Scientists

contribute

fill in

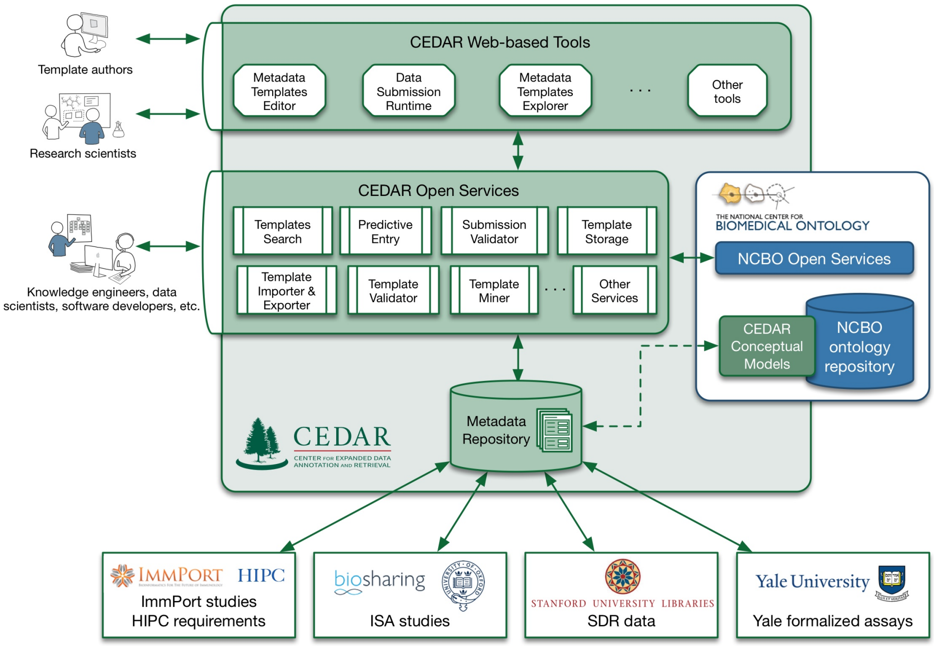
Exploration and Reuse of Datasets through Metadata

search, reuse

Metadata repository

CEDAR technology will give us

- Mechanisms
  - To author metadata template elements
  - To assemble them into composite templates
  - 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 the National Center for Biomedical Ontology to ensure that metadata are encoded using appropriate ontology terms



Template authors

Research scientists

Knowledge engineers, data scientists, software developers, etc.

CEDAR Web-based Tools

Metadata Templates Editor

Data Submission Runtime

Metadata Templates Explorer

. . .

Other tools

CEDAR Open Services

Templates Search

Predictive Entry

Submission Validator

Template Storage

Template Importer & Exporter

Template Validator

Template Miner

. . .

Other Services

THE NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY

NCBO Open Services

CEDAR Conceptual Models

NCBO ontology repository

Metadata Repository

CEDAR CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL

IMMPORT HIPC

ImmPort studies HIPC requirements

biosharing

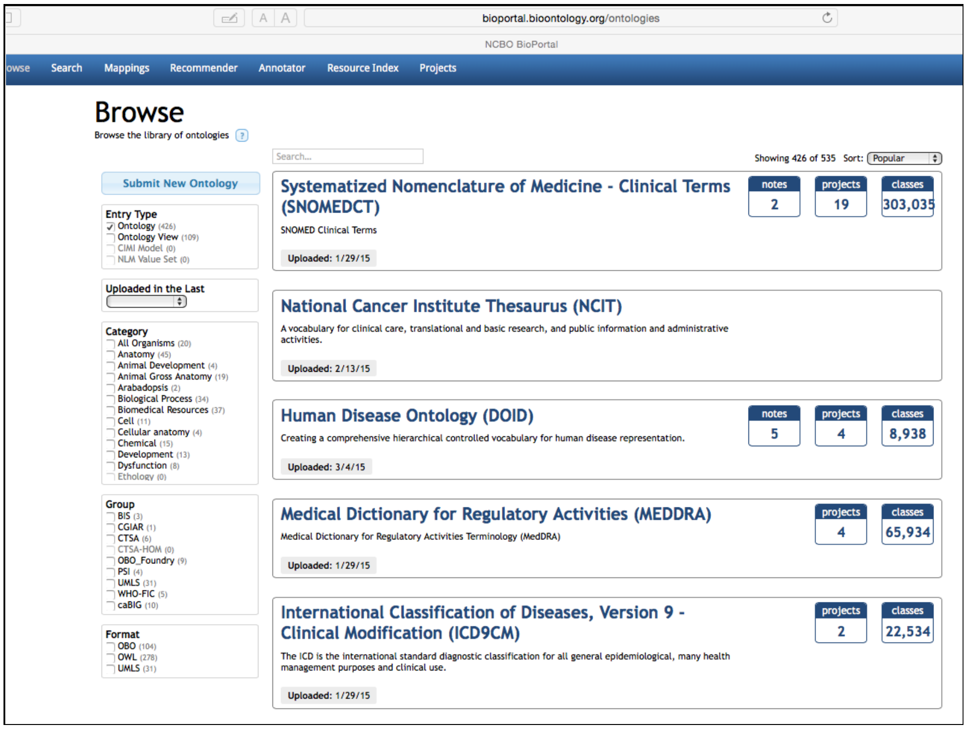
ISA studies

STANFORD UNIVERSITY LIBRARIES

SDR data

Yale University

Yale formalized assays



Browse

Browse the library of ontologies ?

Search...

Showing 426 of 535 Sort: Popular

Submit New Ontology

Entry Type
☑ Ontology (426)
☐ Ontology View (109)
☐ CIMI Model (0)
☐ NLM Value Set (0)

Uploaded in the Last

Category
☐ All Organisms (20)
☐ Anatomy (45)
☐ Animal Development (4)
☐ Animal Gross Anatomy (19)
☐ Arabadopsis (2)
☐ Biological Process (34)
☐ Biomedical Resources (37)
☐ Cell (11)
☐ Cellular anatomy (4)
☐ Chemical (15)
☐ Development (13)
☐ Dysfunction (8)
☐ Ethology (0)

Group
☐ BIS (3)
☐ CGIAR (1)
☐ CTSA (6)
☐ CTSA-HOM (0)
☐ OBO_Foundry (9)
☐ PSI (4)
☐ UMLS (31)
☐ WHO-FIC (5)
☐ caBIG (10)

Format
☐ OBO (104)
☐ OWL (278)
☐ UMLS (31)

Systematized Nomenclature of Medicine - Clinical Terms (SNOMEDCT)

SNOMED Clinical Terms

Uploaded: 1/29/15

notes	projects	classes
2	19	303,035

National Cancer Institute Thesaurus (NCIT)

A vocabulary for clinical care, translational and basic research, and public information and administrative activities.

Uploaded: 2/13/15

Human Disease Ontology (DOID)

Creating a comprehensive hierarchical controlled vocabulary for human disease representation.

Uploaded: 3/4/15

notes	projects	classes
5	4	8,938

Medical Dictionary for Regulatory Activities (MEDDRA)

Medical Dictionary for Regulatory Activities Terminology (MedDRA)

Uploaded: 1/29/15

projects	classes
4	65,934

International Classification of Diseases, Version 9 - Clinical Modification (ICD9CM)

The ICD is the international standard diagnostic classification for all general epidemiological, many health management purposes and clinical use.

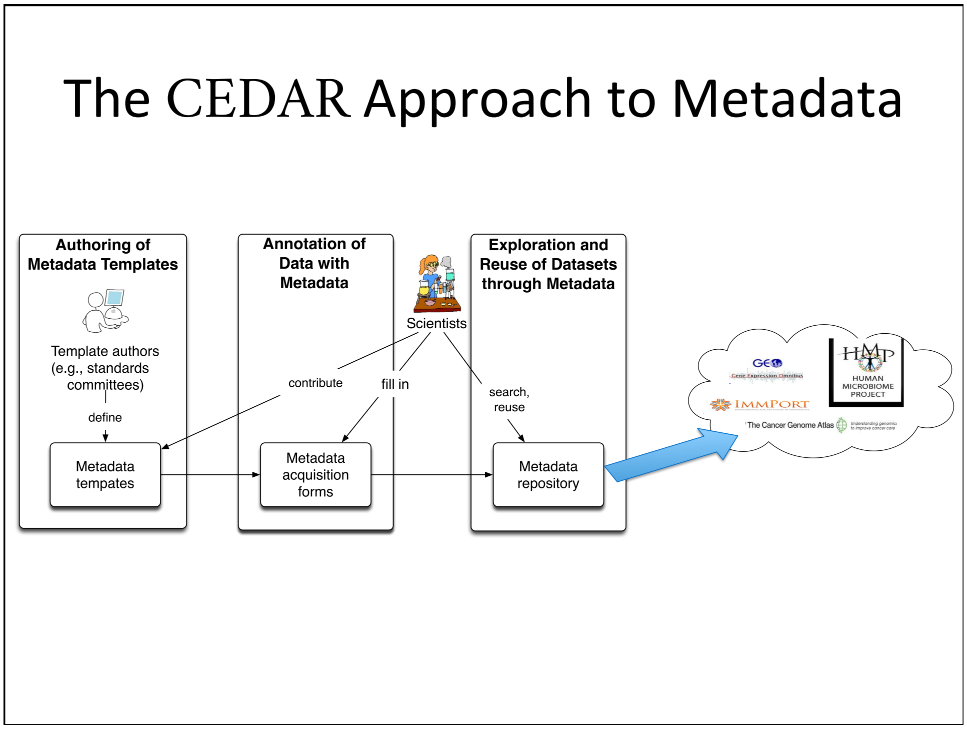
Uploaded: 1/29/15

projects	classes
2	22,534

The National Center for Biomedical Ontology

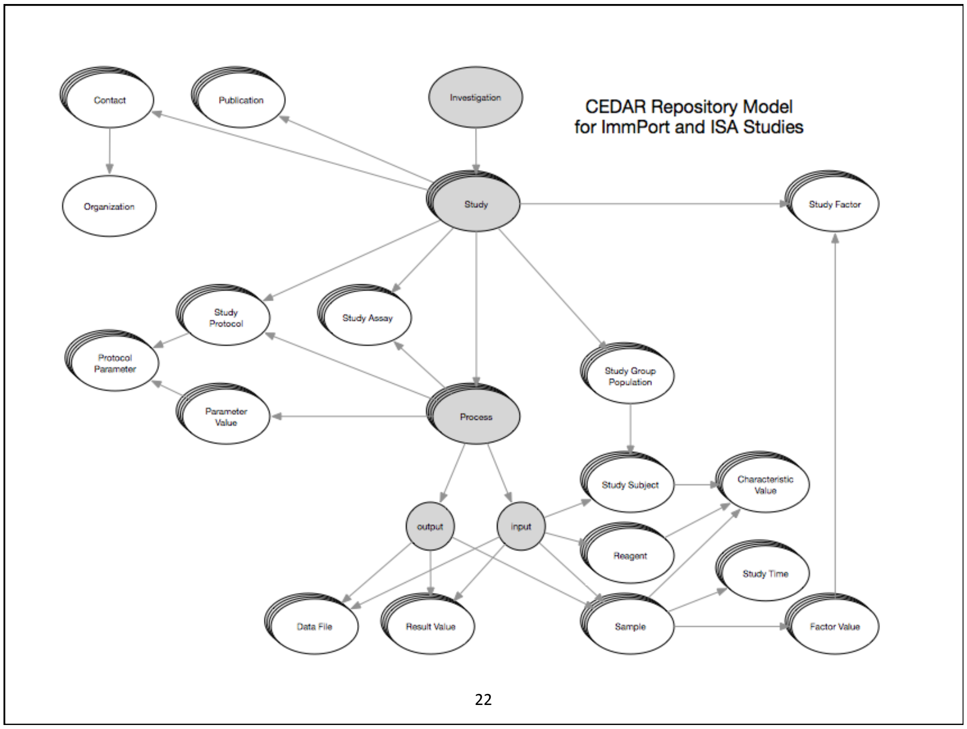
- We **create and maintain a library** of biomedical ontologies and terminologies.
- We **build tools and Web services** to enable the use of ontologies and terminologies.
- We **collaborate with scientific communities** that develop and use ontologies and terminologies in biomedicine.

NATIONAL CENTER FOR
BIOMEDICAL ONTOLOGY



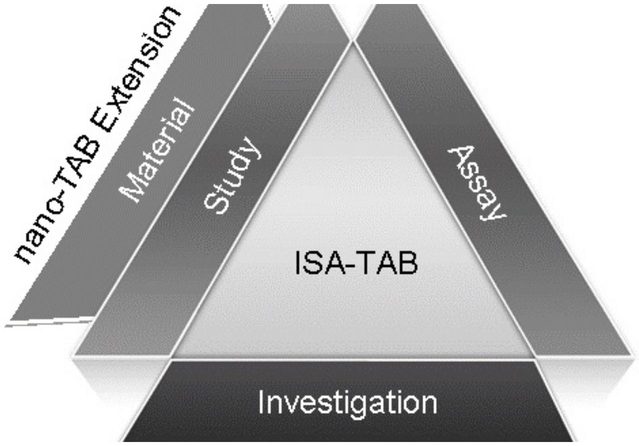
The CEDAR Approach to Metadata

Authoring of Metadata Templates
Template authors (e.g., standards committees)
define
Metadata tempates
contribute
Annotation of Data with Metadata
Scientists
fill in
Metadata acquisition forms
Exploration and Reuse of Datasets through Metadata
search, reuse
Metadata repository
GEO
Gene Expression Omnibus
HMP
HUMAN MICROBIOME PROJECT
IMMPORT
The Cancer Genome Atlas
Understanding genomics to improve cancer care

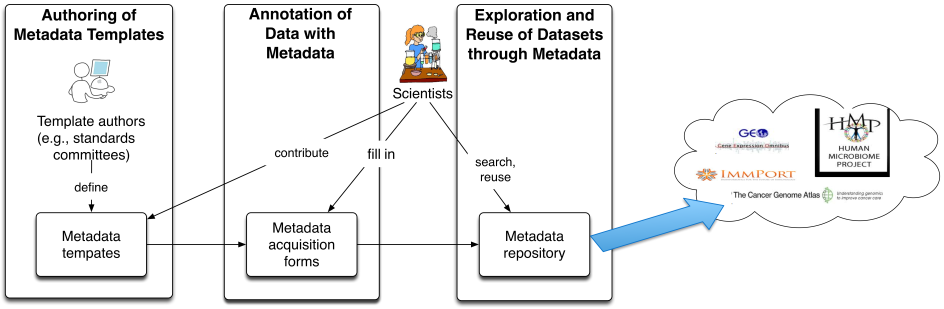


CEDAR Repository Model
for ImmPort and ISA Studies

22



The CEDAR Approach to Metadata



Authoring of Metadata Templates

Template authors (e.g., standards committees)

define

Metadata tempates

Annotation of Data with Metadata

contribute

fill in

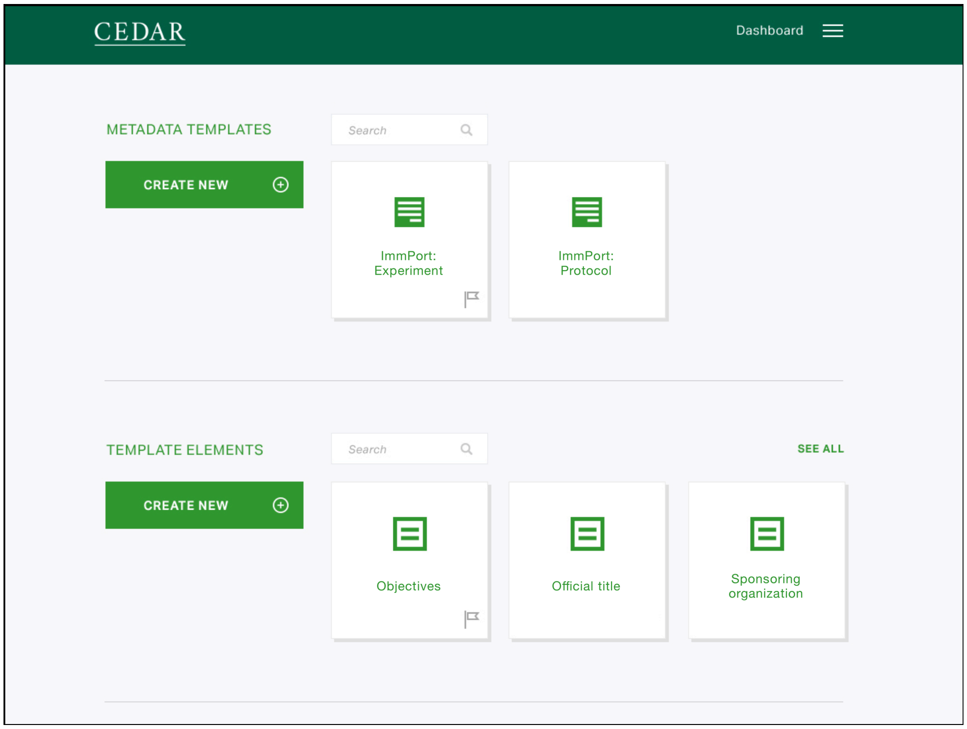
Metadata acquisition forms

Scientists

Exploration and Reuse of Datasets through Metadata

search, reuse

Metadata repository



CEDAR

Dashboard

METADATA TEMPLATES

Search

CREATE NEW

ImmPort: Experiment

ImmPort: Protocol

TEMPLATE ELEMENTS

Search

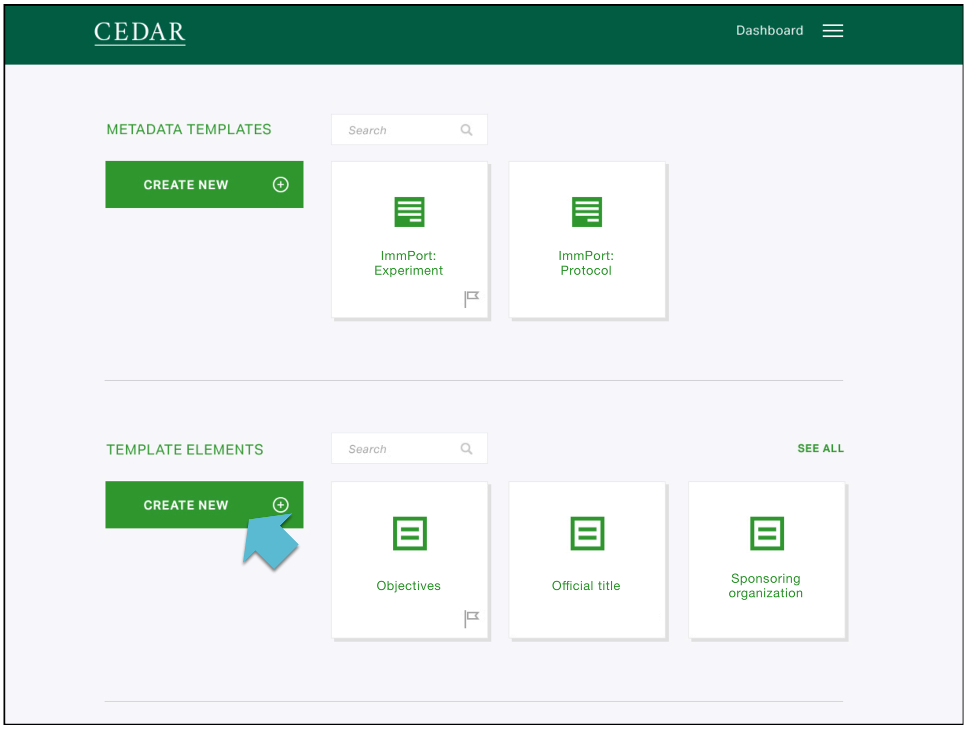
SEE ALL

CREATE NEW

Objectives

Official title

Sponsoring organization



CEDAR

Dashboard

METADATA TEMPLATES

Search

CREATE NEW

ImmPort: Experiment

ImmPort: Protocol

TEMPLATE ELEMENTS

Search

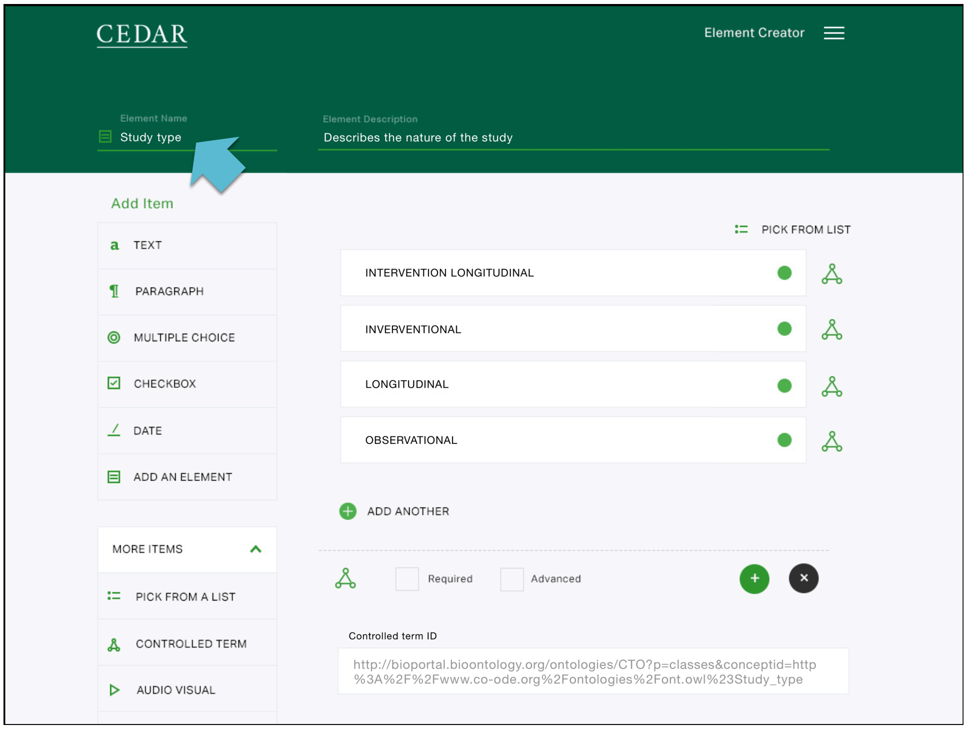
SEE ALL

CREATE NEW

Objectives

Official title

Sponsoring organization

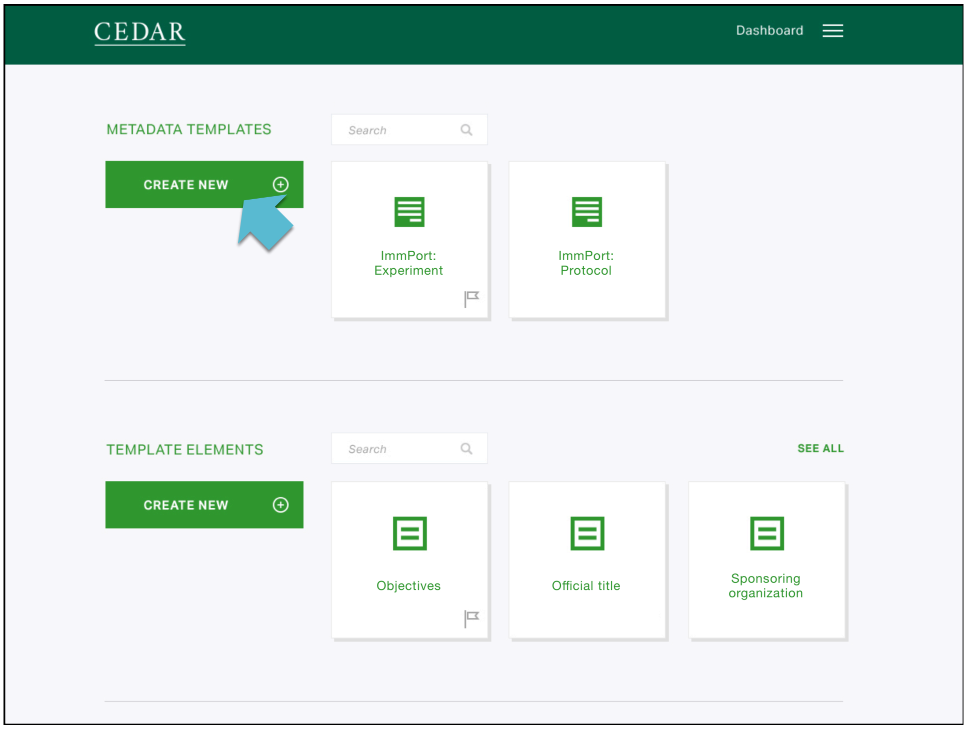


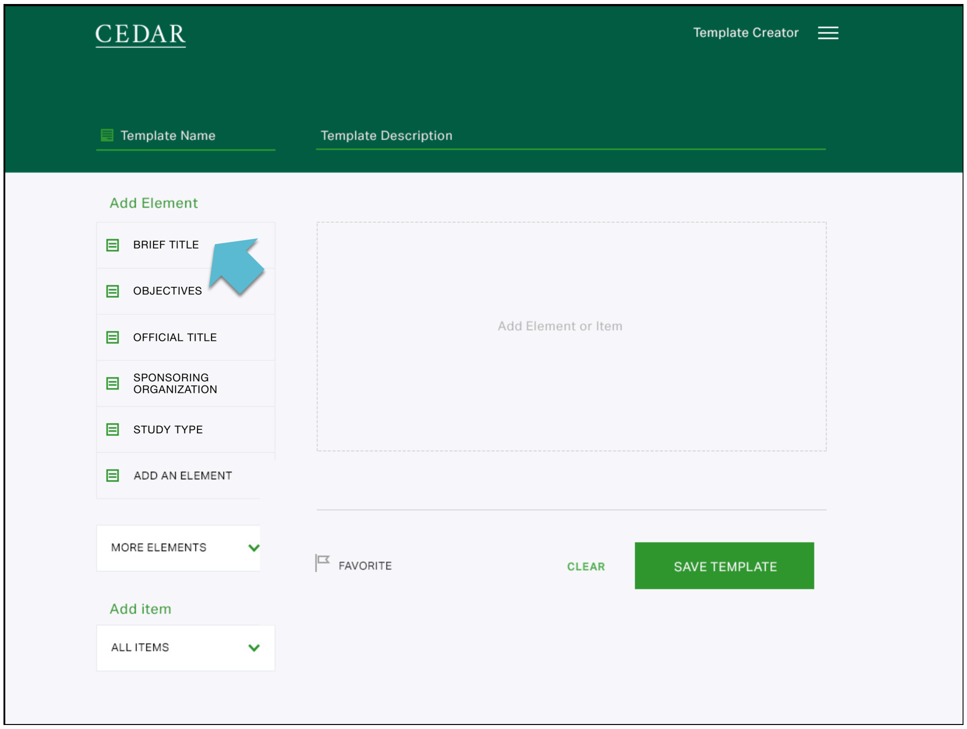
Element Name: Study type

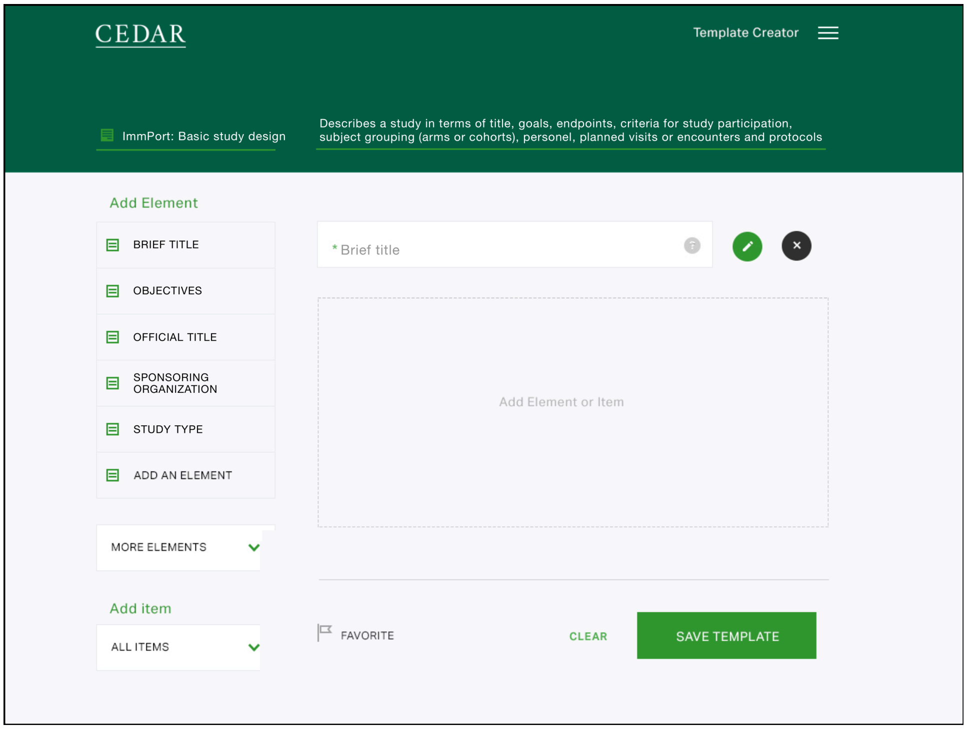
Element Description: Describes the nature of the study

- INTERVENTION LONGITUDINAL
- INVERVENTIONAL
- LONGITUDINAL
- OBSERVATIONAL

Controlled term ID: http://bioportal.bioontology.org/ontologies/CTO?p=classes&conceptid=http%3A%2F%2Fwww.co-ode.org%2Fontologies%2Font.owl%23Study_type



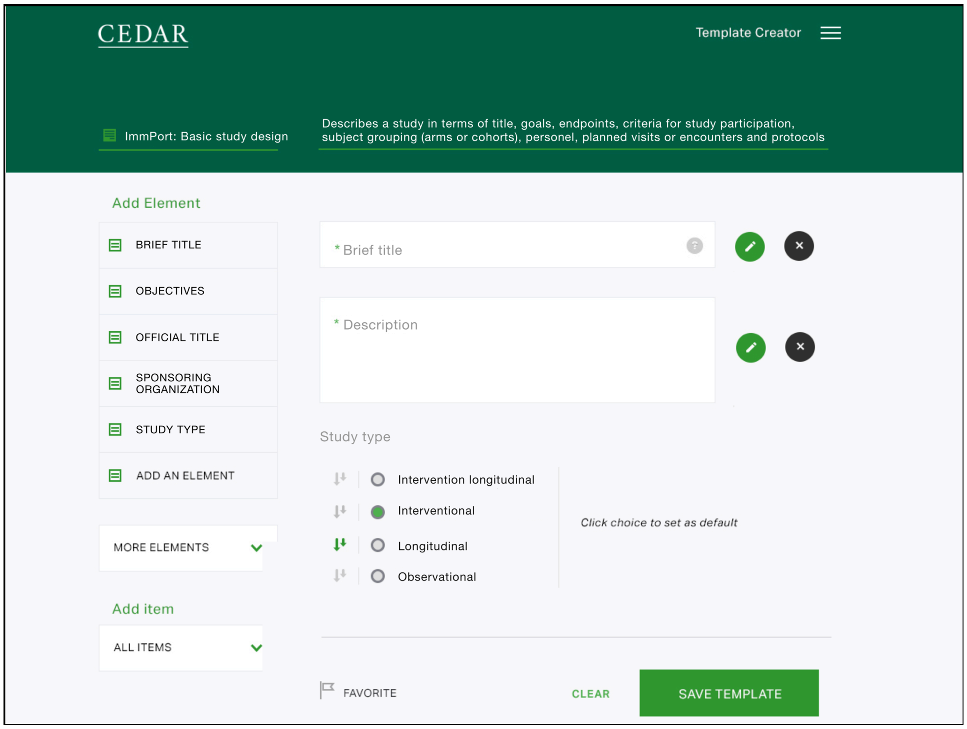




ImmPort: Basic study design

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

* Brief title



CEDAR

Template Creator

ImmPort: Basic study design

Describes a study in terms of title, goals, endpoints, criteria for study participation, subject grouping (arms or cohorts), personel, 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

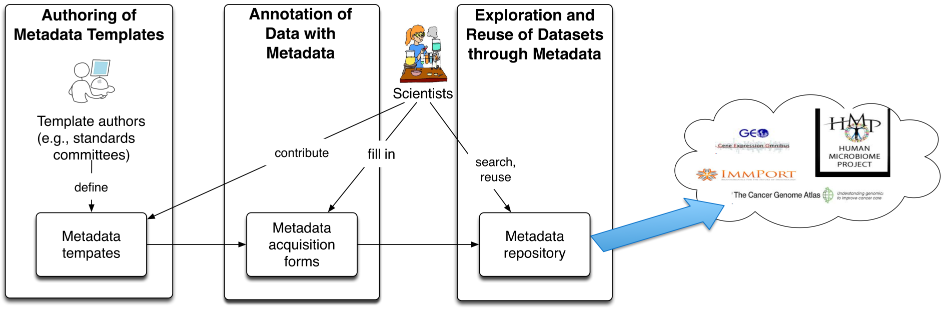
Click choice to set as default

FAVORITE

CLEAR

SAVE TEMPLATE

The CEDAR Approach to Metadata



Authoring of Metadata Templates

Template authors (e.g., standards committees)

define

Metadata tempates

Annotation of Data with Metadata

Scientists

contribute

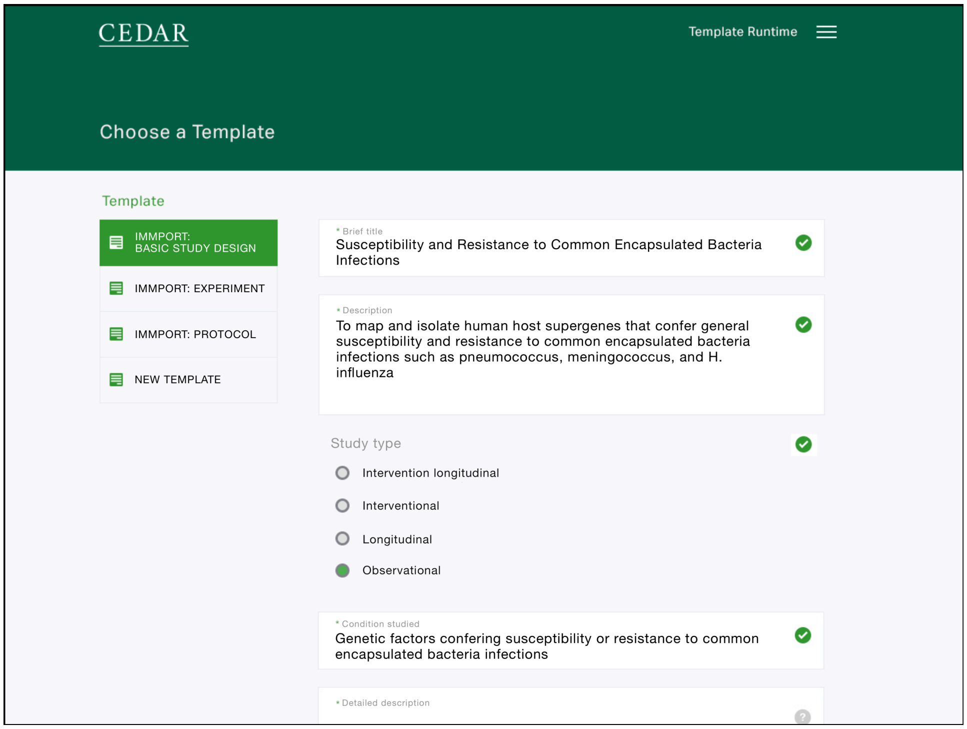
fill in

Metadata acquisition forms

Exploration and Reuse of Datasets through Metadata

search, reuse

Metadata repository



CEDAR

Template Runtime

Choose a Template

Template

- IMMPORT: BASIC STUDY DESIGN
- IMMPORT: EXPERIMENT
- IMMPORT: 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

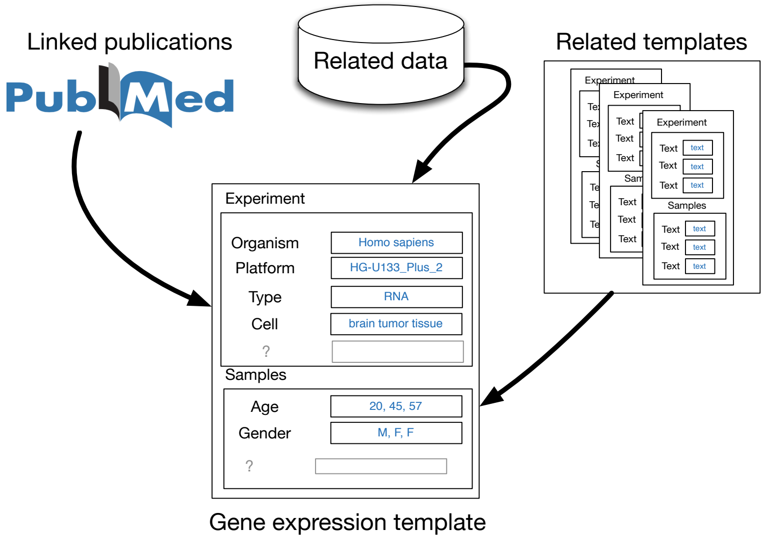
○ Longitudinal

● Observational

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

* Detailed description

Learning for Predictive Metadata Entry



Experiment

Organism	Homo sapiens
Platform	HG-U133_Plus_2
Type	RNA
Cell	brain tumor tissue
?	

Samples

Age	20, 45, 57
Gender	M, F, F
?	

Gene expression template

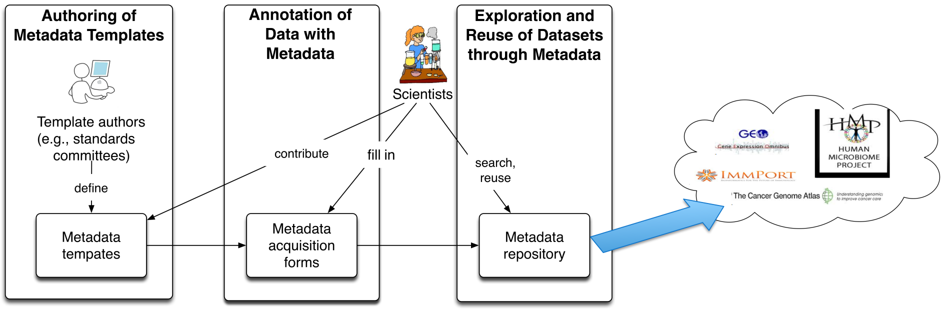
How can we make metadata authoring better?

- Create an ecosystem based on searchable, "smart" metadata templates
- Predefine standard value sets to fill in the blanks
- Use machine learning to enable predictive metadata entry
- Use text processing to acquire metadata from written descriptions of the experiment (e.g., from PubMed and PubMed Central)

How can we make metadata themselves better?

- Mirror metadata authored with CEDAR tools in our own metadata repository
- Augment those metadata with links to the published literature (including secondary analyses and retractions!)
- Augment those metadata with links to follow-up experiments (in online databases and in the literature)
- Allow the scientific community to comment on the experiment through structured metadata
- Learn from the metadata repository to ease the authoring of new metadata

The CEDAR Approach to Metadata



Authoring of Metadata Templates

Template authors (e.g., standards committees)

define

Metadata tempates

Annotation of Data with Metadata

Scientists

contribute

fill in

Metadata acquisition forms

Exploration and Reuse of Datasets through Metadata

search, reuse

Metadata repository

http://metadatacenter.org

CEDAR

CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL