

# The Center for Expanded Data Annotation and Retrieval

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

Stanford University  
musen@Stanford.EDU



# CEDAR Partners

- Stanford University School of Medicine
- University of Oxford e-Science Centre
- Yale University School of Medicine
- Northrup Grumman Corporation
- Stanford University Libraries

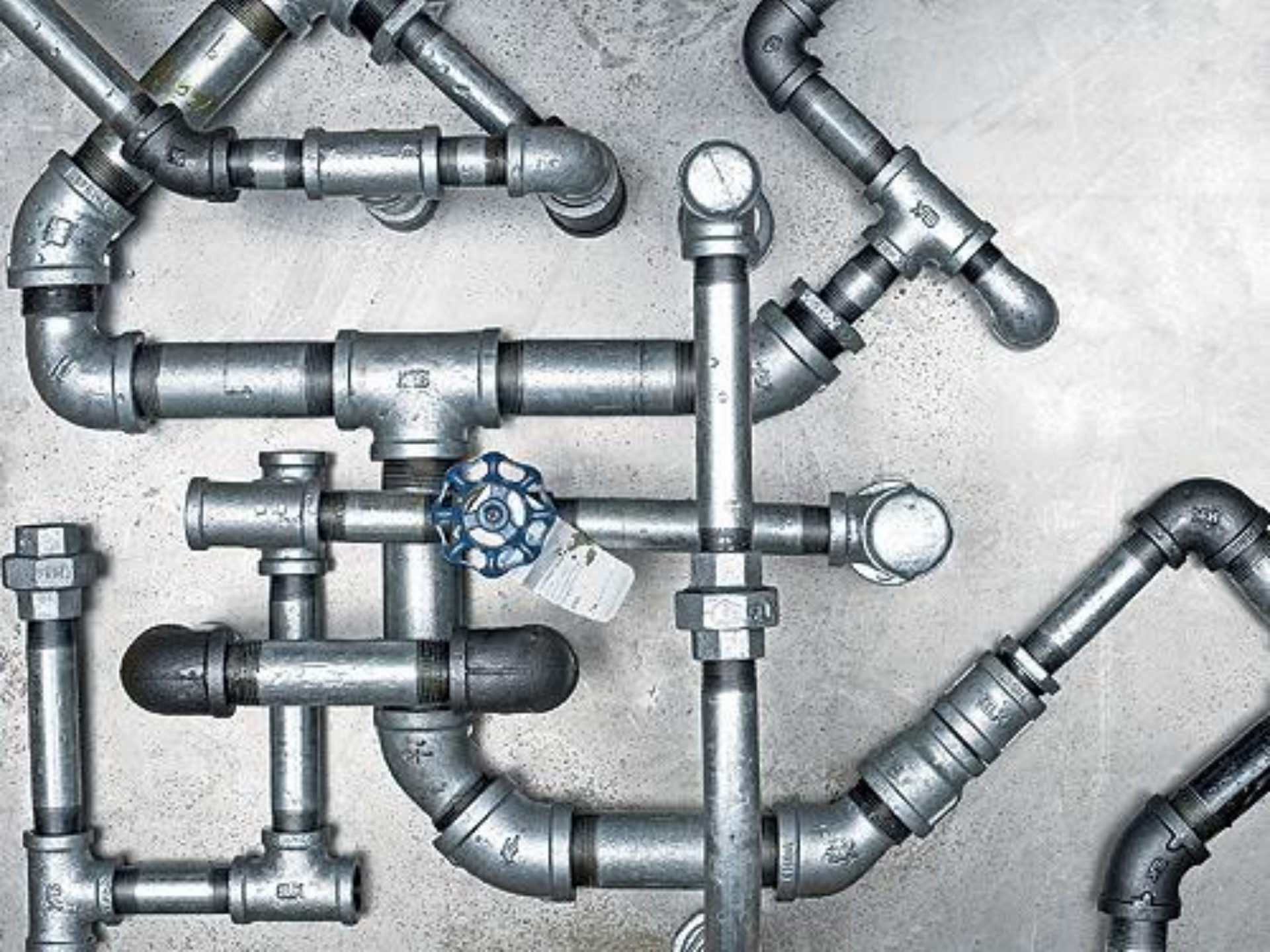


The Yale School  
of Medicine

**NORTHROP  
GRUMMAN**



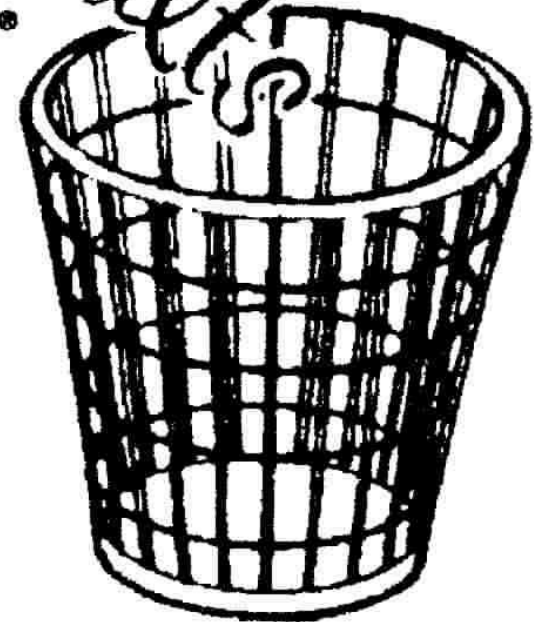
**SUL**



JOURNAL OF  
*Irreproducible Results*

*Official organ of the Society for Basic Irreproducible Research* (ISSN 0022-2038) ●

**Welcome to our 48th  
Year of Publication**



open  
data  
is about  
MORE  
THAN  
DISCLOSURE  
it must be  
Fair

- Findable
- Accessible
- Interoperable
- Reusable

# R Minimum Information About a Microarray Experiment

Abbreviation: MIAME

mibbi REPORTING GUIDELINE

## General Information

MIAME is intended to specify all the information necessary for an unambiguous interpretation of a microarray experiment, and potentially to reproduce it. MIAME defines the content but not the format for this information.

**Homepage** <http://www.fged.org/projects/miame/>

**Developed in** United Kingdom , France , Germany , Netherlands , Belgium , United States of America

**Created in** 1999

**Taxonomic range**

All

## Scope and data types

Microarray Data Genome DNA DNA Microarray Transcriptome RNA Nucleic Acid Hybridization

Record updated: March 11, 2016, 5:33 p.m. by [The BioSharing Team](#).

## Recommended by

EMBO Press

Scientific Data

## In Collections

National Child Development Study (UK)

DNA Microarray

## Related Standards

### Reporting Guidelines

Minimum Information about an ENVIRONMENTAL transcriptomic experiment  
Minimal Information about a high throughput SEQUENCING Experiment  
Minimum Information About a Microarray Experiment Involving Plants  
Minimum Information about a Nutrigenomic experiment  
Minimum Information about a array-based

## Implementing Databases (4)

### ArrayExpress

ArrayExpress is a database of functional genomics experiments that can be queried and the data downloaded. It includes gene expression data from microarray and high throughput sequencing studies. Data is collected to MIAME and MINSEQE standards. Experiments are submitted directly to ArrayExpress or are imported from the NCBI GEO database.

### Gene Expression Omnibus

The Gene Expression Omnibus (GEO) is a public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomic data submitted by the scientific community. In addition to data storage, a collection of web-based interfaces and applications are available to help users query and download the studies and gene expression patterns stored in GEO.

The Immunology Database and Analysis Portal

## Implementing Policies

Scientific Data's Recommended Data Repositories  
EMBO Press Recommended Databases and Data Standards

## Templates

LINCS CELL LINE

MORE TEMPLATES



## DISEASE, CENTER, PROVIDER



### DISEASE

CL\_Disease

CL\_Disease\_Detail

CL\_Disease\_Site\_Onset

CL\_Disease\_Age\_Onset



### CENTER

CL\_Center\_Name

CL\_Center\_Specific\_ID\*



### PROVIDER

CL\_Provider\_Name\*

CL\_Provider\_Catalog\_ID\*

## Templates

LINCS CELL LINE

MORE TEMPLATES



## DISEASE, CENTER, PROVIDER

### DISEASE

CL\_Disease

melanoma

Malignant **melanoma** of other specified sites of skin

Malignant **melanoma** of skin

Malignant **melanoma** of skin of ear and external auditory canal

Malignant **melanoma** of skin of eyelid, including canthus

Malignant **melanoma** of skin of lip

Malignant **melanoma** of skin of lower limb, including hip

Malignant **melanoma** of skin of other and unspecified parts of face

### CENTER

CL\_Center\_Name

CL\_Center\_Specific\_ID\*

### PROVIDER

CL\_Provider\_Name\*

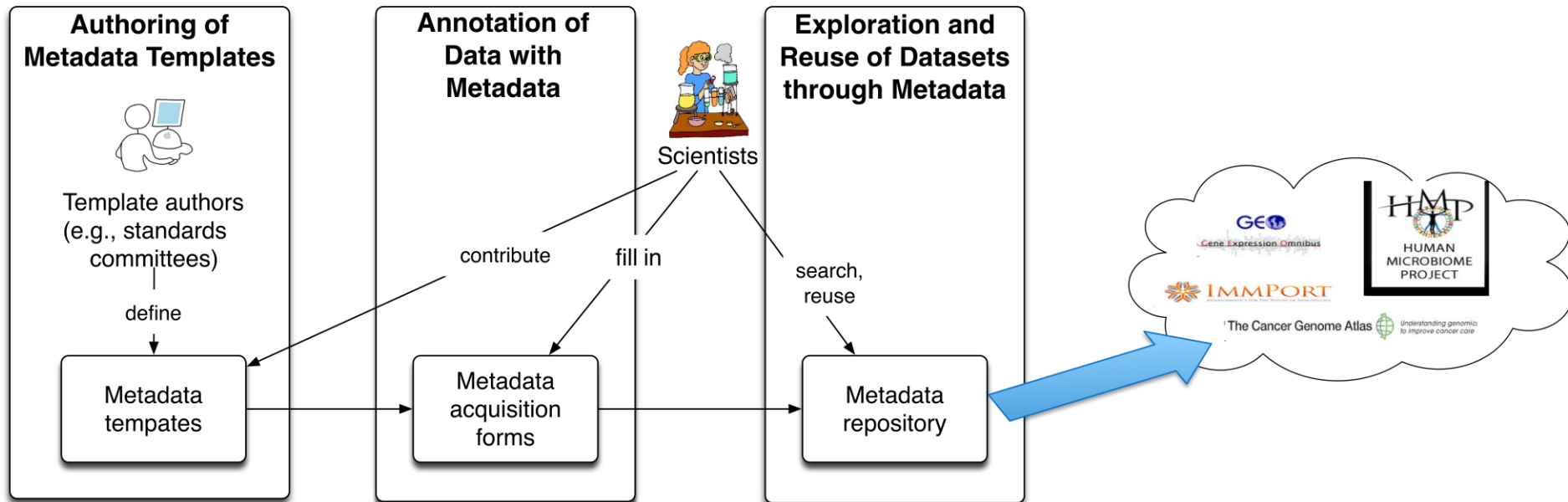
CL\_Provider\_Catalog\_ID\*



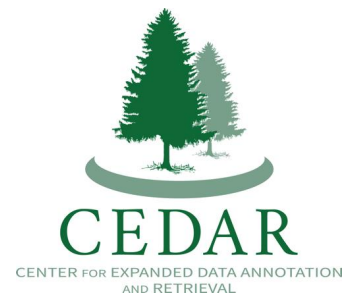
# Some key features of CEDAR

- 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

# The CEDAR Approach to Metadata

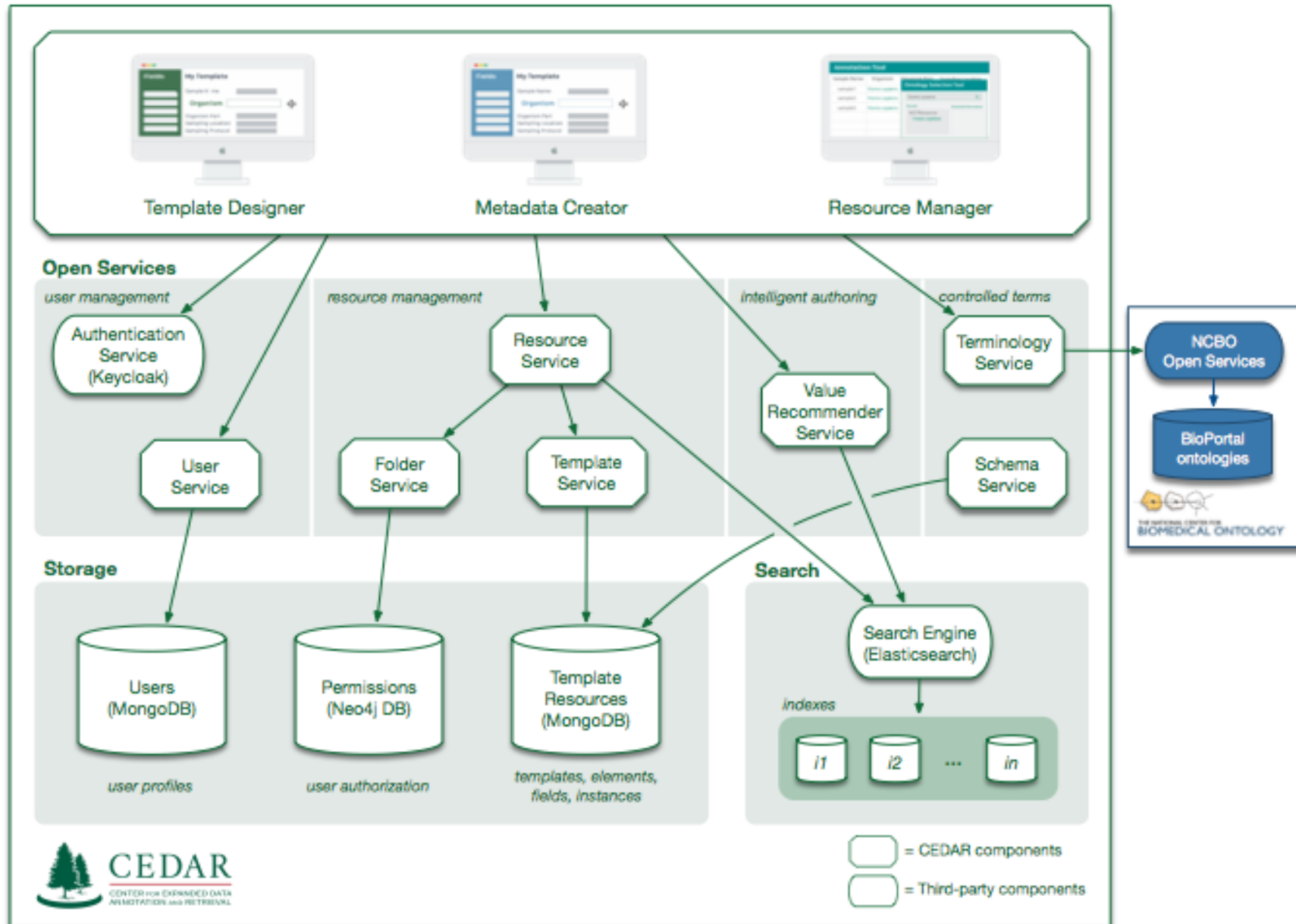


# A Metadata Ecosystem



- **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

# Architectural enhancements!



# A new CEDAR desktop!

The screenshot displays the CEDAR desktop interface. At the top left is the CEDAR logo. A search bar is located at the top right. Below the search bar is a dark navigation bar showing the user's path: "All / Users / John Graybeal". To the right of the navigation bar are icons for a menu, information, a double-headed arrow, and a user profile. On the left side, there is a "FILTER BY" sidebar with a "RESET" button. The sidebar includes sections for "TYPE" (with three circular icons), "AUTHOR", "STATUS", and "TERM", each with a dropdown arrow. The main area contains a "Test Metadata" folder icon and a grid of 12 data element cards. Each card features an icon (document or tag) and a title. An orange circular button with a white plus sign is located in the bottom right corner of the main area.

**CEDAR** Search

All / Users / John Graybeal

**FILTER BY** [RESET](#)

**TYPE**

**AUTHOR**

**STATUS**

**TERM**

Test Metadata

- Demonstration Element
- Enhanced Gene Expre...
- Copy of Gene Expressi...
- Demo for CBIIT metadata
- CDD Test metadata
- CDD Test
- Copy of Data Element
- Copy of Data Element ...
- Demo for CBIIT
- Copy of Gene Expressi...
- Super Enhanced Gene ...
- Publication

# Improved ontology search and term selection!

Injury Type ⚙️ 🔍

498 results for the query 'Injury Type'. Click on a term below to select it

TERM	DEFINITION	TYPE	SOURCE	ID
specified type of transport injury event				
Injury	Damage inflicted on the body as the direct or indirect result of an external force, with or without disruption of...	Class	NCIT	C3671 <span>👤 Hide Details</span>
Injury	-	Class	LOINC	LA17713-1
injury	Damage inflicted on the body as the direct or indirect result of an external force, with or without disruption of...	Class	HUPSON	SCAIVPH_00000269

**Ontology: NCIT**

**TERM DETAILS** | ONTOLOGY DETAILS

**Name** Injury

**Id** <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C3671>

**Definition** Damage inflicted on the body as the direct or indirect result of an external force, with or without disruption of structural continuity.

**Ontology Tree:**

- Adverse Event
  - Patient-Device Interaction
  - Chemical Imbalance
  - Stigmata Of Chronic Liver Disease
  - Injury**
  - Chemical Exposure
  - Ciliary Motility Defect
  - Fungal Colonization
  - Foreign Body Sensation
  - Amniotic Band
  - Obstruction
  - Failure To Thrive
  - Drug Exposure
  - Hypothermia

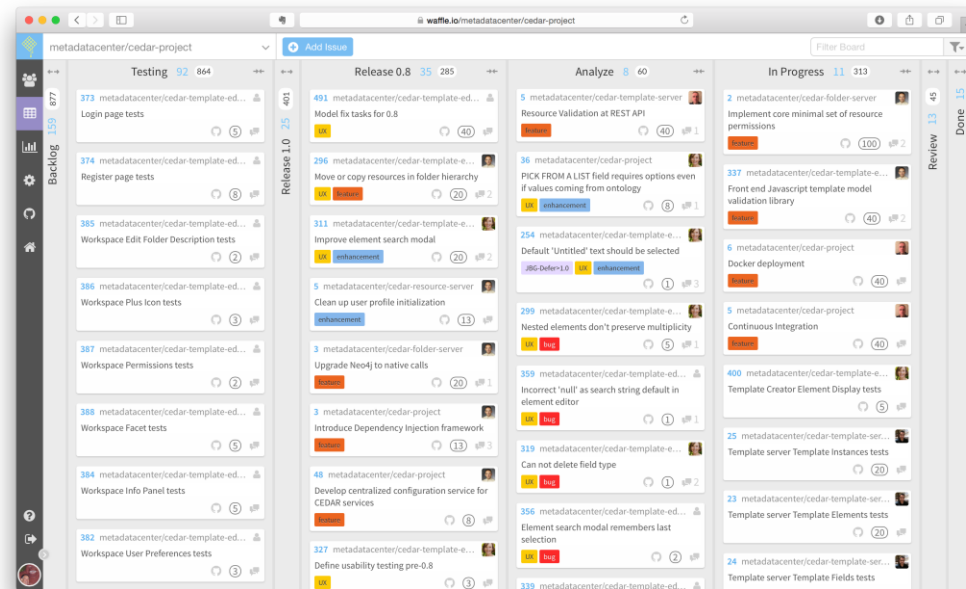
**TERM** | BRANCH | ONTOLOGY

**Term Id** <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C3671>

**Term Name** Injury

# Operations and Outreach!

- User support email lists
- User contact from Web site
- Open task tracking on GitHub at 'metadacenter'
- System monitoring
- Docker packaging



# Lots of kinds of metadata!

- Metadata describing **data sets** —



- Metadata describing **data types** and **value sets** —



- Metadata describing **data resources** —







## 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 Submission  
& Results](#)

1. Which Data Submission Templates do you need?  
Please contact us by email at [BISC\\_Helpdesk@niaid.nih.gov](mailto: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 Data Metadata Fields
Basic Study	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 using a single worksheet. A study design should be uploaded first.	<a href="#">basic_study_design.xls</a> <a href="#">basic_study_design.txt</a> 	<ul style="list-style-type: none"> <li>▶ Study</li> <li>▶ Title</li> <li>▶ Desc</li> <li>▶ End</li> </ul>



# NIH LINCS

## PROGRAM

LIBRARY

[HOME](#)

[CENTERS](#)

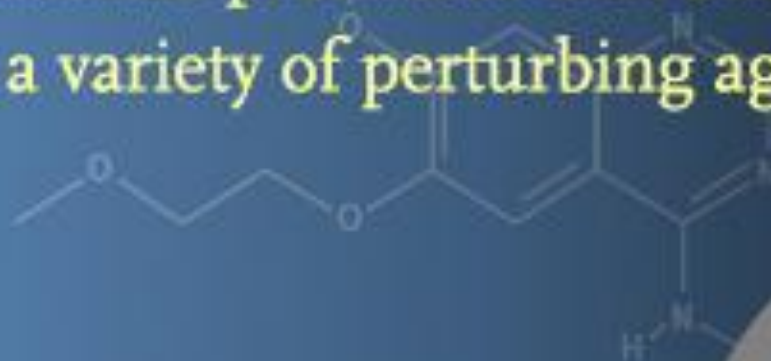
[DATA](#)

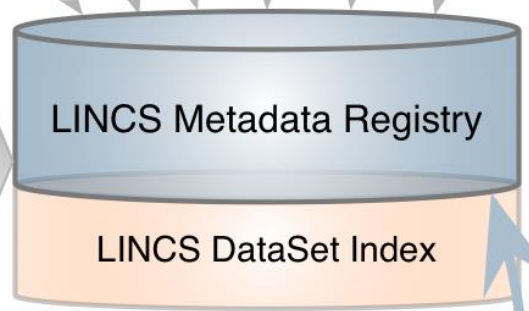
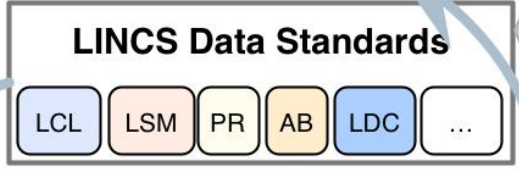
[COMMUNITY](#)

[PUBLICATIONS](#)

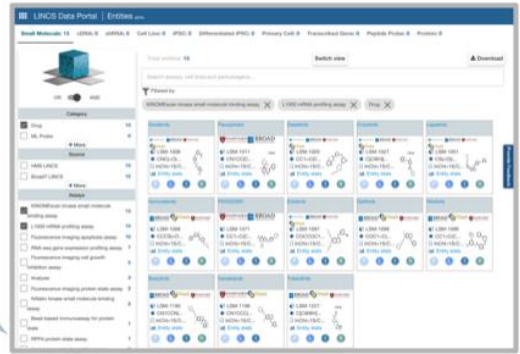
[NEWS](#)

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





**LINCS Data Portal**



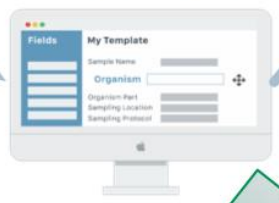
Create LINCS Standards Templates

Give Users Templates to Fill Out

Generate & Verify LINCS-importable Metadata

Explore Using LINCS & Other Templates

**Template Designer**



**Metadata Editor**

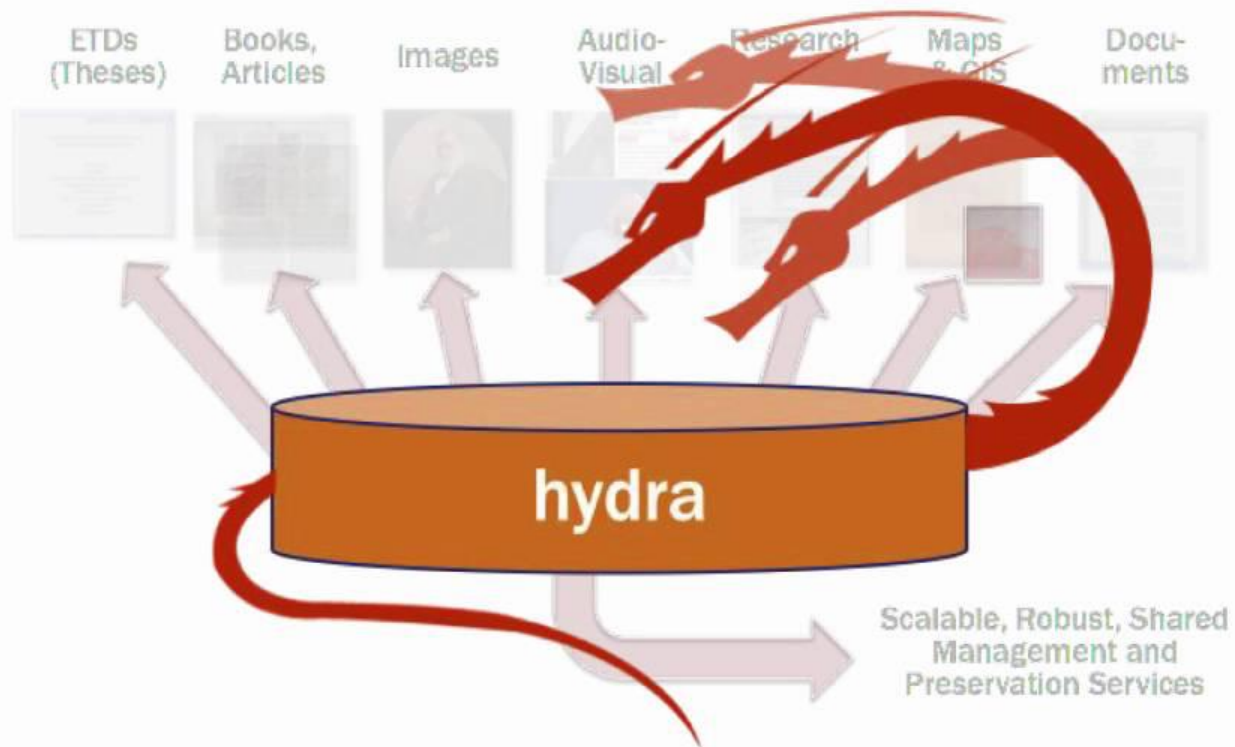


**Metadata Explorer**



# Hydra-in-a-Box

## One Body, Many Heads



# Lots of kinds of metadata!

- Metadata describing **data sets** —



- Metadata describing **data types**  
and **value sets** —



- Metadata describing **data resources** —



# NCI uses “common data elements” as metadata for fields in CRFs



## CDE Browser



[Admin Tool](#) [Curation Tool](#) [NCI Metathesaurus](#) [NCI Terminology Server](#) [Sentinel Tool](#) [UML Model Browser](#) [What's new](#) [Available Downloads](#) **New!**

Data Element Search

### Search for Data Elements

54040 Matches

[Search preferences](#)

[Advanced search](#)

#### caDSR Contexts

- Exact phrase
- All of the words
- At least one of the words

Name

Tip: This is an exact match search. To search for partial words or phrases use the \* as a wildcard.

Note: Default settings exclude Test and Training Context views from the tree and certain 'non-released' Workflow and Registration statuses. Click the 'Search Preferences' link above to view or change the exclusion criteria. Search Preferences will be reset to default settings when the 'New Search' button is clicked on the search results page or 'caDSR Context' in the Tree.

[Search](#) [Clear](#) [New Search](#)

### Search Results [Search within results](#)

Results fewer than expected? [Check Search Preferences](#)

[\[Download Data Elements to Prior Excel\]](#) [\[Download Data Elements to Excel\]](#) [\[Download Data Elements as XML\]](#)  
[\[Download CDE Browser DTDs\]](#)

Sort order : (Default) Registration Status>>Workflow Status>>Long Name [Ascending]

[Add to CDE Cart](#) [Add to CDE compare list](#) [Compare CDEs](#)

1 - 100 of 54040 [Next 100](#)

<input type="checkbox"/>	Long Name	Preferred Question Text	Owned By	Used By Context	Registration Status	Workflow Status	Public ID	Version
<input type="checkbox"/>	<a href="#">Access Route of Administration Text Code</a>	Route	CCR	AECC,BOLD,CITN,DCI,DCP,LCC,NCIP,NHC-NCI,NINDS,NRG,OHSU Knight,PBTC,SPOREs,USC/NCCC	Standard	RELEASED	2003586	6.0
<input type="checkbox"/>	<a href="#">Address Additional Identifier Text</a>	Additional Geographical Place, Name, Service	NCIP		Standard	RELEASED	2405182	1.0
<input type="checkbox"/>	<a href="#">Address Blocks Text Identifier</a>	Address Blocks	NCIP	CCR	Standard	RELEASED	2405421	1.0
<input type="checkbox"/>	<a href="#">Address Canton Identifier</a>	Address Canton	NCIP	CCR	Standard	RELEASED	2405438	1.0

Refresh tree

- caDSR Contexts
  - [ABTC \(Adult Brain Tumor Consortium\)](#)
  - [AECC \(Albert Einstein Cancer Center\)](#)
  - [Alliance \(Alliance\)](#)
  - [BBRB \(Biorepositories and Biospecimen Research Branch\)](#)
  - [BOLD \(Breast Oncology Local Disease\)](#)
  - [BRIDG \(BRIDG Collaboration\)](#)
  - [caCORE \(NCI Core Infrastructure\)](#)
  - [CCR \(NCI Center for Cancer Research\)](#)
  - [CDC/PHIN \(Centers for Disease Prevention and Control - P\)](#)
  - [CDISC \(Clinical Data Interchange Standards Consortium\)](#)
  - [CIP \(NCI Cancer Imaging Program\)](#)
  - [CITN \(Cancer Immunotherapy Trials Network\)](#)
  - [COG \(Children's Oncology Group\)](#)
  - [CTD-2 \(Cancer Target Discovery and Development\)](#)
  - [CTEP \(NCI Cancer Therapy Evaluation Program\)](#)
  - [DCI \(Duke Cancer Institute\)](#)
  - [DCP \(NCI Division of Cancer Prevention\)](#)
  - [ECOG-ACRIN \(ECOG-ACRIN\)](#)
  - [EDRN \(NCI Early Detection Research Program\)](#)
  - [GDC \(Genomic Data Commons\)](#)
  - [IcaRe2 \(Buffett Cancer Center\)](#)
  - [LCC \(Lombardi Cancer Center\)](#)
  - [MCL \(Molecular and Cellular Characterization of Screened\)](#)
  - [NCIP \(NCI cancer Biomedical Informatics Grid\)](#)
  - [NCIP CDE Data Standards \(Shortcut\)](#)
  - [NHC-NCI \(Norton Cancer Institute\)](#)
  - [NHLBI \(National Heart, Lung and Blood Institute\)](#)
  - [NICHD \(National Institute of Child Health and Development\)](#)
  - [NIDA \(National Institute on Drug Abuse\)](#)
  - [NIDCR \(National Institute of Dental and Craniofacial Resea\)](#)
  - [NINDS \(National Institute of Neurological Disorders and St\)](#)
  - [NRDS \(Network RAVE Data Standards\)](#)
  - [NRG \(NRG Oncology Group\)](#)

# CDEs are based on ISO/IEC 11179 model



## Data Element Details

<b>Public ID:</b>	4983471
<b>Version:</b>	1.0
<b>Long Name:</b>	Adrenal Cortical Cancer American Joint Committee on Cancer (AJCC) Edition 7 Clinical Tumor T Stage
<b>Short Name:</b>	4982812v1.0:4983378v1.0
<b>Preferred Question Text:</b>	Clinical T Stage
<b>Definition:</b>	Extent of the primary adrenal cortical cancer based on evidence obtained from clinical assessment parameters determined prior to treatment.
<b>Value Domain:</b>	Adrenal Cancer American Joint Committee on Cancer (AJCC) Edition 7 T Stage
<b>Data Element Concept:</b>	Adrenal Cortex Carcinoma Clinical T Stage
<b>Context:</b>	NCIP
<b>Workflow Status:</b>	RELEASED
<b>Origin:</b>	AJCC Based:Staging Criteria Based on American Joint Committee on Cancer System
<b>Registration Status:</b>	Standard
<b>Direct Link:</b>	<a href="https://cdebrowser.nci.nih.gov/CDEBrowser/search?elementDetails=9&amp;FirstTimer=0&amp;PageId=ElementDetailsGroup&amp;publicId=4983471&amp;version=1.0">https://cdebrowser.nci.nih.gov/CDEBrowser/search?elementDetails=9&amp;FirstTimer=0&amp;PageId=ElementDetailsGroup&amp;publicId=4983471&amp;version=1.0</a>

## Reference Documents

Document Name	Document Type	Document Text	Context	URL
Clinical T Stage	Preferred Question Text	Clinical T Stage	NCIP	

## Alternate Names and Definitions

## Other Versions

Version	Long Name	Workflow Status	Registration Status	Context
No other versions available				

RAI  
2.16.840.1.113883.3.26.2 ✓

Public ID  
4983471 ✓

Long Name  
Adrenal Cortical Cancer American Joint Committee on Cancer (AJCC) Edition 7 Clinical Tumor T Stage ✓

Preferred Name  
4982812v1.0:4983378v1.0 ✓

Preferred Definition  
Extent of the primary adrenal cortical cancer based on evidence obtained from clinical assessment parameters determin ✓

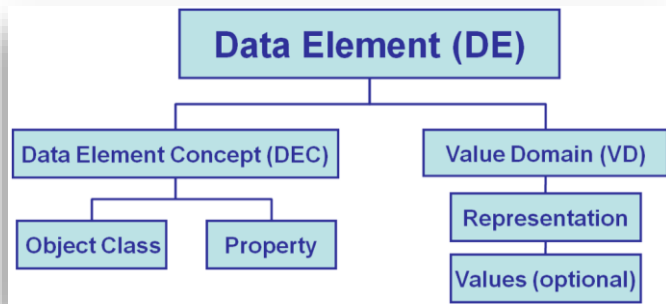
Version  
1 ✓

Workflow Status  
RELEASED ✓

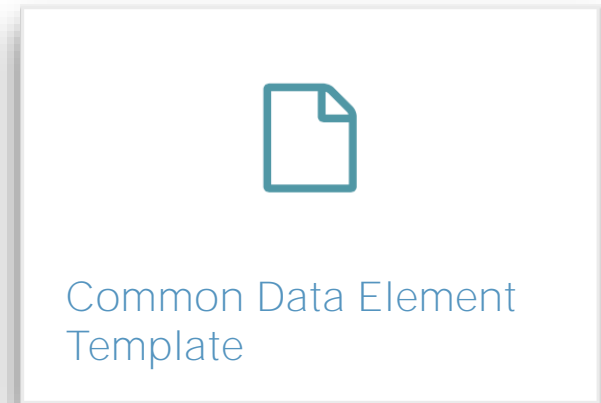
Context Name  
NCIP ✓



# Describing CDEs in CEDAR



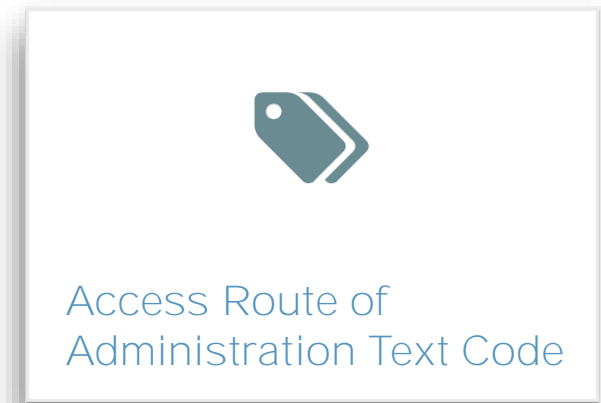
Data Element model



CEDAR Template

<input type="checkbox"/>	<b>Long Name</b>
<input type="checkbox"/>	<u>Access Route of Administration Text Code</u>


CDE Instance





CEDAR Metadata Instance











FILTER BY [RESET](#)

TYPE 

AUTHOR 

STATUS 

TERM 

	Title	Created	Modified	
	3925241 : Pleurodesis Performed Less Initial Pathologic Diagnosis Date Calculation Value	7/29/16 7:11 PM	7/29/16 7:11 PM	
	3479270 : Performance Status Initial Score Less Initial Pathologic Diagnosis Date Calculated Day Value	7/29/16 7:11 PM	7/29/16 7:11 PM	
	3302264 : Randomization To Event Calculation Day Month Duration Value	7/29/16 7:11 PM	7/29/16 7:11 PM	
	3302260 : Most Extensive Surgical Procedure To Event Calculation Day Month Duration Value	7/29/16 7:11 PM	7/29/16 7:11 PM	
	3302242 : Registration To Event Calculation Day Month Duration Value	7/29/16 7:11 PM	7/29/16 7:11 PM	
	2870015 : Document Identifier Identifier ISO21090.II.v1.0	7/29/16 7:11 PM	7/29/16 7:11 PM	

# Representing CDEs in CEDAR will allow authoring of CEDAR templates that can provide the basis for eCRFs



Case Report Form



Template

# Lots of kinds of metadata!

- Metadata describing **data sets** —



- Metadata describing **data types**  
and **value sets** —



- Metadata describing **data resources** —



## DATA REPOSITORY INFORMATION

A repository or catalog of datasets



Data Repository Information



Data Repository ID



Data Repository Name



Data Repository Description



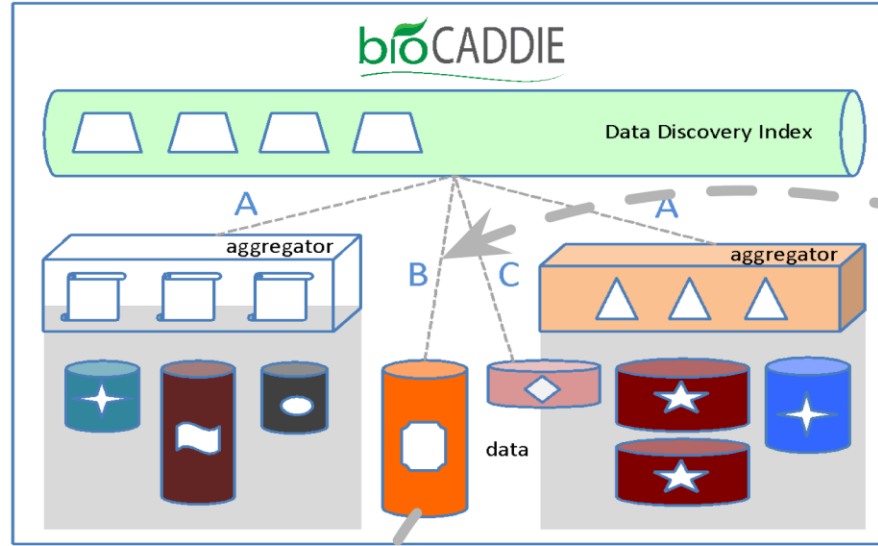
Data Repository Homepage



a



[A] Major aggregator services  
[B] Independent data sets



(4) Index data source(s)

Data source indexer

Data source description retriever

(3) Retrieve data source descriptions

bioCADDIE Authority

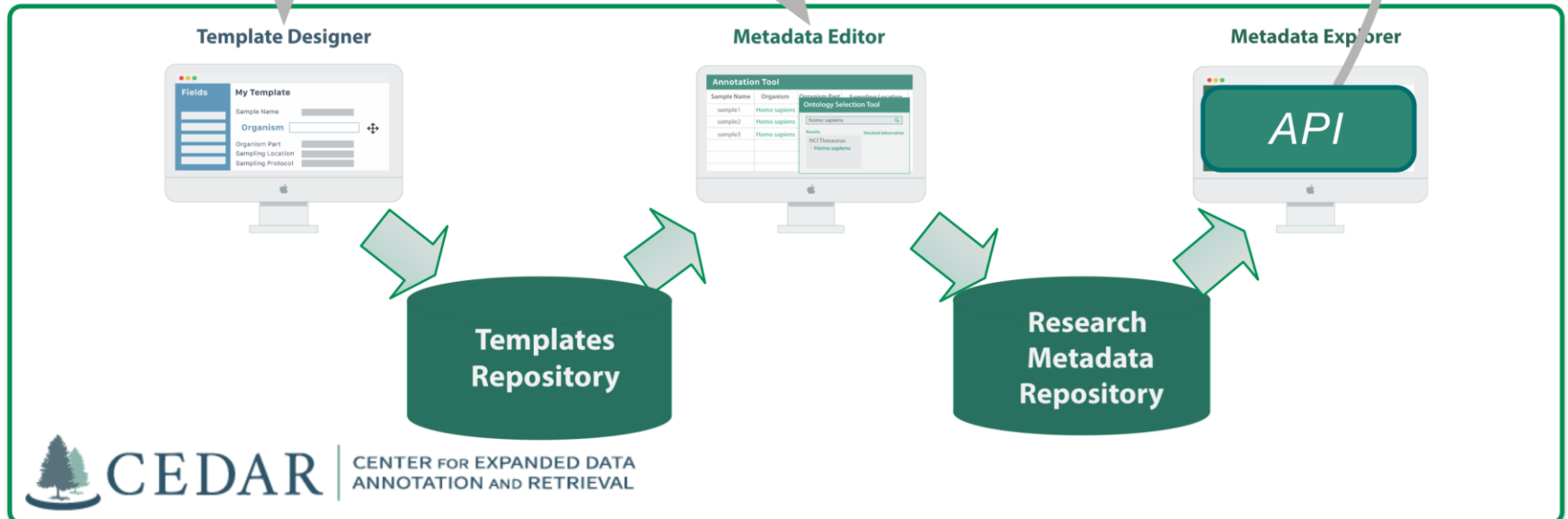


(1) Define metadata for data sources

Data source provider



(2) Describe independent data source



CEDAR

CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL

# Lots of kinds of metadata!

- Metadata describing **data sets** —



- Metadata describing **data types**  
and **value sets** —



- Metadata describing **data resources** —

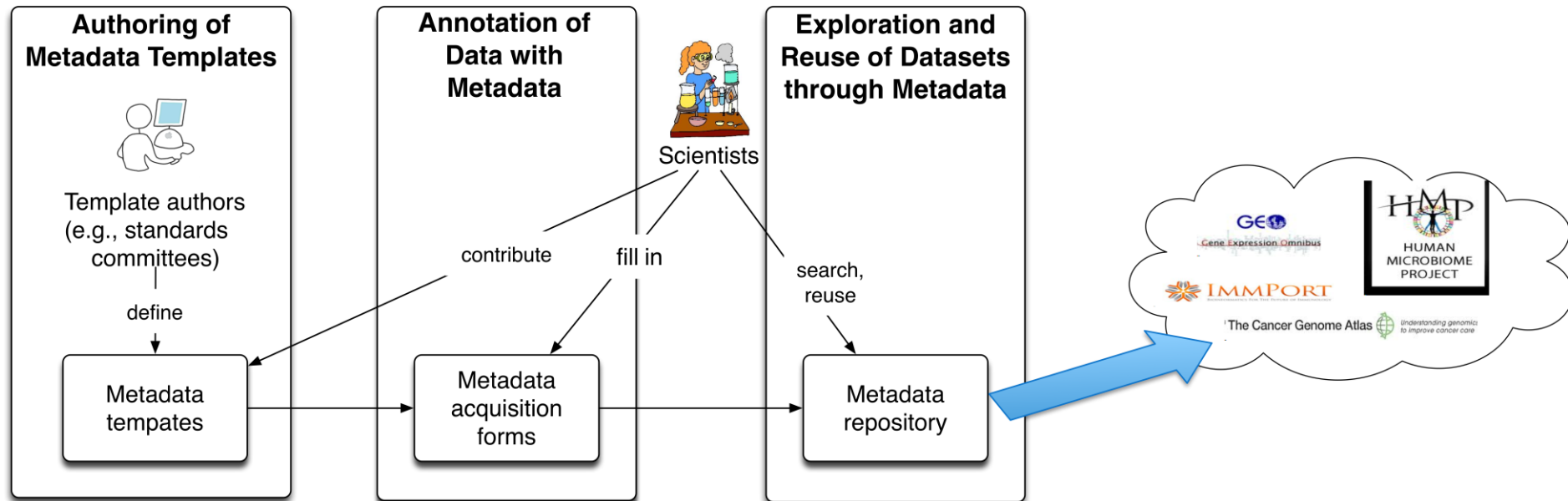


# CEDAR Status

- CEDAR Version 0.8 is now released; Version 1.0 is coming soon
- First “Bring your own data” event in the planning stage
- Ongoing collaboration with ImmPort
- Proof of concept with NCI CDEs
- New collaboration with LINCS
- Potential application with Hydra-in-a-Box
- Expanding opportunities with BioCADDIE



# The CEDAR Approach to Metadata



# Opportunities for Collaboration

- Everybody needs metadata authoring
- Everybody needs *smart* metadata authoring
- Everybody needs to use ontologies to clarify the semantics of data sets
- Everybody wants to mine, to explore, and to capitalize on patterns in (meta)data
- Everybody wants to publish scientific knowledge in machine-processable form





<http://metadatacenter.org>