CEDAR

Making Data FAIR Through Better Metadata

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

CEDAR
CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL

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

FAIR Data Principles

- Findable
- Accessible
- Interoperable
- Reusable

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

HOME SEARCH SITE MAP

NCBI > GEO > **Accession Display**

GEO help: Mouse over screen elements for information.

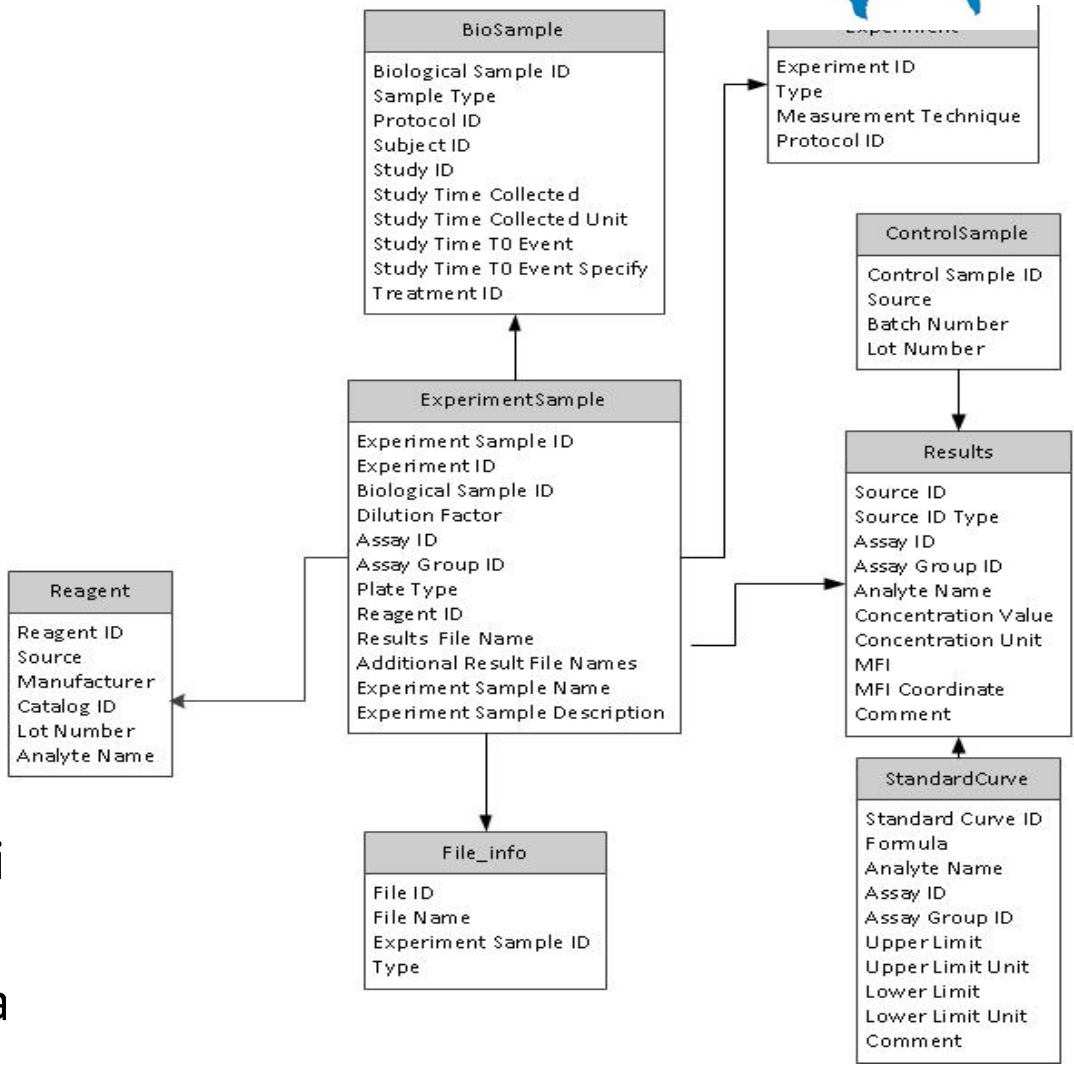
Scope: Self Format: HTML Amount: Quick

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 develop 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 b assess how centrosome loss or centrosome a 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 Dr biological sample, material dissected fror expression of the mutant strains was compar
Contributor(s)	Baumbach J, Levesque MP, Raff JW
Citation(s)	Baumbach J, Levesque MP, Raff JW. Centrosc dramatically perturb global gene expression i 15;1(10):983-93. PMID: 23213376

Human Immunology Project Consortium

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



Example entity–relationshi for describing metadata for multiplex bead array a (e.g., Luminex)

Data Submission / Resource / Data Submission Templates

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

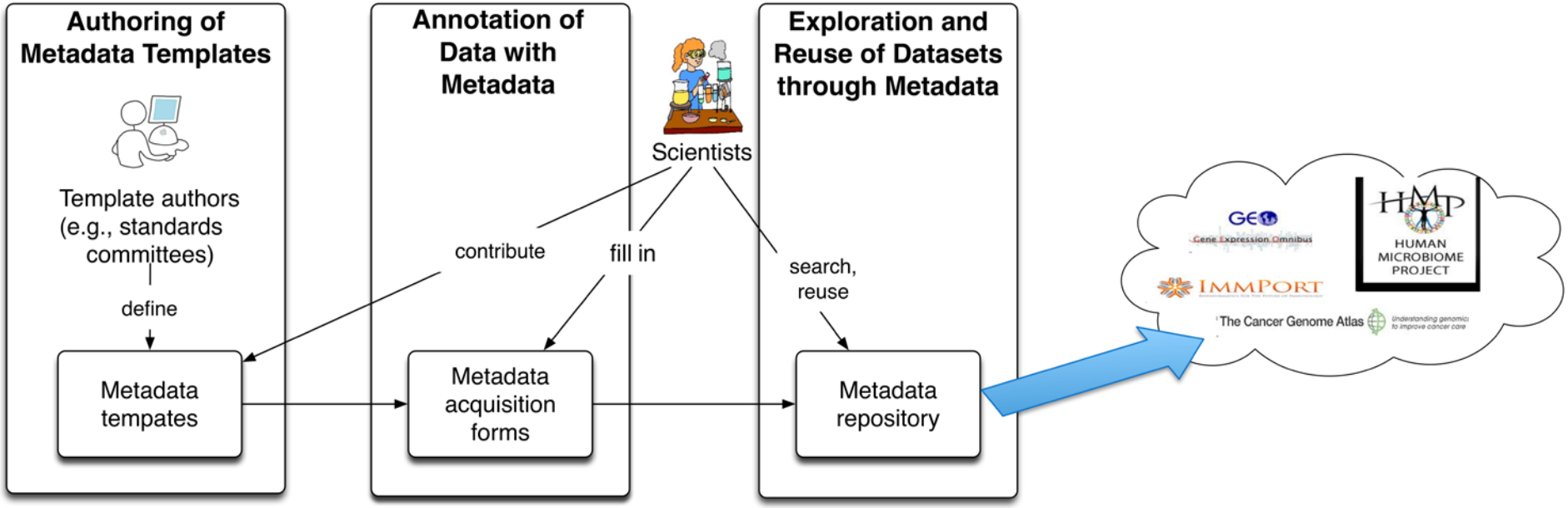
1. Which Data Submission Templates do you need?
   Please contact us by email at BISC_Helpdesk@niaid.nih.gov.
   The User Guide is a reference you can use to determine which templates need to be completed.
2. Complete the templates that are needed.
   Note: Please save spreadsheet .xls templates as tab 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

- **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

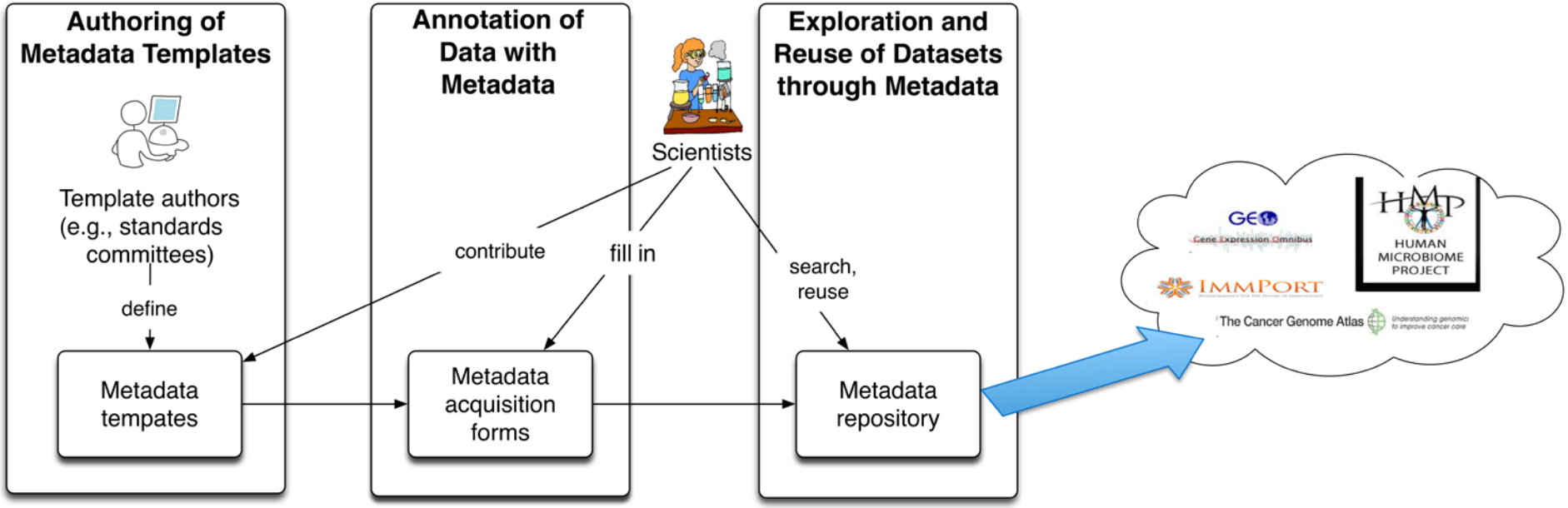
The CEDAR Approach to Metadata

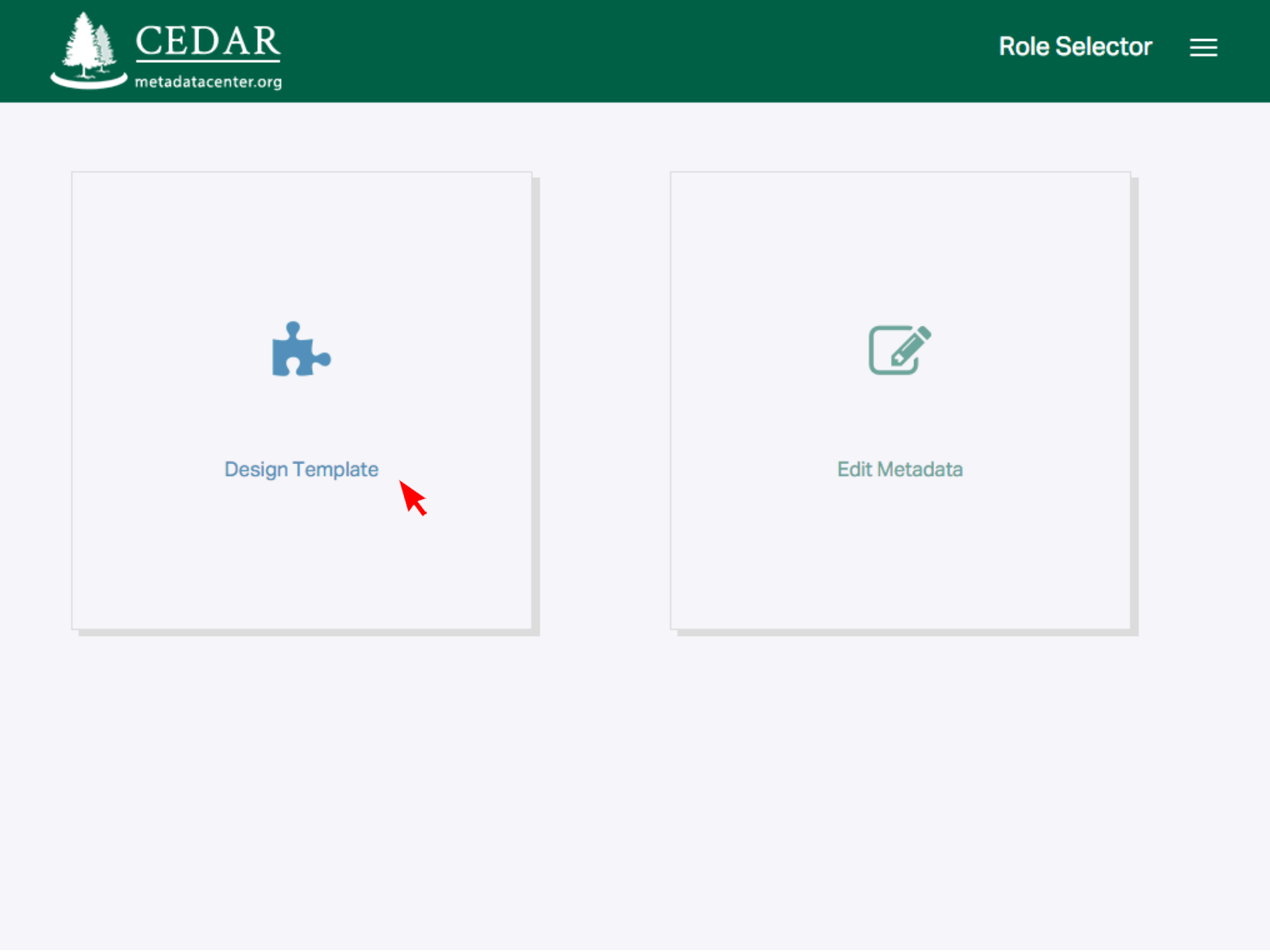


CEDAR technology will give us

- Mechanisms
  - To author metadata template elements
  - To assemble them into composite templates that reflect community standards
  - 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

The CEDAR Approach to Metadata




CEDAR
metadatacenter.org
Role Selector
Design Template
Edit Metadata

CEDAR
metadatacenter.org

Templates Listing

METADATA TEMPLATES

Search

BACK

CREATE NEW

LINCS Cell Line

CEDAR
metadatacenter.org

LINCS Cell Line

Template Designer

Add Element

- CELL LINE INFORMATION - SECONDARY
- CENTER
- CORE CELL LINE INFORMATION

MORE ELEMENTS

Add Field

ALL FIELDS

- CORE CELL LINE INFORMATION
- DISEASE
- CENTER
- PROVIDER
- CELL LINE INFORMATION- SECONDARY
- DONOR

CLEAR

SAVE TEMPLATE

CEDAR
metadatacenter.org

Disease

Element Designer

Add Field

- TEXT
- PARAGRAPH
- MULTIPLE CHOICE
- CHECKBOX
- DATE

ADD AN ELEMENT

Add an existing element or create a new item.

TEXT

Enter Field Title
CL_Disease

Enter Field Description (Help Text)
If the cell line came from a particular diseased tissue, the disease should be noted in terms of a con

Required Cardinality

CLEAR

SAVE ELEMENT

CEDAR
metadatacenter.org

Disease

Element Designer

Add Field

- TEXT
- PARAGRAPH
- MULTIPLE CHOICE
- CHECKBOX
- DATE

ADD AN ELEMENT

Required Cardinality

select value types

BioPortal Ontologies BioPortal Value Sets

Start Over

Search
diseases

Source Name	Type	Categories	Size
International Classification of Diseases, Version 9 - Clinical Modification	Ontology		22533
International Classification of Diseases, Version 10 - Clinical Modification	Ontology	Health	92168
International Classification of Diseases, Version 10 - Procedure	Ontology	Health	178319

CLEAR

SAVE ELEMENT

CEDAR
metadatacenter.org

Disease

Element Designer

Add Field

- TEXT
- PARAGRAPH
- MULTIPLE CHOICE
- CHECKBOX
- DATE

ADD AN ELEMENT

International Classification of Diseases, Version 10 - Clinical Modification

Class Details | **Ontology Details**

- Certain conditions originating in the
- Certain infectious and parasitic dise
- Congenital malformations, deforma
- Diseases of the blood and blood-fo
- Diseases of the circulatory system
- Diseases of the digestive system (K
- Diseases of the ear and mastoid pr
- Diseases of the eye and adnexa (H0
- Diseases of the genitourinary syste
- Diseases of the musculoskeletal sy
- Diseases of the nervous system (G
- Diseases of the respiratory system
- Diseases of the skin and subcutane
- Endocrine, nutritional and metaboli

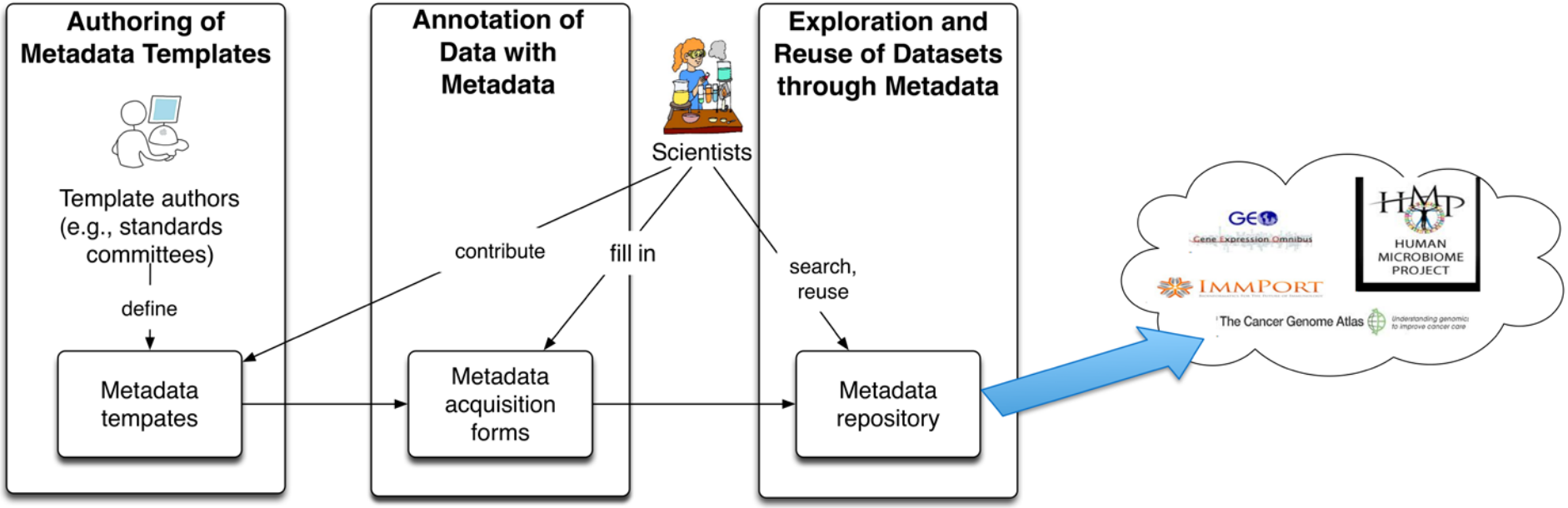
NAME	INTERNATIONAL CLASSIFICATION OF DISEASES, VERSION 10 - CLINICAL MODIFICATION
Acronym	ICD10CM
@id	http://data.bioontology.org/ontologies/ICD1
Description	
# of Classes	92168

ADD ONTOLOGY

CLEAR

SAVE ELEMENT

The CEDAR Approach to Metadata



CEDAR
metadatacenter.org

Role Selector

Design Template

Edit Metadata

CEDAR
metadatacenter.org

LINCS Cell Line

Metadata Editor

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*

CEDAR
metadatacenter.org

LINCS Cell Line

Metadata Editor

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

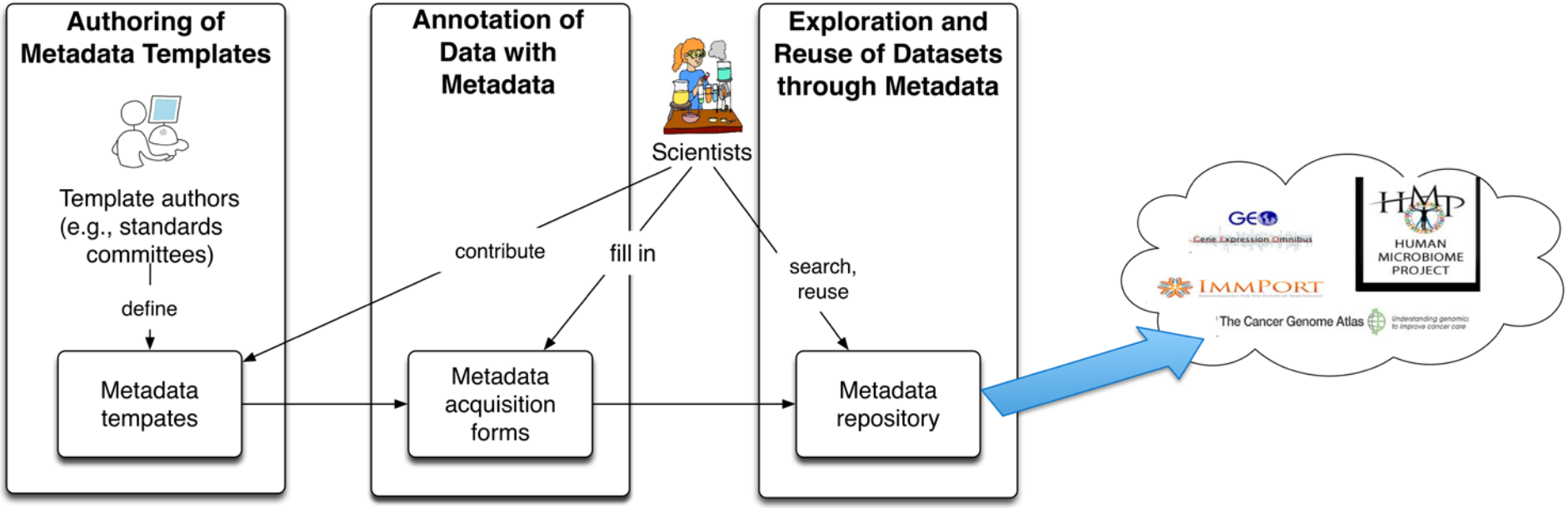
CL_Center_Specific_ID*

PROVIDER

CL_Provider_Name*

CL_Provider_Catalog_ID*

The CEDAR Approach to Metadata



FAIR Data Principles

- Findable
- Accessible
- Interoperable
- Reusable

How will CEDAR make data *FAIR*?

Data will be more <u>findable</u> because

- Datasets will be associated with rich metadata
  - Use of community-based templates will set expectations for the metadata elements that need to be specified
  - Use of predictive metadata entry will encourage the creation of more complete metadata specifications
- Metadata will use standard terminologies and value sets, that are themselves based on FAIR principles
- Metadata, and the data sets that they reference, will be given unique identifiers

How will CEDAR make data *FAIR*?

Data will be more <u>accessible</u> because

- All metadata and their associated data sets will have unique identifiers
- The CEDAR metadata archive will be persistent, even if the target data repositories are ever decommissioned

How will CEDAR make data *FAIR*?

Data will be more <u>interoperable</u> because

- CEDAR metadata are stored in a standard, widely used knowledge-representation language (JSON-LD)
- Use of community-based content standards (for templates, terminologies, and value sets) assures shared representation of metadata

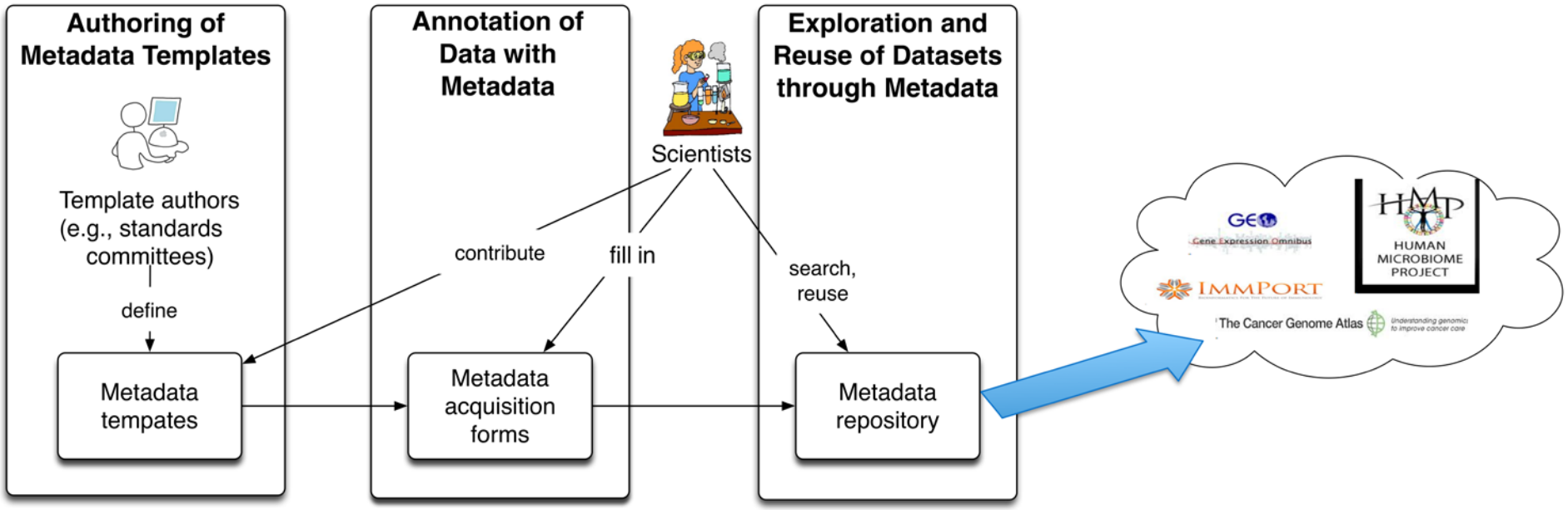
How will CEDAR make data *FAIR*?

Data will be more <u>reusable</u> because

- Metadata will include provenance information for the data sets that they reference and for the metadata themselves
- Metadata will be more comprehensive and more complete because
  - They will be easier to author
  - They will be constructed with community buy-in, due to the use of community-developed standards

biosharing

CEDAR: A Data Ecosystem Based on *FAIR*ness



http://metadatacenter.org

CEDAR

CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL