CEDAR

CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL

NIH Big Data to Knowledge (BD2K)

metadatacenter.org

(2014 -2018)

CEDAR
CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL

Steering Committee

Stanford MEDICINE

Yale University School of Medicine

UNIVERSITY OF OXFORD

NORTHROP GRUMMAN

Mark Musen,
Principal Investigator

Carol Bean,
Project Manager

Michel Dumontier

Olivier Gevaert

Purvesh Khatri

Steven Kleinstein

Kei-Hoi Cheung

Susanna-A Sansone

Jeffrey Wiser

Community-developed *content standards*

- To ***structure***, ***enrich*** and ***report*** the description of the datasets and the experimental context under which they were produced
- To facilitate **discovery**, **sharing**, **understanding** and **reuse** of datasets

formats

Including ***conceptual model***, ***conceptual schema*** from which an exchange format is derived to allow data to flow from one system to another

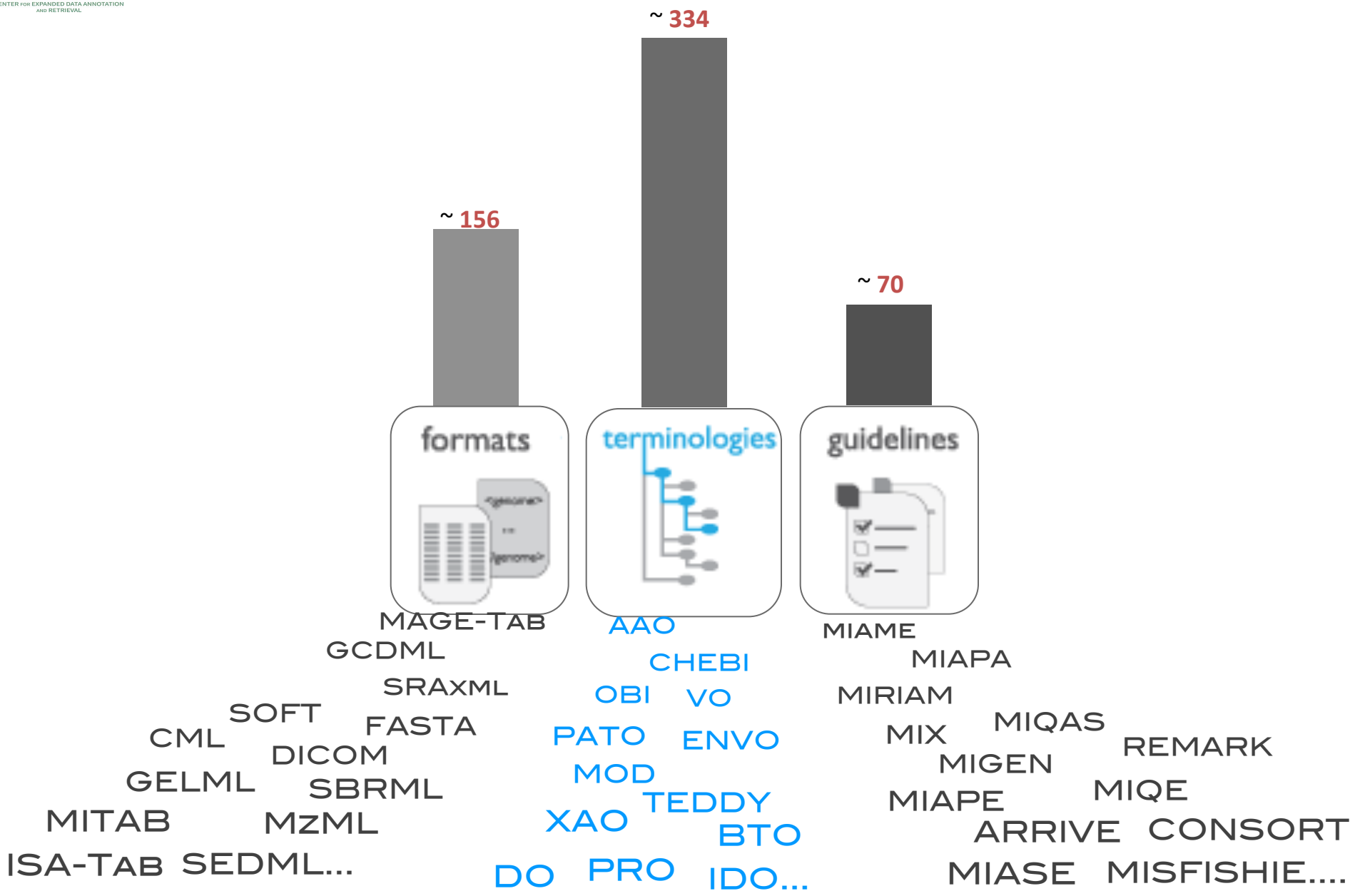
terminologies

Including ***controlled vocabularies***, ***taxonomies***, ***thesauri***, ***ontologies*** etc. to use the same word and refer to the same 'thing'

guidelines

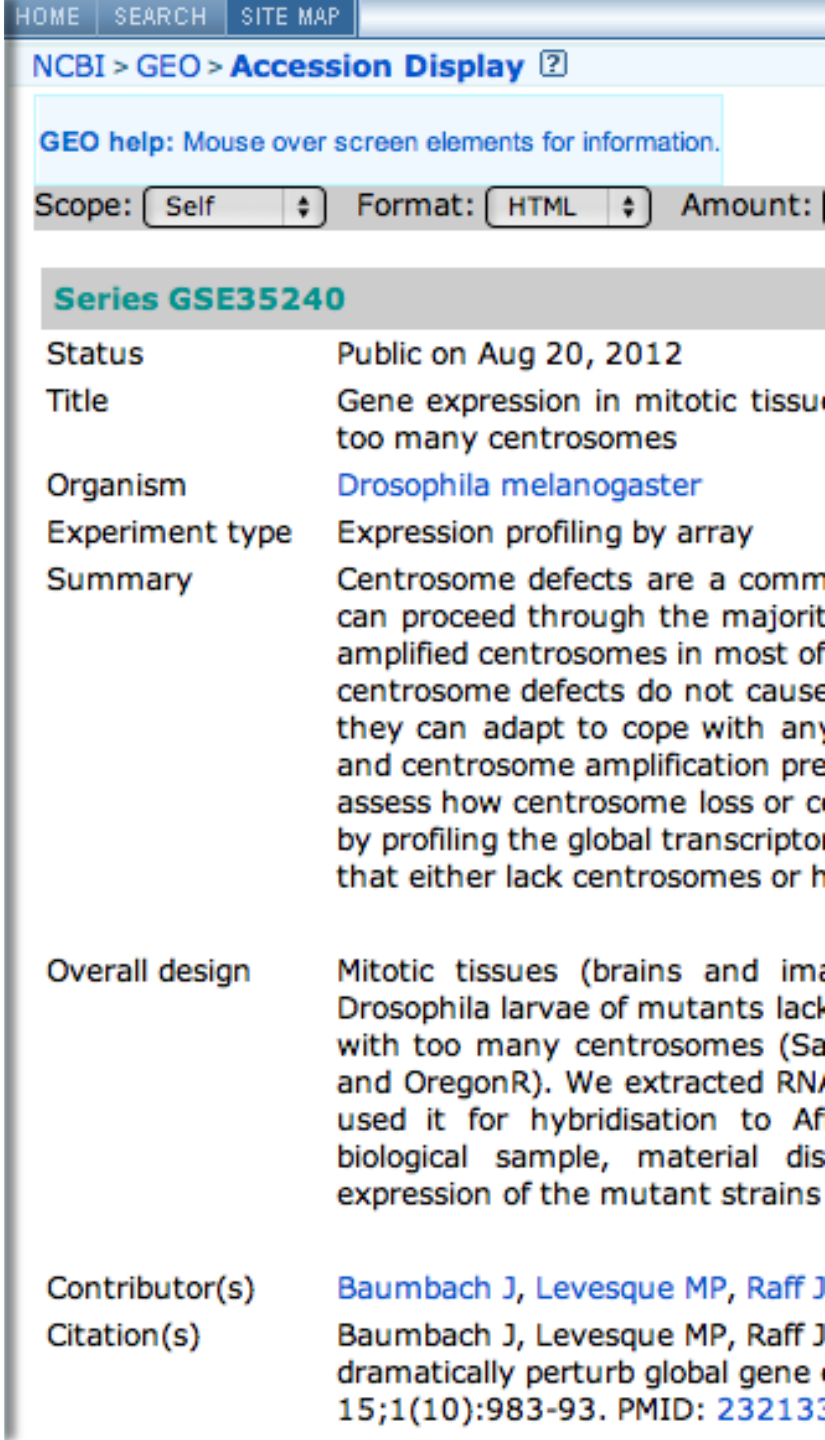
Including ***minimum information reporting requirements***, or ***checklists*** to report the same core, essential information

In the life sciences alone…..almost 600!



Researchers *hate* standards!

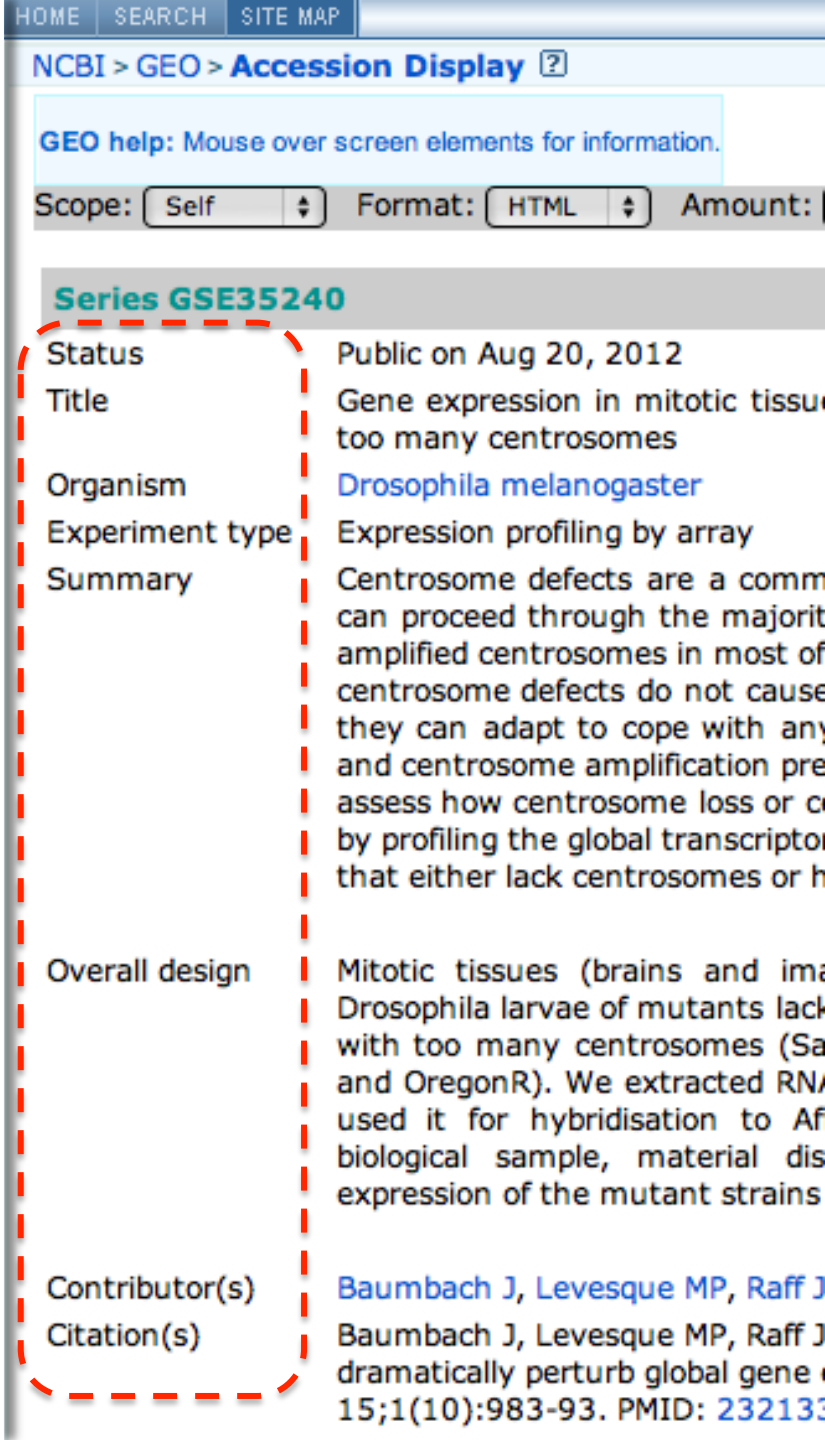
- Most researchers understand the value of standardized descriptions, when using third-party datasets
- But when asked to structure their datasets, they view requests for even “minimal” information as burdensome



Researchers *hate* standards!

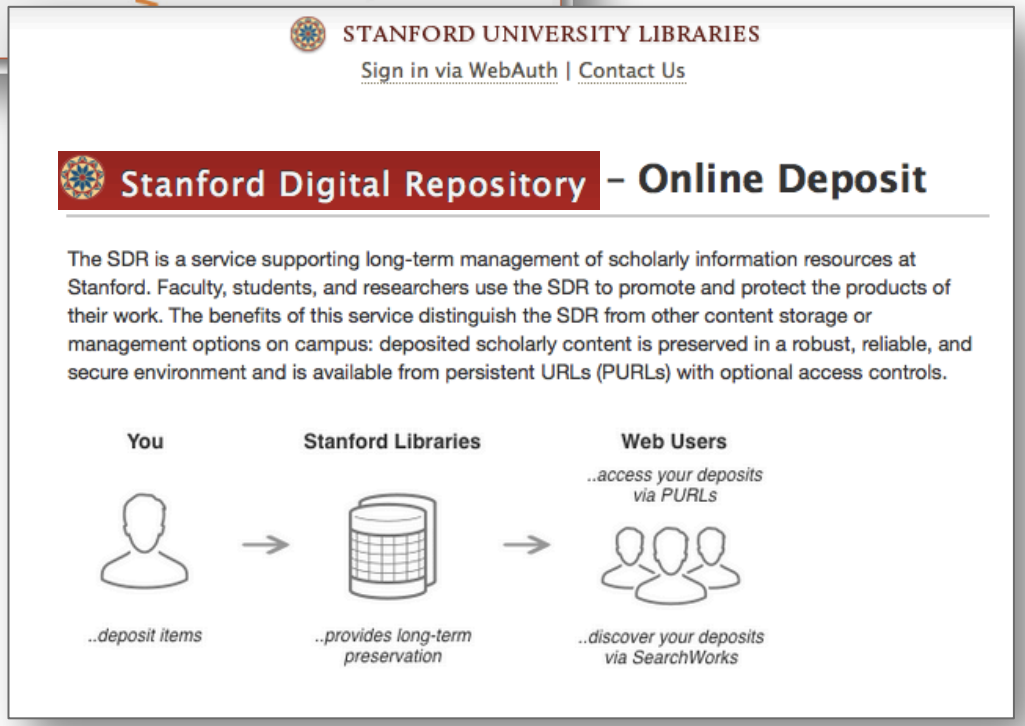
- Most researchers understand the value of standardized descriptions, when using third-party datasets
- But when asked to structure their datasets, they view requests for even "minimal" information as burdensome

➢ There is an urgent need to lower the bar for authoring good metadata



CEDAR
CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL

Our two initial uses cases



1. Map the landscape of content standards

biosharing.org

POLICIES REGISTRY

A catalogue of data preservation, management and sharing policies from international funding agencies and regulators.

STANDARDS REGISTRY

FORMATS TERMINOLOGIES REPORTING GUIDELINES

A catalogue of reporting standards and organizations that develop these.

Partly compiled by linking to

BioPortal mibbi equator network

DATABASES REGISTRY

BioDBcore

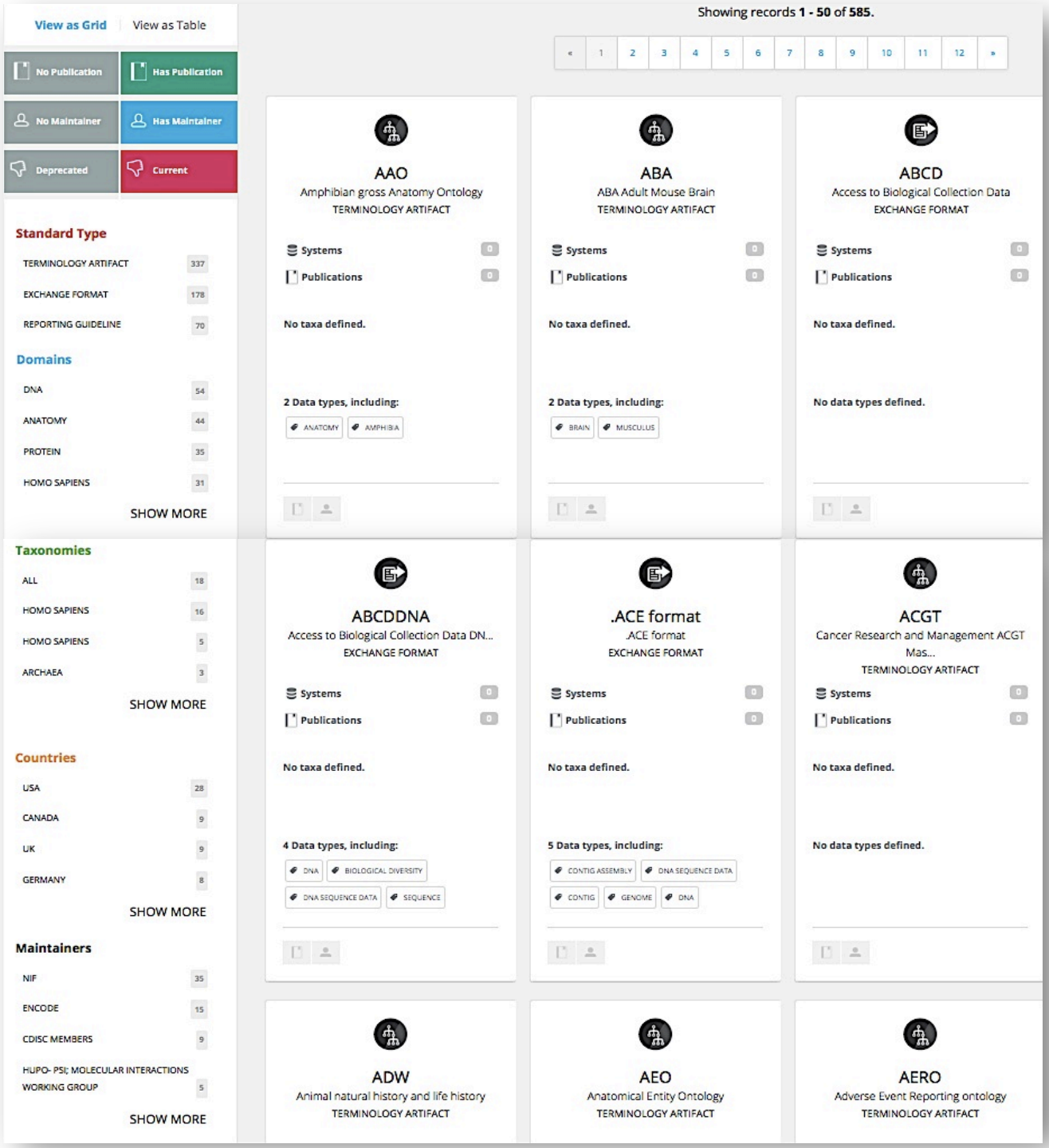
A catalogue of databases, described according to the BioDBcore guidelines, along with the standards used within them.

Partly compiled with the support of

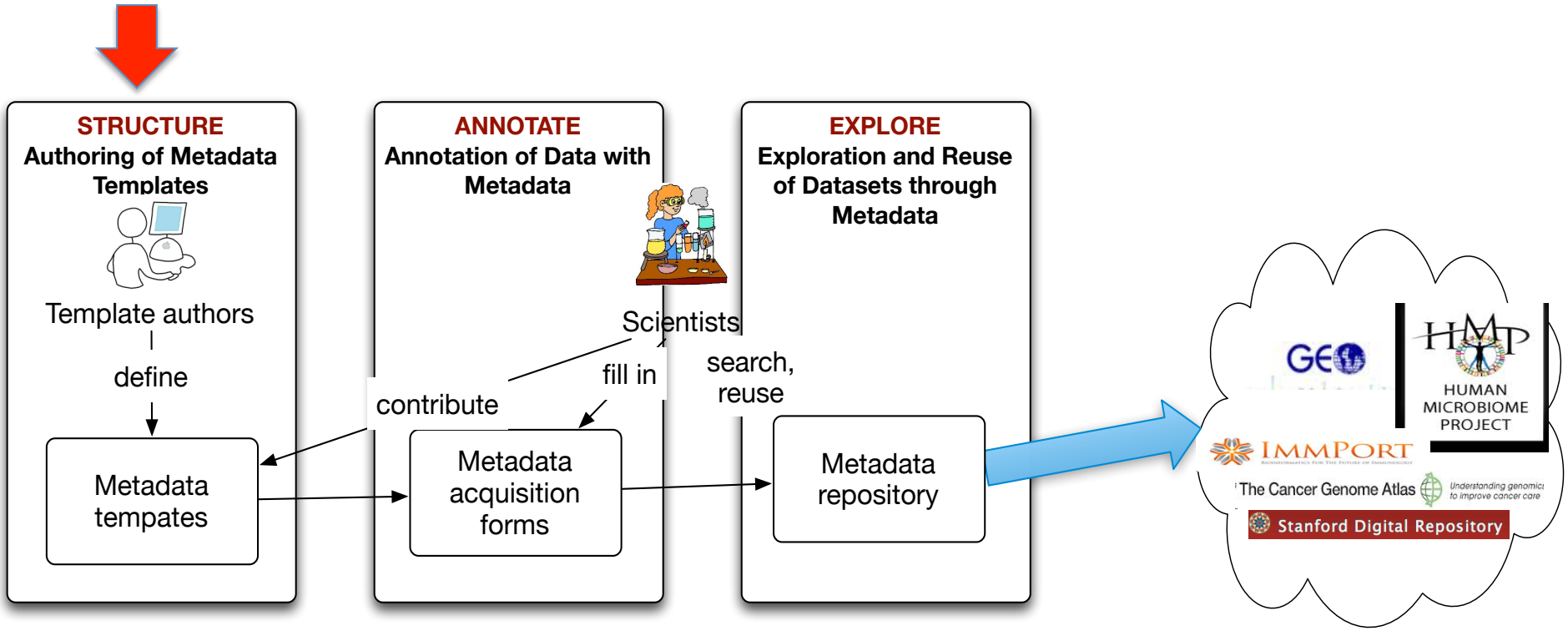
OXFORD UNIVERSITY PRESS re3data.org

biosharing.org

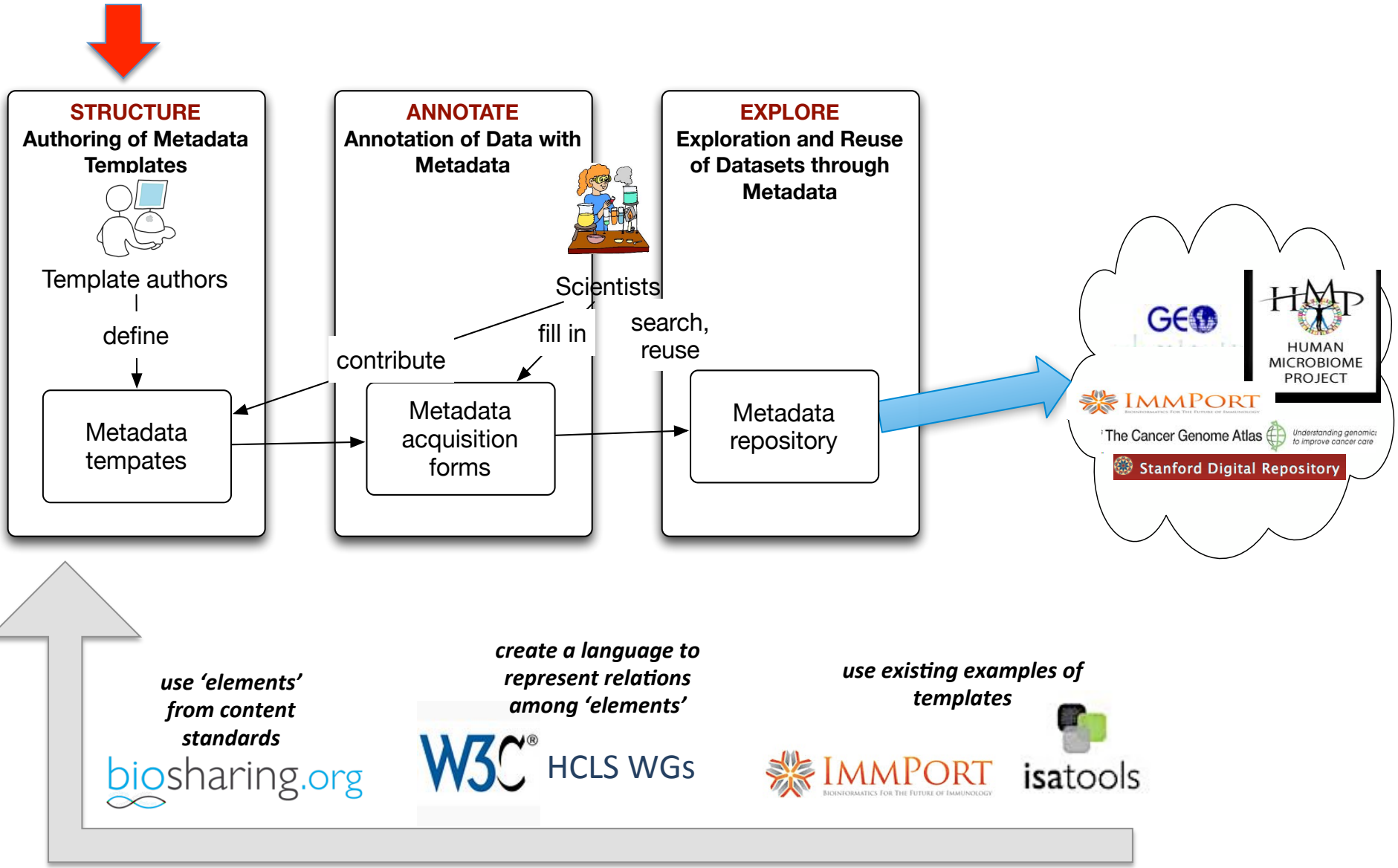
Almost 600!



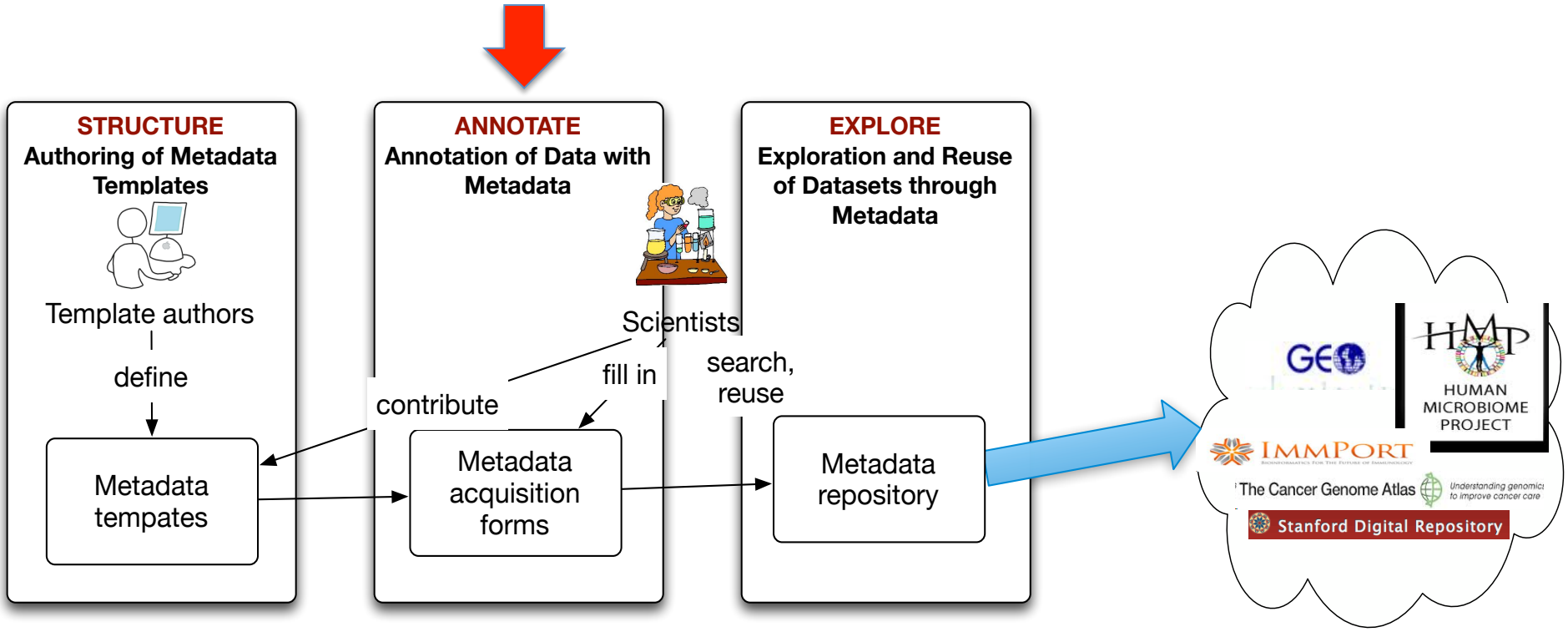
2. Develop methods for creating templates



2. Develop methods for creating templates

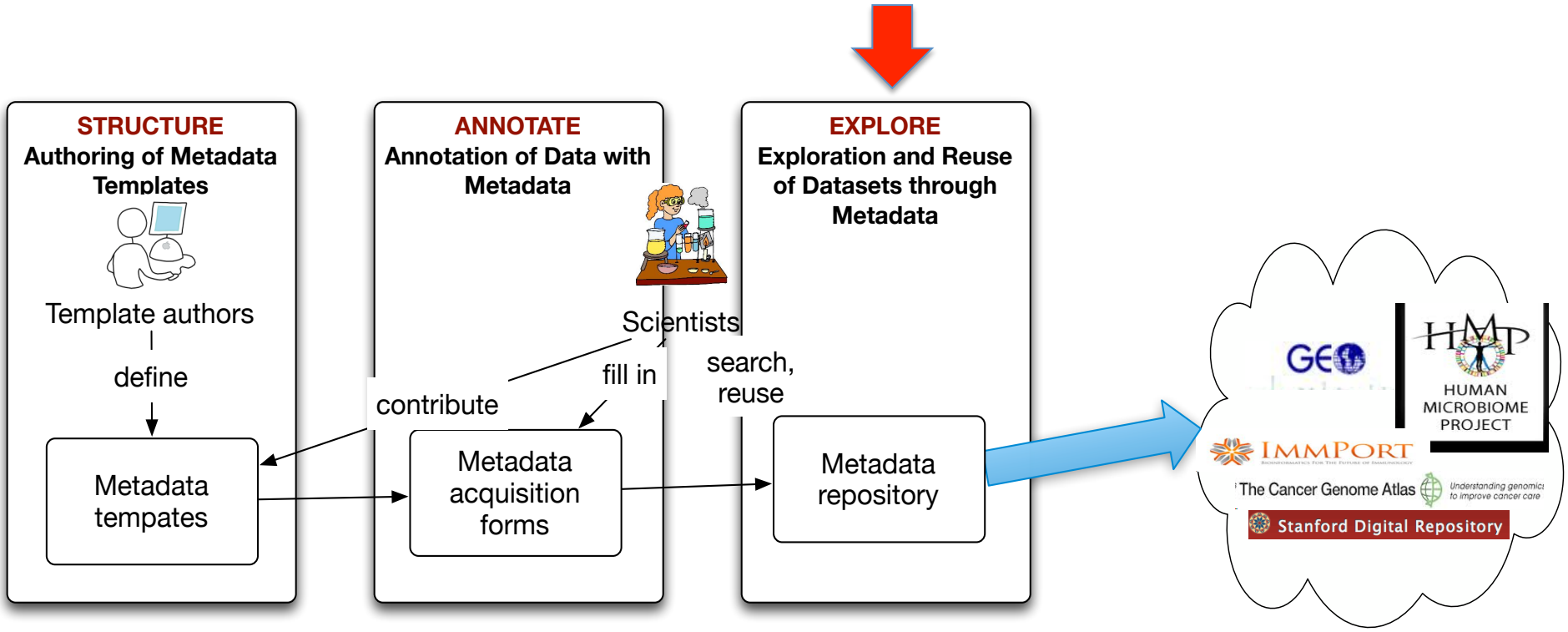


3. Develop methods to ease use of templates



- Enable researchers to help us creating templates appropriate to their needs
- Help researchers to ***find*** and ***use*** these templates to describe their experiments, and ***populate*** them with appropriate values (e.g. terms from ontologies)

4. Create a repository of populated templates



CEDAR repository will:

- ***store*** the experimental descriptions
- ***facilitate*** submission of datasets to our two case study repositories and progressively to other recognized online repositories

CEDAR
CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL

5. Exploring ways to enhance metadata

- Analyze the CEDAR repository to reveal ***patterns*** in the metadata that will enable the metadata tools to use ***predictive data entry*** to ease the task of filling out the templates
- 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 how to ease the authoring of metadata, using community standards, to enhance the richness of the experimental descriptions

CEDAR

CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL

NIH Big Data to Knowledge (BD2K)

metadatacenter.org

(2014 -2018)