



biomedical and healthCare
Data Discovery Index Ecosystem

Enabling the Big Data Commons
through indexing of data and
their interactions

2nd BD2K all-hands meeting
Bethesda
11/12/15

Aims

1. Help users find accessible data
2. Assist data producers on how to publish data for maximal discoverability
3. Build a prototype/platform to dock related products

PubMed of Data = *DataMed*

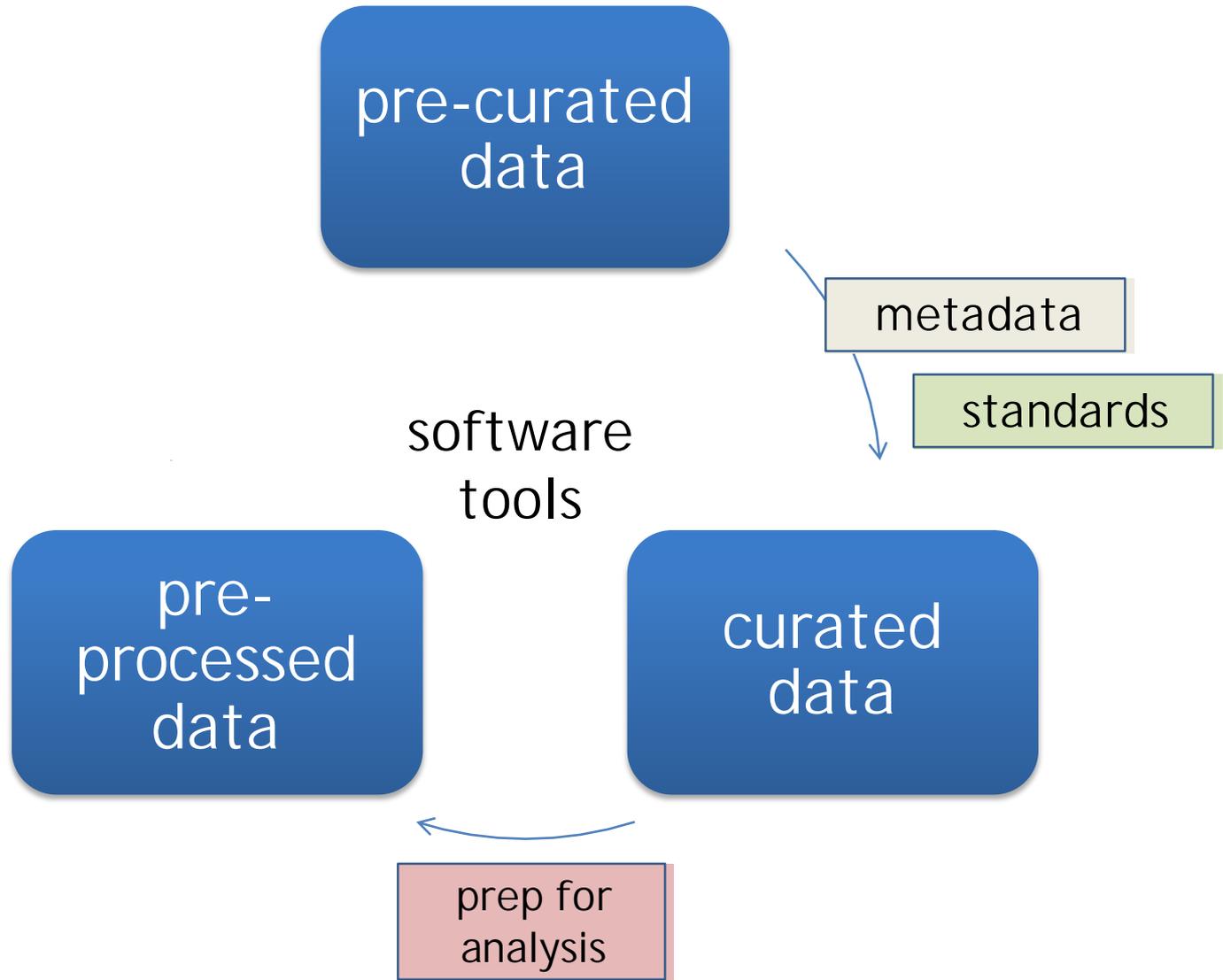
- ❖ The data ecosystem
 - ◆ Data, derived data, metadata
 - ◆ Stakeholders
 - ❖ Nuts and Bolts
 - ◆ Components: metadata, search tool
 - ◆ Plan and timelines
 - ❖ How to participate
 - ◆ Working groups
 - ◆ Pilots
 - ◆ Collaborations
-

What does it take to use big data?

- ❖ Find the data (across various resources)
- ❖ Find the tools that operate on the data
- ❖ Find the appropriate computational environment

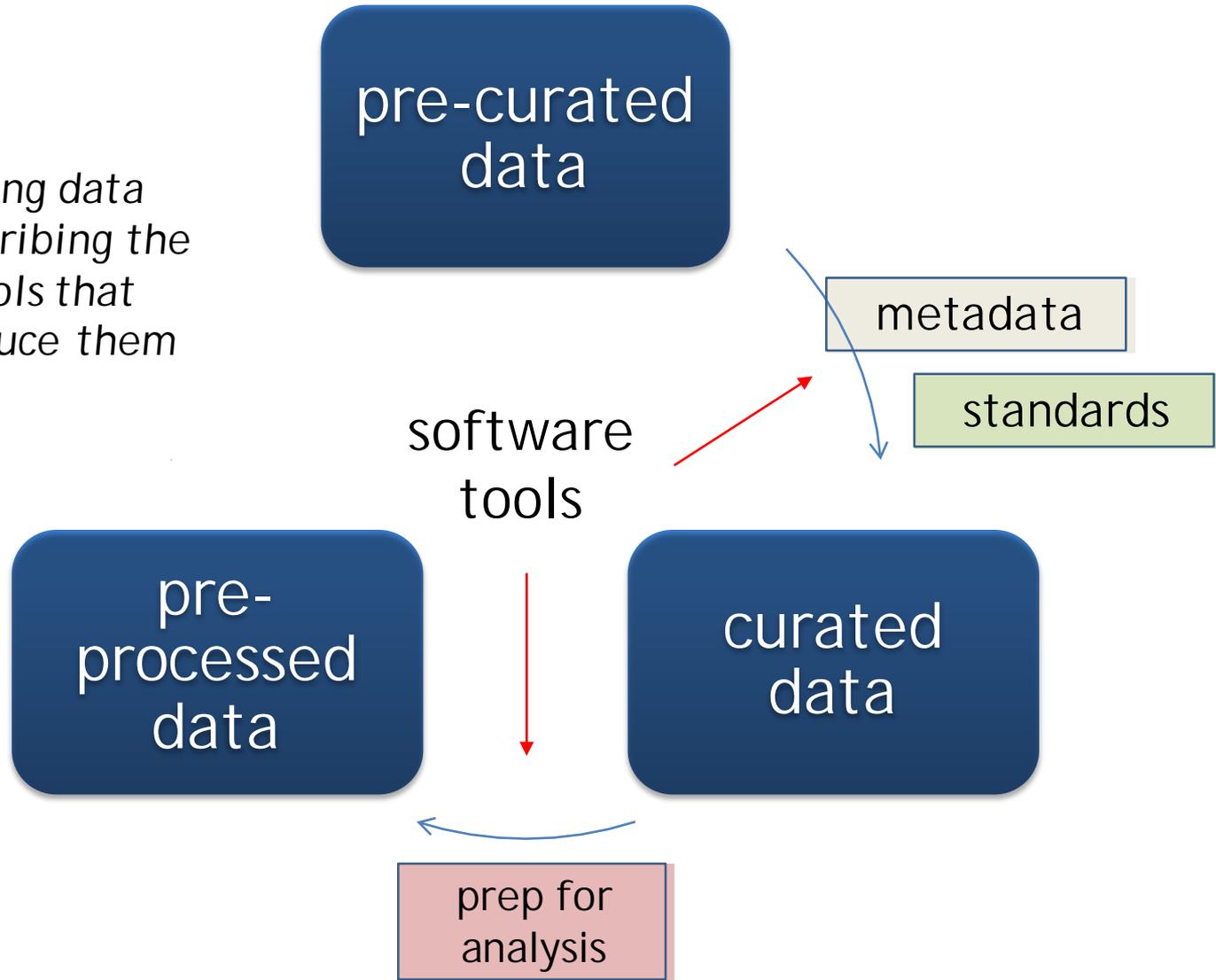


- ❖ Access the data
 - ❖ Access the tools (software/systems)
 - ❖ Access the computational environment
-



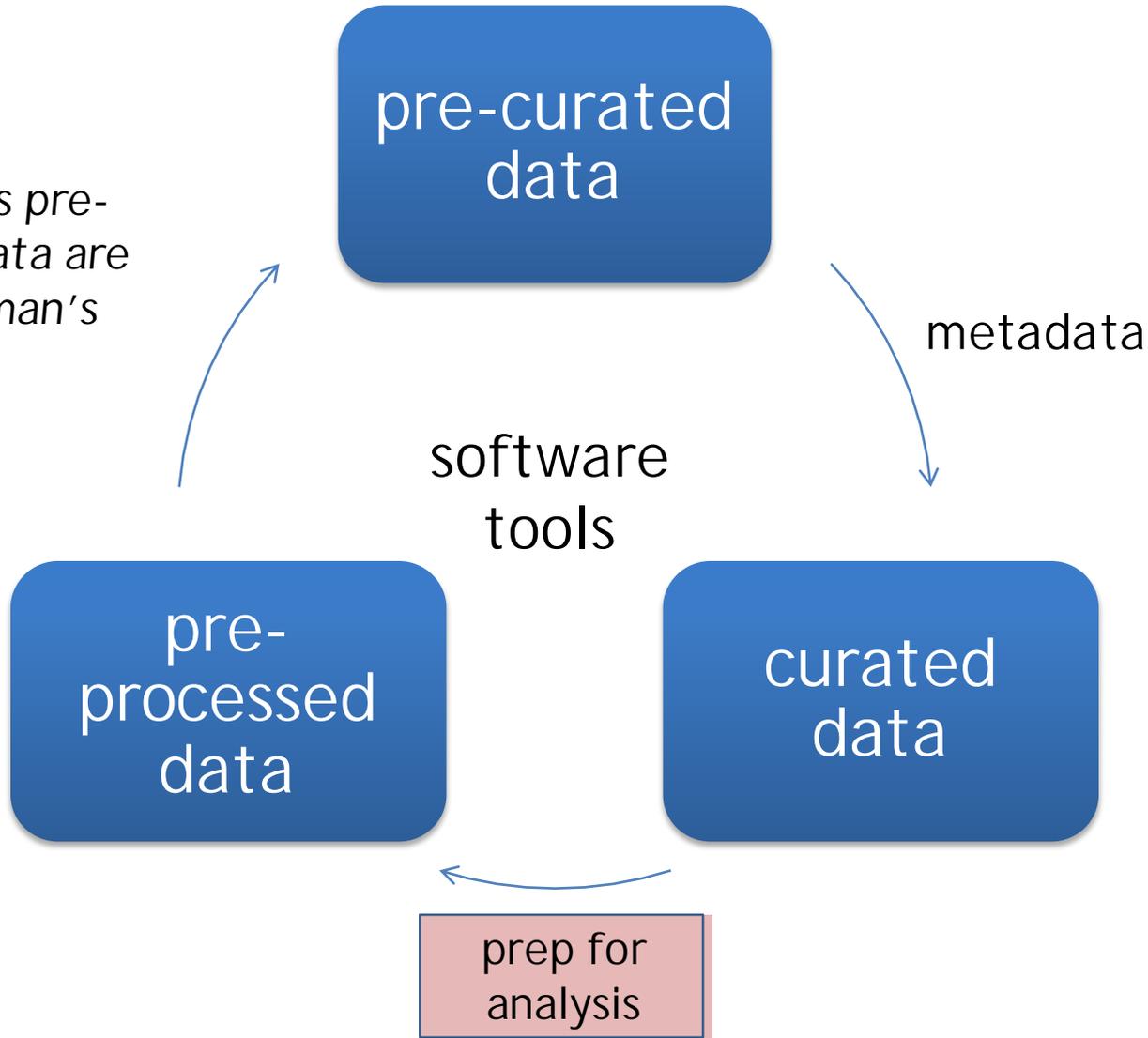
data ecosystem

Characterizing data implies describing the software tools that helped produce them



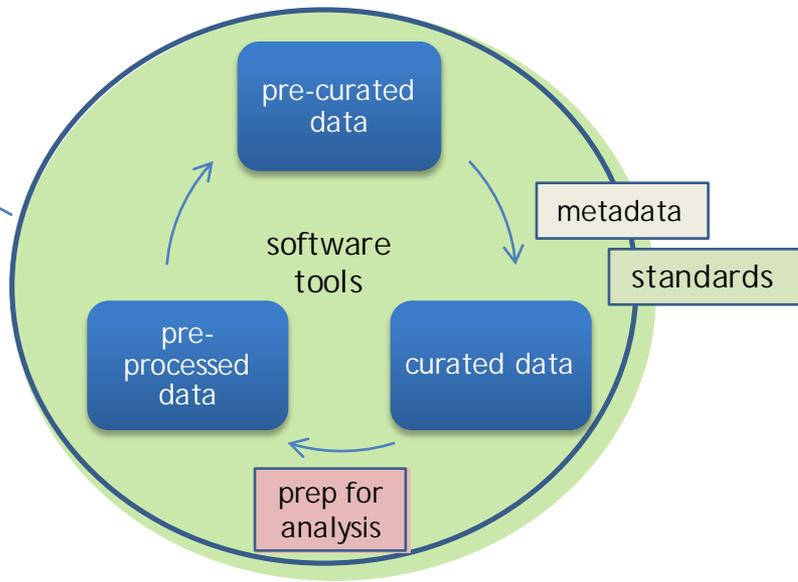
data ecosystem

one woman's pre-processed data are another woman's "raw" data



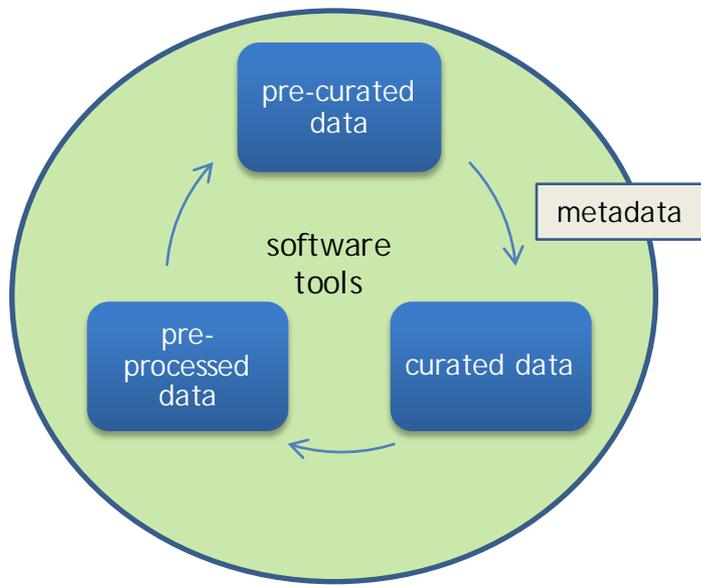
data ecosystem

hosting in a
cloud, cluster,
server

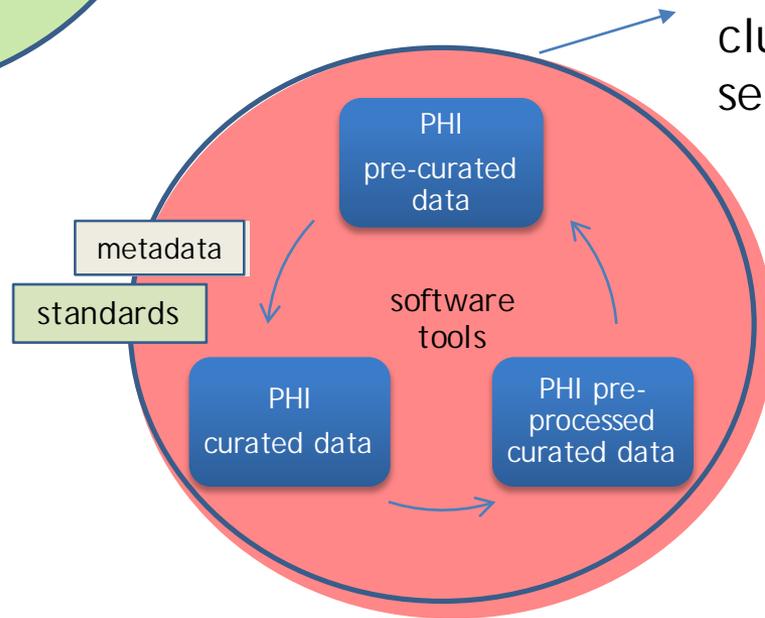


*understanding which
computational
environment is best
for the combination of
data and relevant
tools is important
(e.g., HPC, GPU)*

data ecosystem



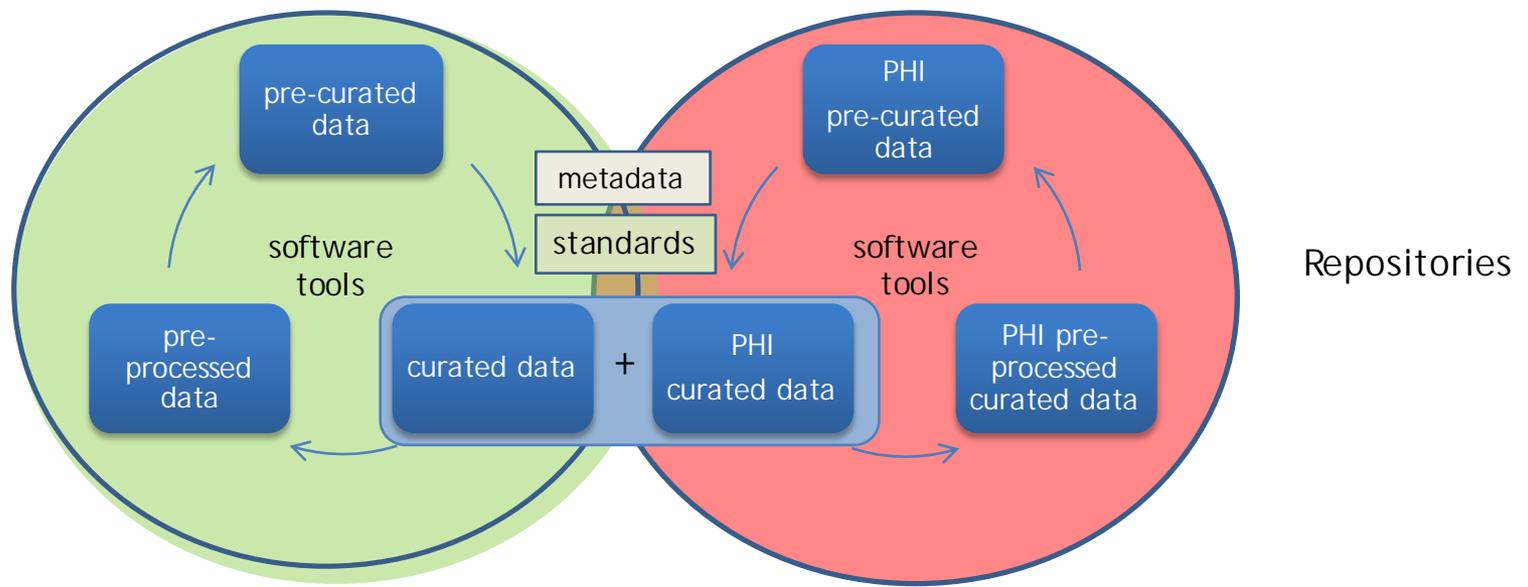
Protected Health Information (PHI) hosting in a HIPAA cloud, cluster, server



selecting the right computational environment for the right type of data is important

understanding the conditions of accessibility is also important

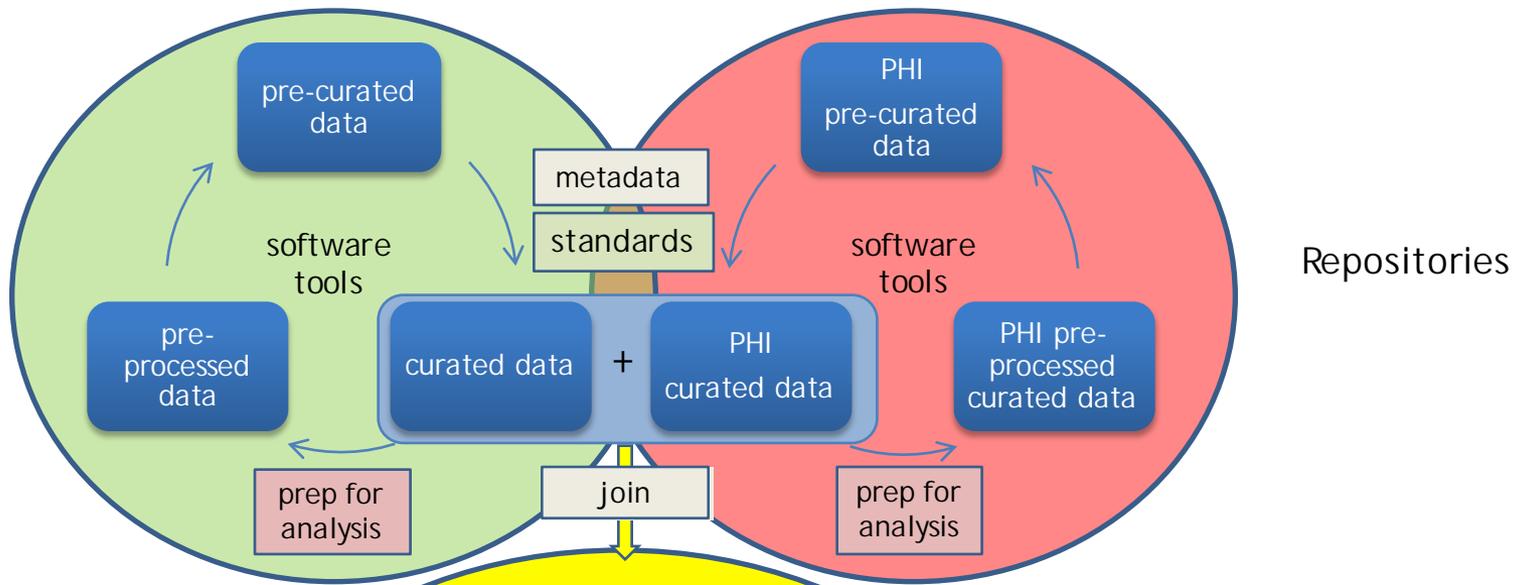
data ecosystem



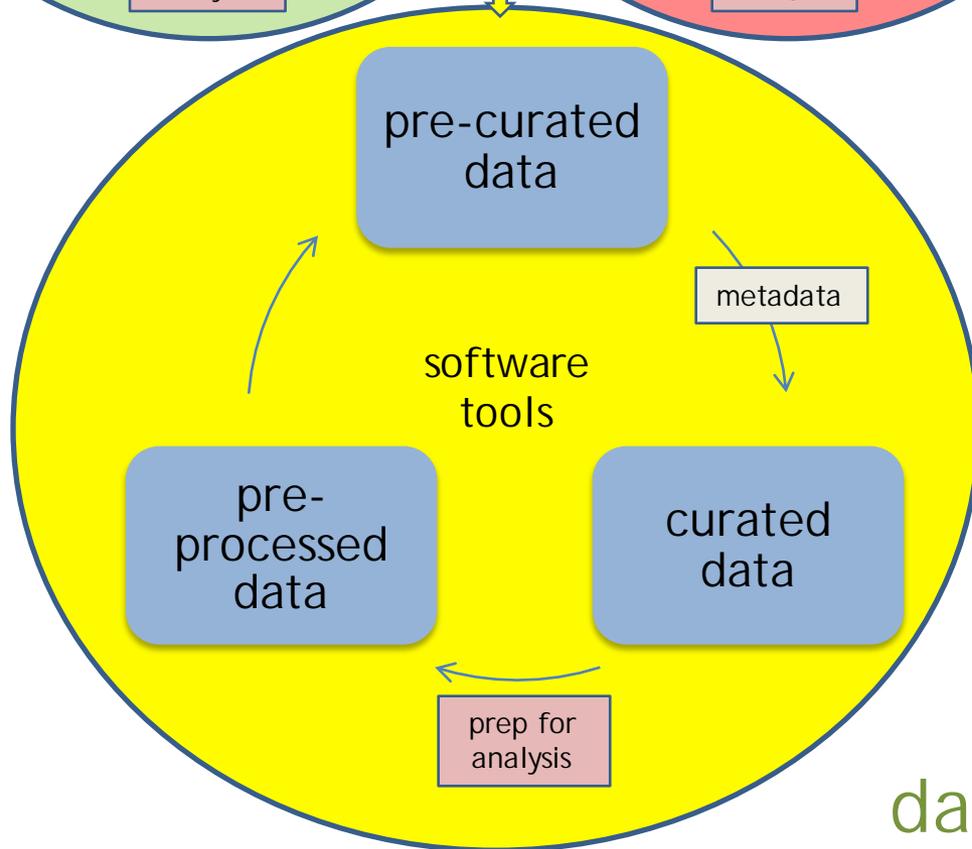
big data analytics depend on

1. *merging data from several different sources (e.g., reference databases, molecular data repositories, clinical repositories),*
2. *proper software, and the*
3. *proper computational environment*

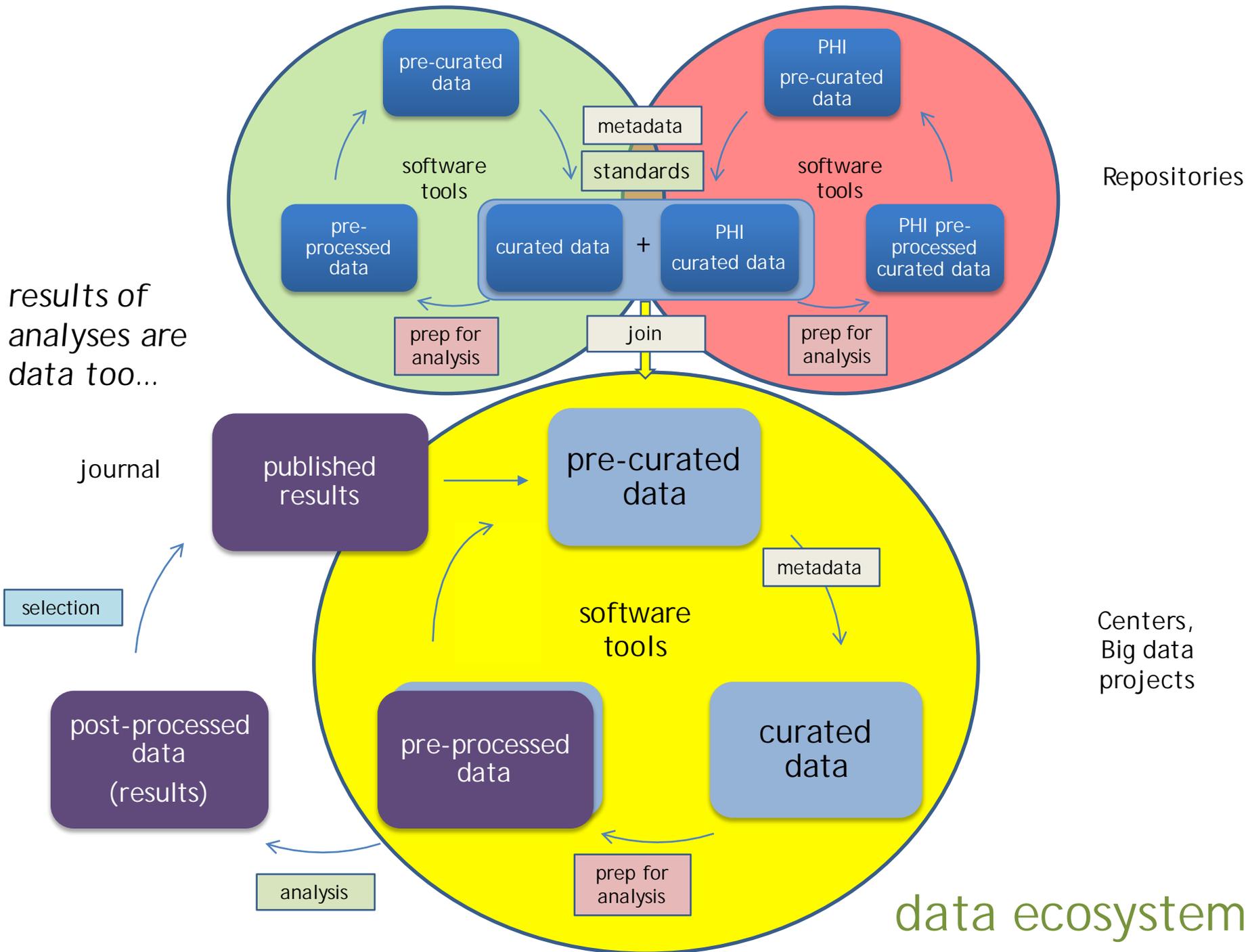
data ecosystem



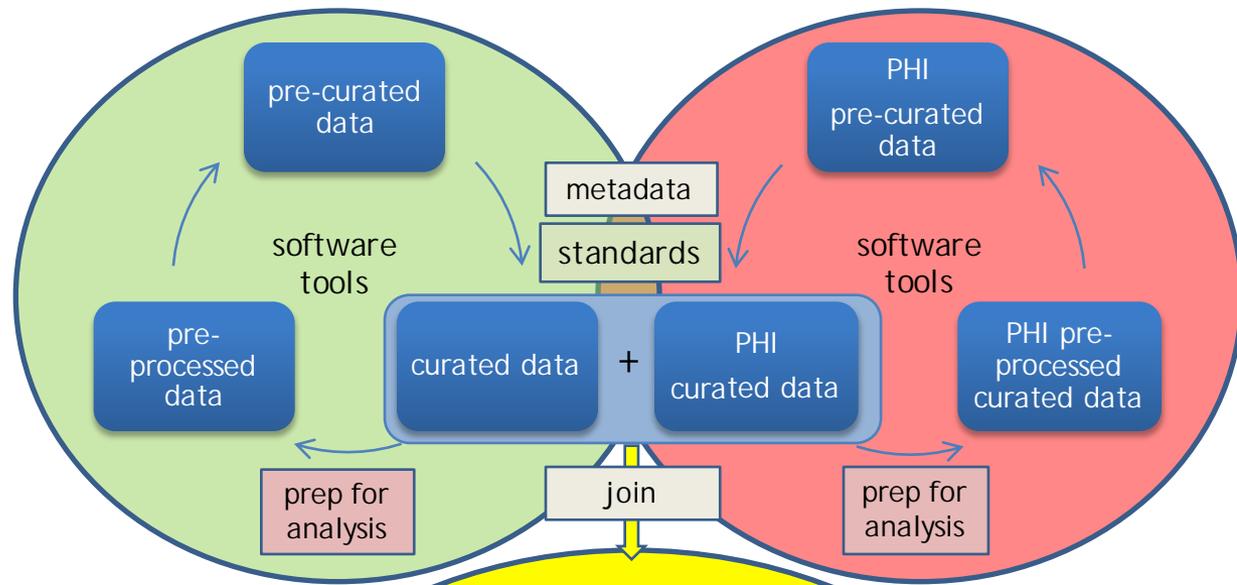
big data projects use several types of digital objects and they are inter-related



data ecosystem

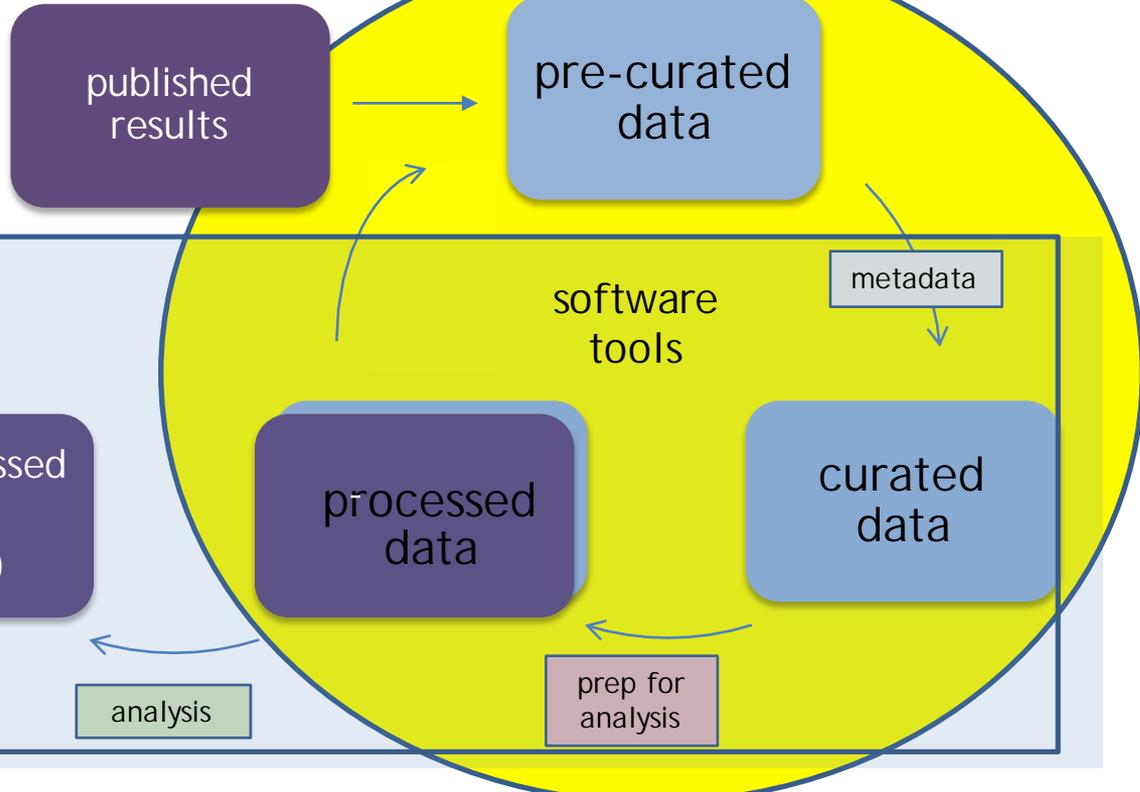


new types of "publications" have emerged



Repositories

journal



Centers,
Big data
projects

data
ecosystem

selection

post-processed
data
(results)

analysis

pre-curated
data

processed
data

curated
data

prep for
analysis

software
tools

metadata

published
results

curated data

PHI
curated data

prep for
analysis

prep for
analysis

metadata
standards

software
tools

software
tools

pre-curated
data

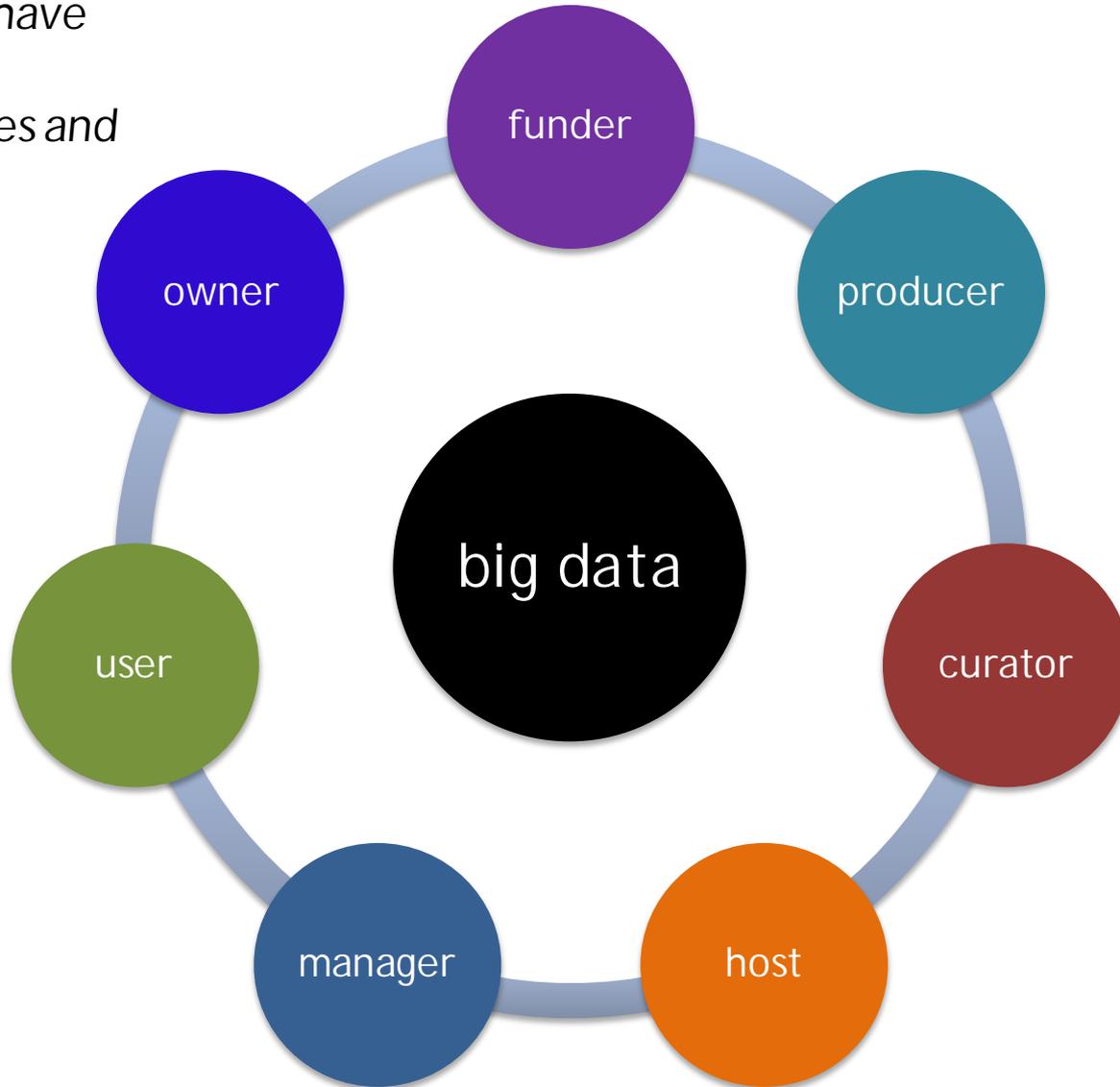
PHI
pre-curated
data

pre-
processed
data

PHI pre-
processed
curated data

Stakeholders

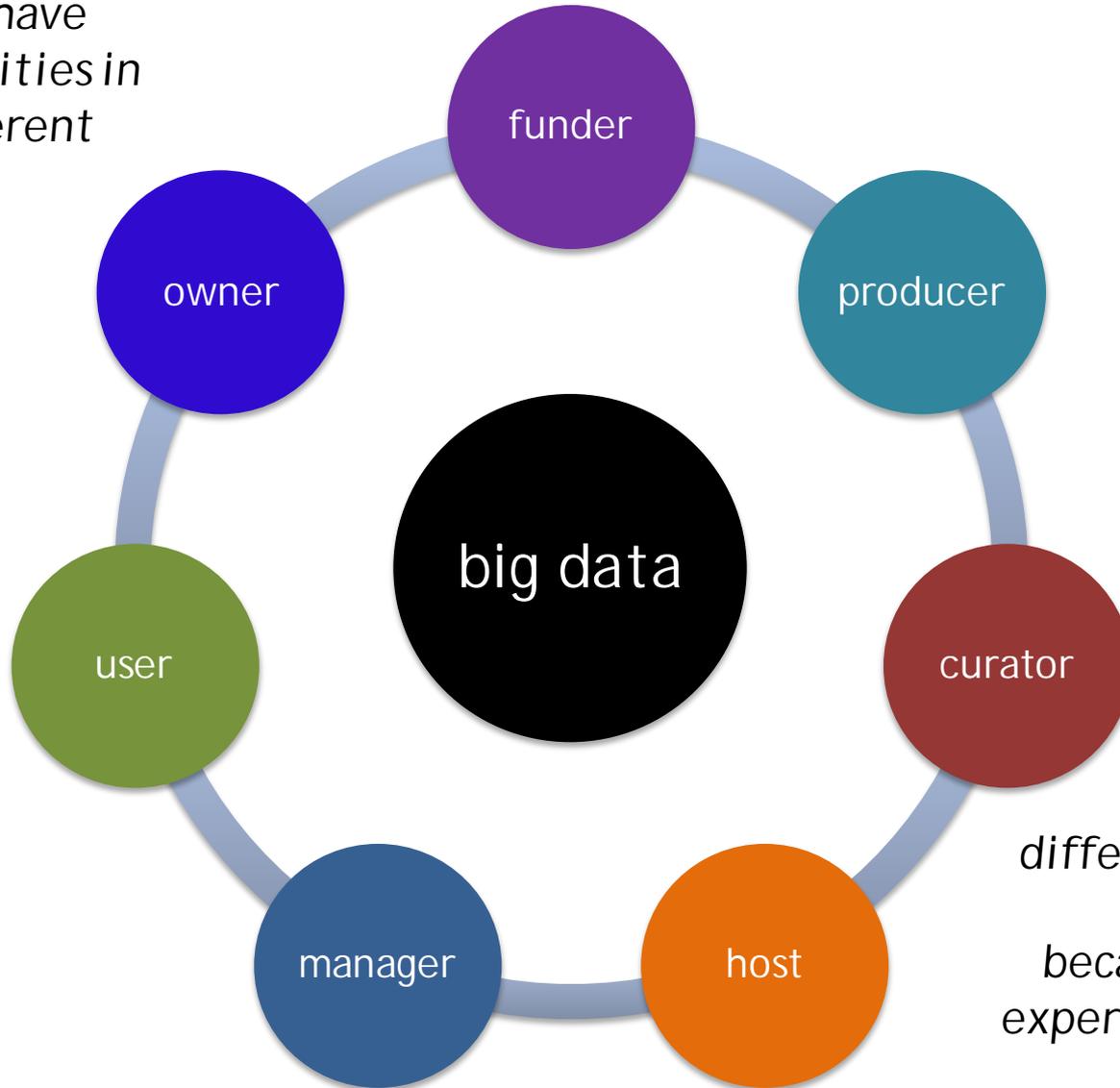
stakeholders have
different
responsibilities and
interests



data ecosystem

Stakeholders

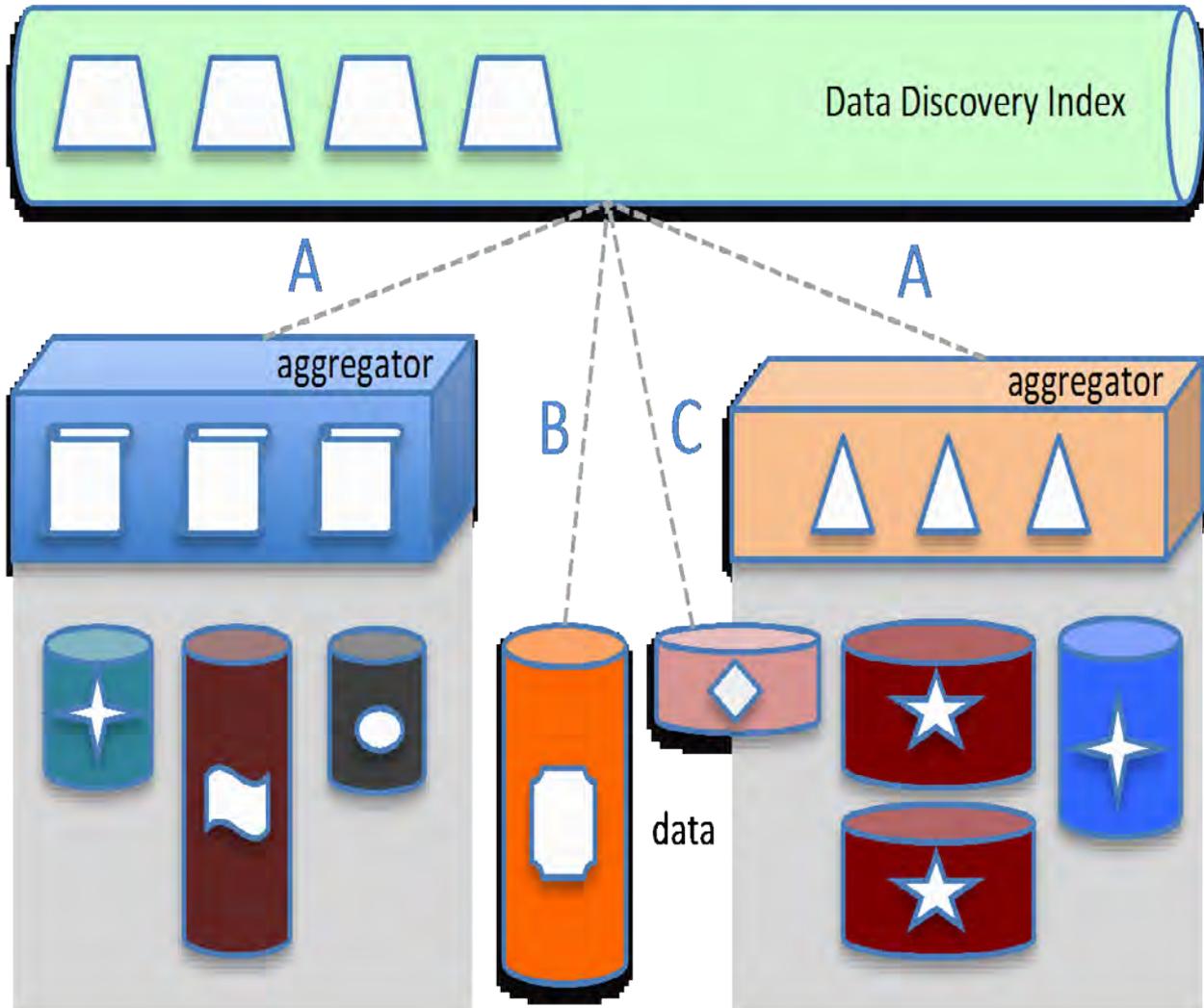
stakeholders have different abilities in indexing different types of data



searching across different resources is time consuming because no one is an expert in all resources

data ecosystem

bioCADDIE



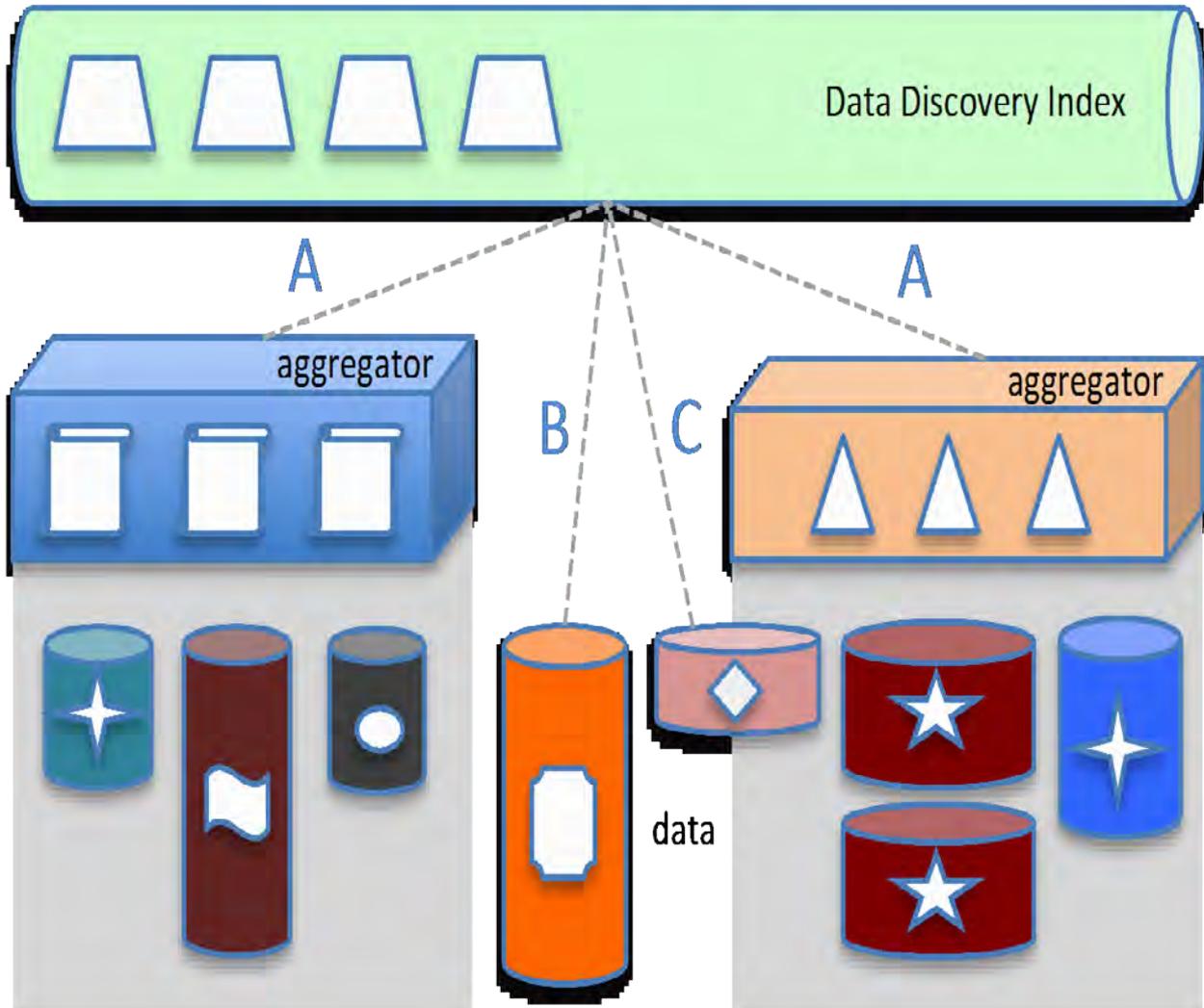
searching across indices and repositories

existing indices can interoperate with the cross-aggregator index

best indexers for data are those who use it all the time, but they may not know as much about other resources

data

bioCADDIE



*"find data on
Kawasaki disease"*

platform and portal

A, B, C: mapping of
metadata, standards,
links to aggregators,
passing of queries

aggregators: various
indices whose
metadata are or can be
mapped into Commons
metadata

data
digital objects

Metadata Model

A set of *metadata specifications*, future-proofed for progressive extensions, to support intended capability of the Data Discovery Index prototype

PHASE 1 OUTPUT:

- NIH BD2K bioCADDIE Data Discovery Index WG3 Metadata Specification v1. 2015. Zenodo. 10.5281/zenodo.28019
 - The WG3-MetadataSpecifications-v1.zip contains a **document**, two **Appendixes**, **JSON schema** and **examples**.

If you wish to provide **comments** on this document, please, use the **live Google version** (no login required). If you are a WG3 member, use the mailing list; if not, please send your comments to [biocaddie\[at\]ucsd.edu](mailto:biocaddie[at]ucsd.edu).

Created using

Competency question

Search for **organism x** in **biological process y** (apoptosis) at **scale z** with an estimate of the **reliability of the annotations**

Search for new **drug x** to predict and track **biological process x** (cardiotoxicity)

Search for **data type x** ('omics correlates) of **biological process** for **drugs related to drug x**

Search for **data types a, b, and c** (EHR data, self-report, sensor) to determine **natural history of patients** given **drugs similar to drug x**

Track **responses to treatment** to ensure detection of **biological process x**

Find **patient data "like these"** with **similar treatments, responses to treatment, genetics**

Search for **studies a-z** with **patient data** with **biological process x** (e.g. **obesity** as measured by BMI) and **interventions a-z**. Then filter on **demographic characteristics**.

bioCADDIE Project: bioCADDIE Collection of Standards biosharing.org
Information Resources

The collection of terminology artifacts and models being considered for the bioCADDIE Metadata Working Group.
This Collection is maintained by: [agbltran](#) [BRCID](#)

Investigation Study Assay Tabular MODEL.FORMAT	MicroArray Gene Expression Tabular Format MODEL.FORMAT	MIAME Notation in Markup Language MODEL.FORMAT	Ontology for Biomedical Investigations TERMINOLOGY ARTIFACT
PRIDE XML Format MODEL.FORMAT	RIF-CS MODEL.FORMAT	Schema.org TERMINOLOGY ARTIFACT	Semanticscience Integrated Ontology TERMINOLOGY ARTIFACT

BioSharing: Content Standards and Databases

bioCADDIE Project: bioCADDIE Collection of Standards

The collection of terminology artifacts and models being considered for the bioCADDIE Metadata Working Group.
This Collection is maintained by: [agbeltran](#)

[Homepage](#) [Reference](#)

biosharing

Investigation Study Assay Tabular
MODEL/FORMAT

- Systems: 5
- Publications: 3
- In Collections: 3

No taxa defined.

9 Data types, including:
REPORT, MATRIX, EXPERIMENT, REAGENT, DEVICE, ASSAY

MicroArray Gene Expression Tabular Format
MODEL/FORMAT

- Systems: 3
- Publications: 1
- In Collections: 1

No taxa defined.

4 Data types, including:
REPORT, EXPERIMENT, DNA MICROARRAY, FILE

MIAME Notation in Markup Language
MODEL/FORMAT

- Systems: 3
- Publications: 0
- In Collections: 2

No taxa defined.

4 Data types, including:
REPORT, EXPERIMENT, FUNCTIONAL GENOMICS, FILE

Ontology for Biomedical Investigations
TERMINOLOGY ARTIFACT

- Systems: 4
- Publications: 1
- In Collections: 0

No taxa defined.

9 Data types, including:
DATA TRANSFER, MATRIX

MIAME REPORTING GUIDELINE

- Systems: 3
- Publications: 1
- In Collections: 1

3 Taxa types, including:
BACTERIA, ARCHAEA, EUKARYOTA

7 Data types, including:
GENOME, DNA, DNA MICROARRAY, TRANSCRIPTOME, RNA

Gene Expression Omnibus

- Standards: 4
- Publications: 1
- In Collections: 6

1 Taxa types, including:
ALL

3 Data types, including:
GENE EXPRESSION, LIFE SCIENCE, GENOME

PRIDE XML Format
MODEL/FORMAT

RIF-CS
MODEL/FORMAT

Schema.org
TERMINOLOGY ARTIFACT

Semantic
TERMINOLOGY ARTIFACT

Data Identifiers

Define a set of best practices and operating procedures for identifiers that support the intended capability of the NIH BD2K Data Discovery Index (DDI) prototype - being designed by the bioCADDIE Core Development Team.

GROUP MEMBERS

Tim Clark - Harvard Medical School, FORCE11 Data Citation Implementation Group
Merce Crosas - Data Science
Michel Dumontier - Center for Expanded Data Annotation and Retrieval, W3C HCLSIG
Claudiu Farcas - University of California, San Diego
Ian Fore - NIH
Carole Goble - The University of Manchester
Alejandra Gonzalez-Beltran - Oxford e-Research Centre, University of Oxford
Jeffrey Grethe - University of California San Diego
John Kunze - California Digital Library
Ron Margolis - NIH
Jo McEntyre - EMBL-EBI, ELIXIR EBI Node, Pub Med Central
Julie McMurry - Monarch Initiative
Philippe Rocca-Serra - Oxford e-Research Centre, University of Oxford
Susanna-Assunta Sansone - University of Oxford and Nature Publishing Group
Heidi Sofia - NIH
Stian Soiland-Reyes - e-Science lab, University of Manchester
Joan Starr - California Digital Library, DataCite
Hua Xu - University of Texas Houston

Check document at biocaddie.org

bioCADDIE Identifiers Working Group WG2 DDI Identifier Recommendations

Deadline for Public Review and Comment: November 9, 2015

Status

Working Group final version released for public comment, 30 October 2015

Introduction

This document is meant to review identifier practices in the context of the bioCADDIE Data Discovery Index (DDI) prototype. The goal of this document is to provide the bioCADDIE Core Development Team the necessary guidelines and operating procedures to produce a functional prototype that enables data discovery and facilitates data sharing and re-use of data found within data repositories and the NIH Commons.

Scope of Identifier Recommendations

The identifier recommendations contained within this document are focused on the need for the DDI to uniquely identify data sets within biomedical data repositories and the NIH Commons. Other aspects of identifiers were deemed out-of-scope for this document. Data citation, for example, is already being addressed by many groups including the FORCE11 Data Citation Implementation Group which is building on the Joint Declaration of Data Citation Principles. The DDI, similar to PubMed, must reference the appropriate citable identifier provided by the data set maintainer.

Intended Audience

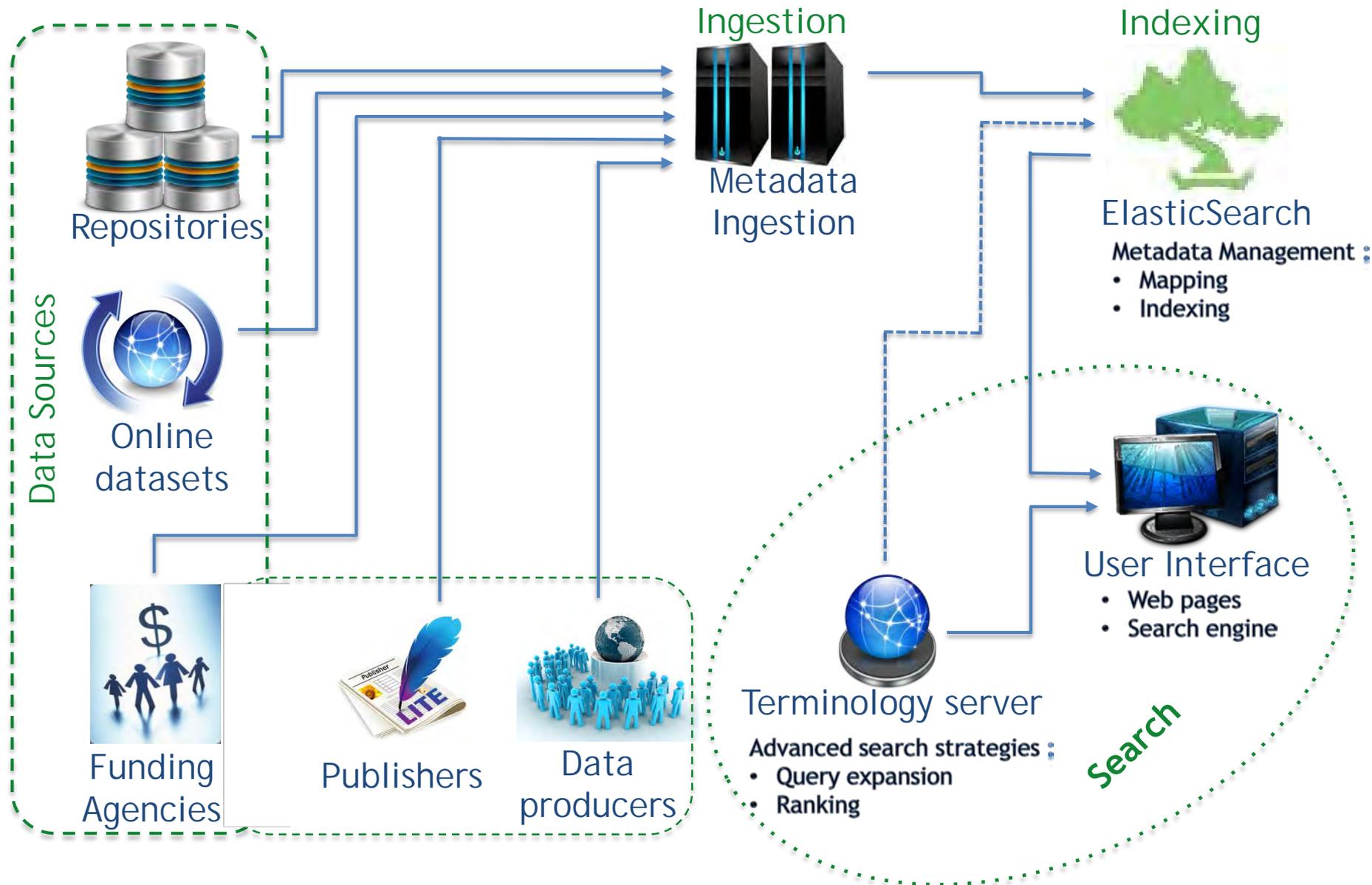
This document is primarily aimed at the bioCADDIE Core Development Team that is implementing the bioCADDIE prototype; however this document may also be informative for other groups such as groups involved with the NIH Common, community aggregators and developers of data harvesting and other tools.

Background

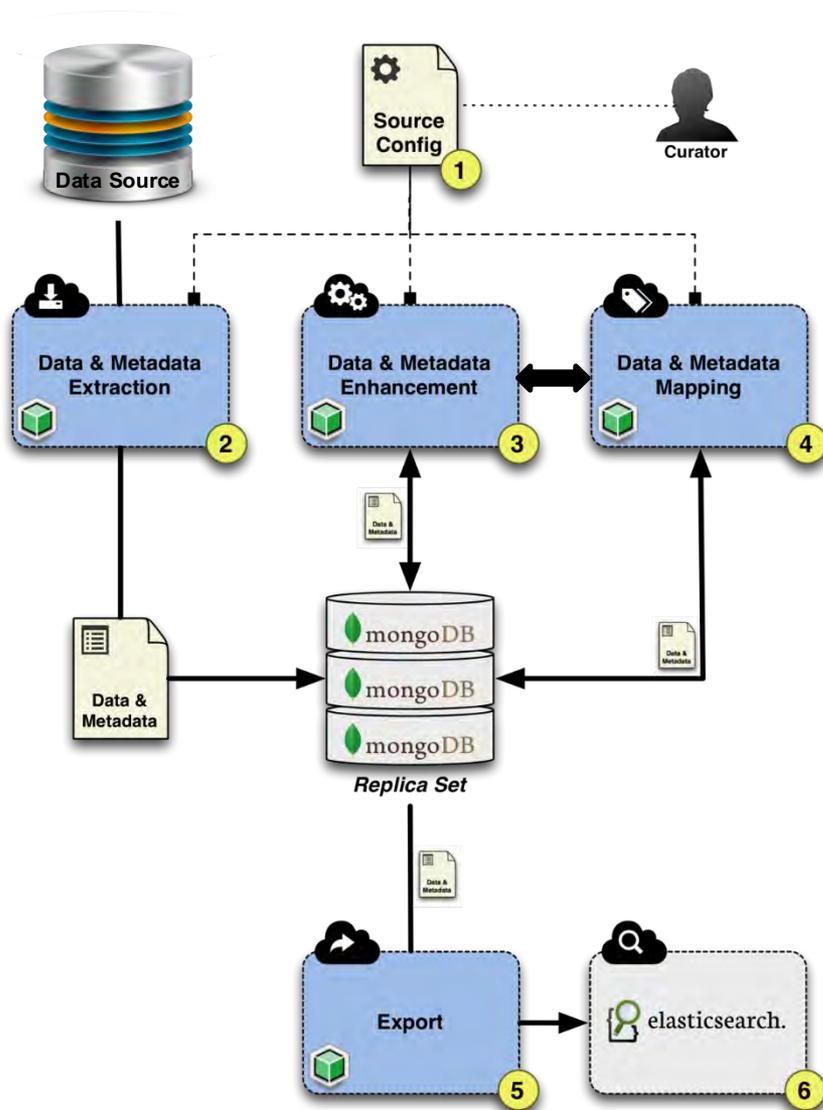
In the context of the NIH BD2K bioCADDIE project, we aim to provide recommendations for the appropriate handling of identifiers for datasets and data repositories (repository as used in this report refers to a source of data/datasets) within the bioCADDIE prototype. This document was generated from the discussions of bioCADDIE Working Group 2. During these discussions a number of identifier related topics were described as being out-of-scope for the current work of this working group – many of these topics potentially relate to the work of other active or future bioCADDIE working groups or broader community working groups.

Attend breakout session on Identifiers

bioCADDIE Prototype

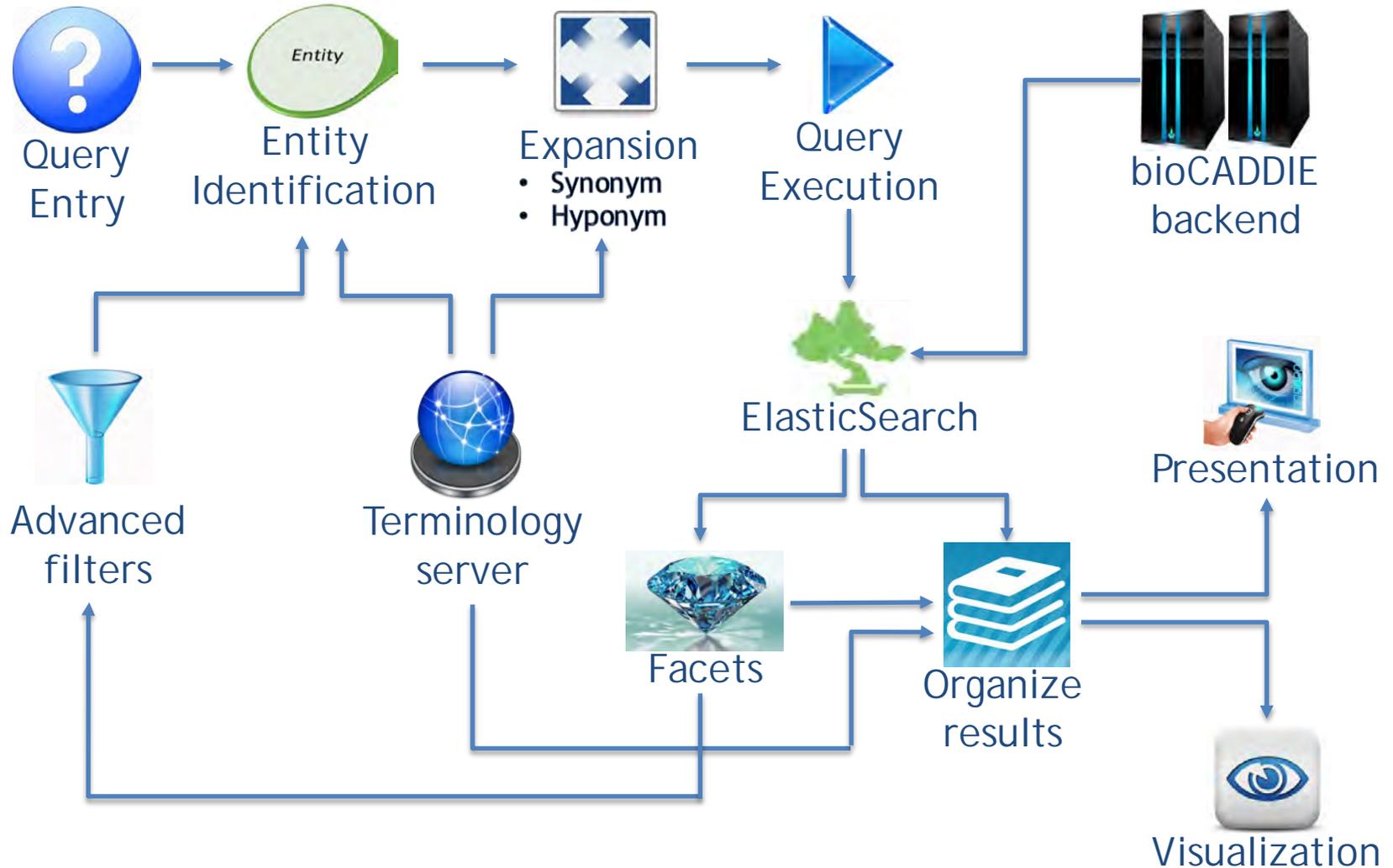


Data Indexing Pipeline



1. Configuration file developed by curator
 - ◆ Cache information for further processing
2. Extraction of metadata/data from data resource or dataset via ingestion module
 - ◆ e.g. ID conversion, keyword extraction, data normalization
3. Process metadata/data via a set of modules
4. Mapping of metadata/data to metadata model(s)
5. Export to target endpoint(s)
6. Search via Elasticsearch APIs

User Interface Workflow



Engaging The Community Toward a Data Discovery Index (v0.2)

Search for data through BioCADDIE



Search Examples: (Breast Cancer, Genetic Analysis Software, Gene EGFR)

Statistics



7 REPOSITORIES



4 DATA TYPES

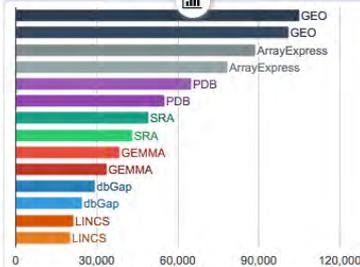


391,776 DATASETS



4 PILOT PROJECTS

Repositories



Latest Datasets

Coming Soon

ramaswamy-cancer



dbGap

A Genome-Wide Association Study of Lung Cancer Risk

GEO

A Genome Wide Scan of Lung Cancer and Smoking

ArrayExpress

Estrogen Receptor Positive Breast Cancer: Aromatase Inhibitor Response Study

PDB

A Multiethnic Genome-wide Scan of Prostate Cancer

LINC

New Features



- September 16, 2015
 - » The interface has been updated
 - » Global statistics feature has been added
- September 16, 2015
 - » First release
- September 16, 2015
 - » The interface has been updated
 - » Global statistics feature has been added

Announcements



- September 16, 2015

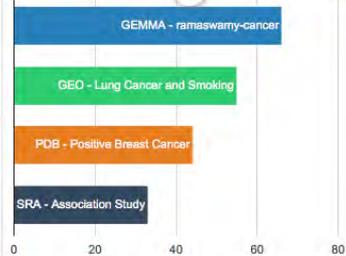
Feedback Update: If you are having problems using our tools, or if you would just like to send us some feedback, please ([Read More...](#))
- October 12, 2015

New Version Released. We just released our new interface design with new features. ([Read More...](#))
- September 16, 2015

Feedback Update: If you are having problems using our tools, or if you ([Read More...](#))

Most Accessed Datasets

Coming soon



Pilot Projects



GWAS Finder

Search literatures for "Genome-Wide Association Studies".



iSEE-DELVE

Search Visualization project for Big Data.



DataBank

Find most suitable datasets for you.



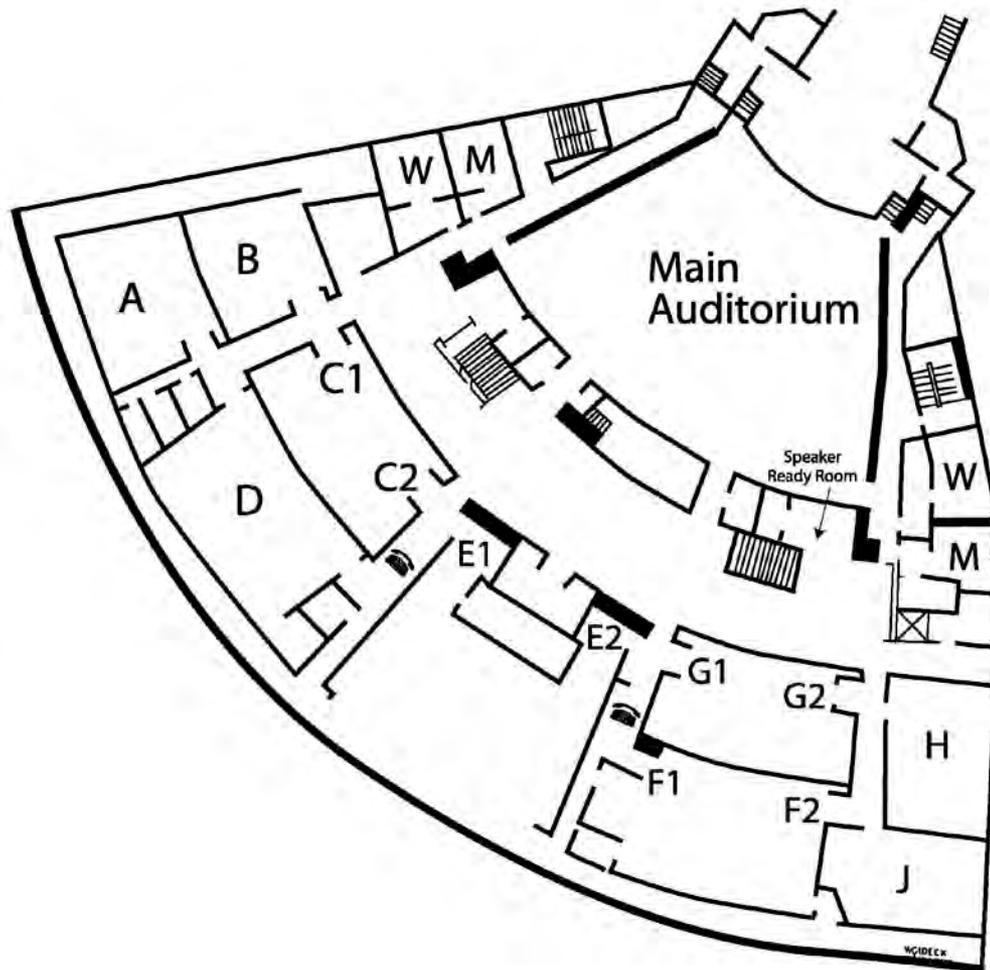
Data Citation Discovery

Coming Soon

Citation and Data Access Metrics Development applied to RCSB Protein Data Bank.

Demo and Posters

Location -
Room E1 2-4pm



Data and
Software
Indexing

E1

Commons &
Interoperability

E2

Core Development Roadmap

DDI architecture

- Setup website for searching for datasets
- Set up infrastructure for web portal

Data identifier

- Implement Data identifier into the DDI

Data indexing

- Set up indexing using metadata from WG 3.0

September 2015

Version 0.1

Data ingestion

- Determine datasets
- Decide on scalable data/metadata input routes
- Metadata mapping

Search function

- Implement the function for 3 repositories

Feedback collection

- Github

RFA for pilot on Harvester for DDI schema

- RFA announced
- Review, selection and award

Wrap up of Y1 pilot projects

- Literature/dataset link: Advanced search
- Recommender System: Ranking results
- iSEE/DELVE: Innovative visualization
- PDB citation pipelines

Core Development Roadmap

Dataset result display

- Sort datasets
- Group metadata

Terminology server

- Import ontology
- Integrate to Scigraph API
- Integrate autocomplete feature to prototype

Interface design

- New interface for prototype v 0.2
- Global statistics

Usability study

- UI Analysis

Ranking algorithm

- Results from pilot project

Search function

- Expand the function to 7 repositories
- Find similar datasets
- Search history

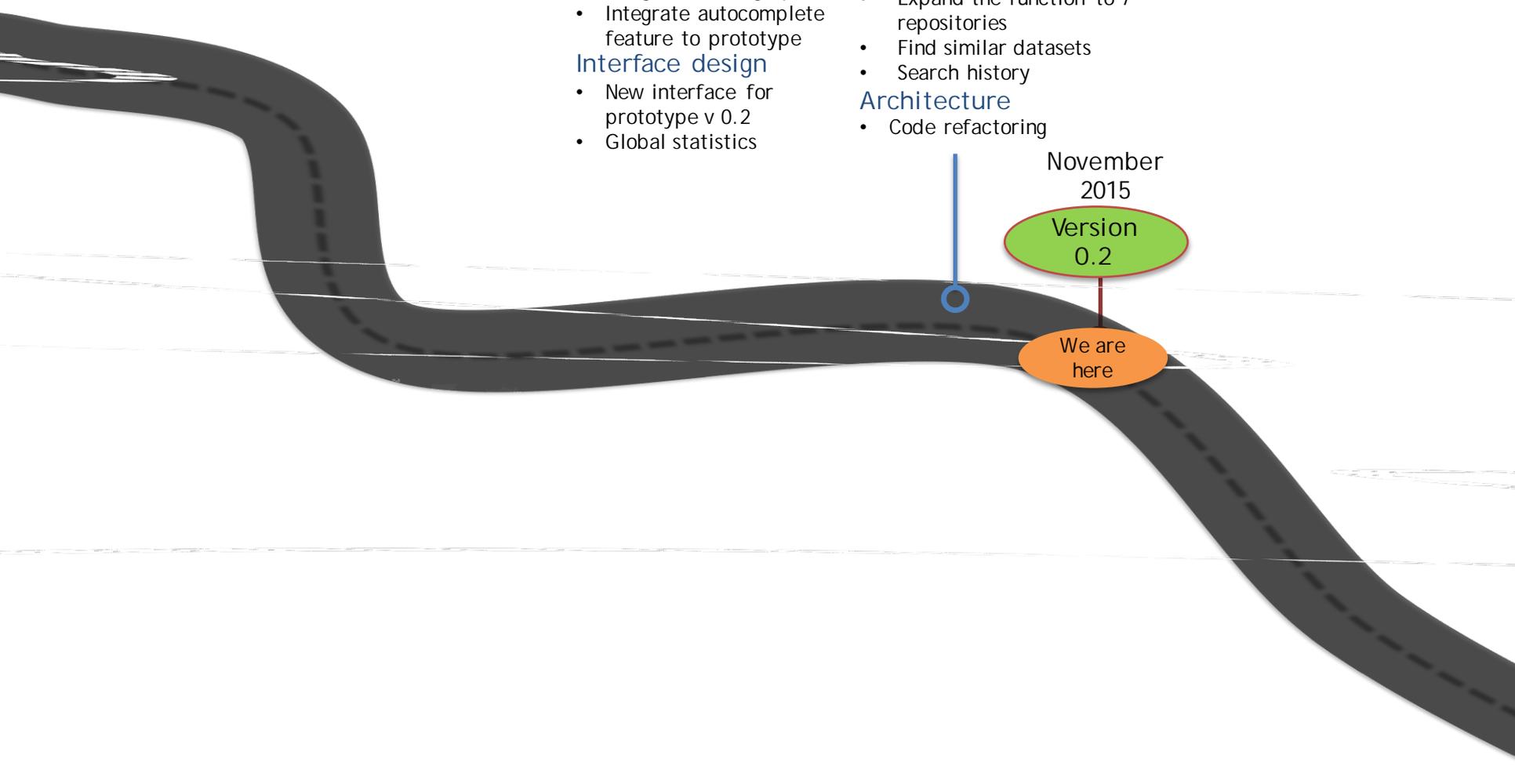
Architecture

- Code refactoring

November
2015

Version
0.2

We are
here



Core Development Roadmap

Personalized search

- Share/save search results
- User account

Link dataset to external resources

- PubMed
- Grants

Search algorithm

- Boolean/advanced search
- Data repository search function

February
2016

Version
0.5

June
2016

Version
1.0

Usability study

- User study
- Track user's action

Ranking algorithm

- Refine search results based on user's selection
- Report from WG 8 on Ranking

Data duplication problem

Metadata management

Participation

- ❖ Working groups
 - ◆ Participate or follow
- ❖ Prototype
 - ◆ Using and providing feedback on the prototype search engine
- ❖ Interoperate with the prototype
 - ◆ Link your favorite index
 - Use or map to metadata
 - Collaborate on APIs
- ❖ Recommend data/repositories for inclusion
 - ◆ New working group

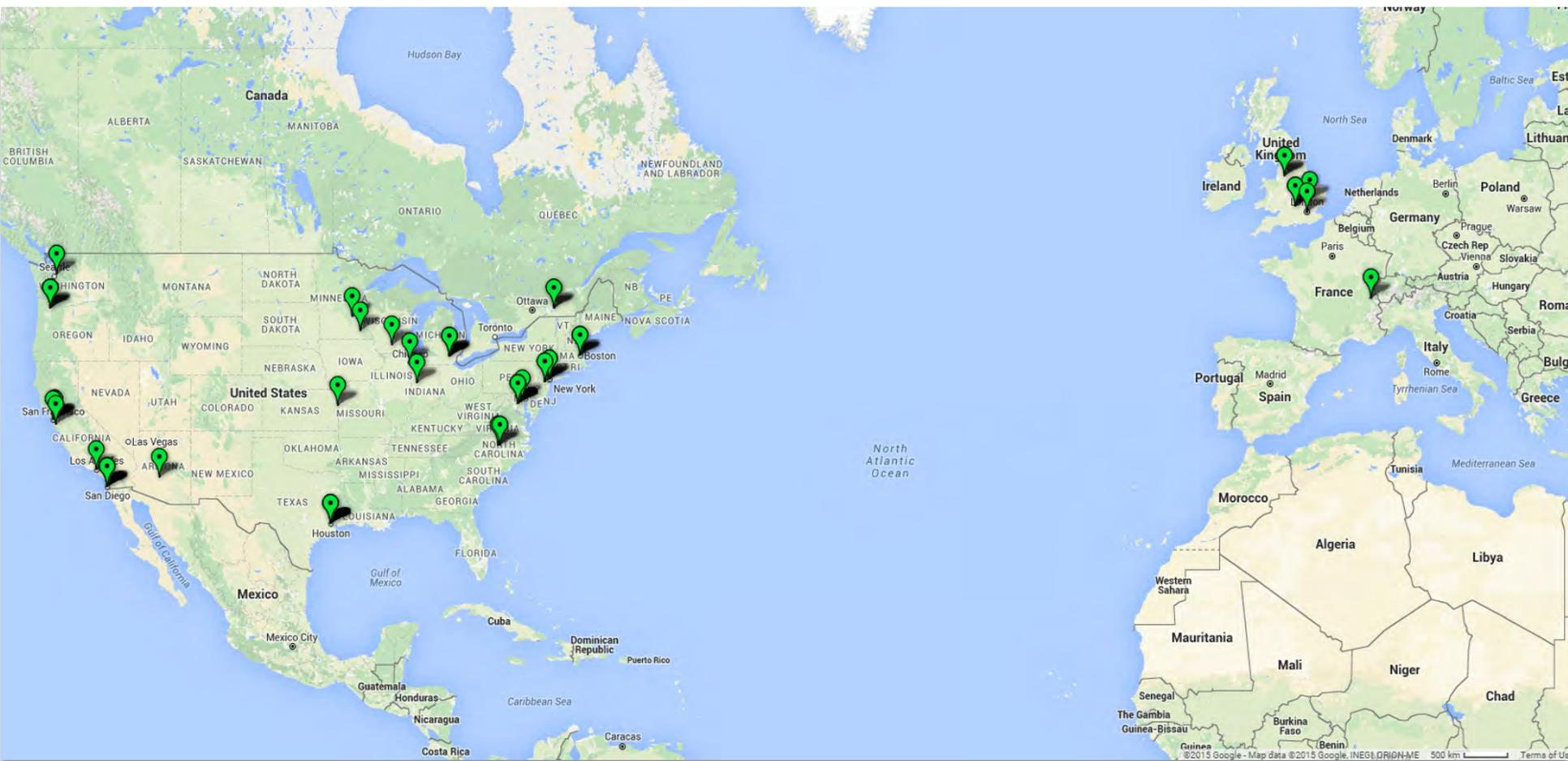
Newly awarded - Metadata Discover

- Distributed data discovery using gym: github, yaml and markdown
Chris Mungall, Lawrence Berkeley National Laboratory
- Feasibility study of indexing clinical research data using HL7 FHIR
Guoqian Jiang, Mayo Clinic College of Medicine
- Metadata discovery and integration to support repurposing of heterogeneous data using the Openfurther platform
Ram Gouripeddi and Julio Facelli, University of Utah

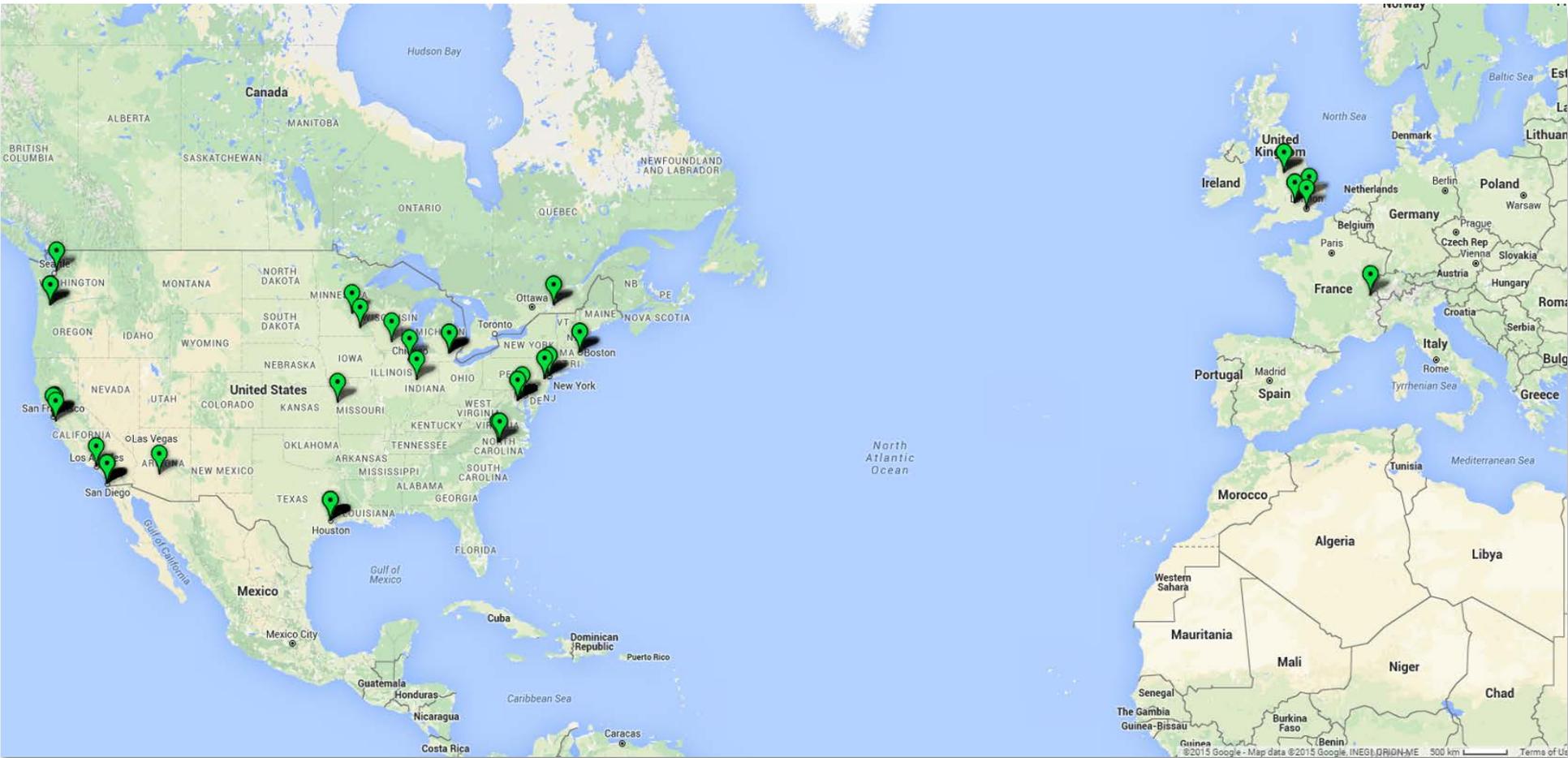
Results of Ranking (1: top proposal)

Project	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6	Reviewer 7	Reviewer 8	Sum
Project 1	2	NR	NR	1	NR	1	1	1	6
Project 2	3	2	NR	2	1	2	NR	NR	10
Project 3	4	1	NR	NR	2	NR	2	3	12
Project 4	1	4	1	NR	4	NR	4	NR	14
Project 5	NR	3	2	NR	3	3	3	NR	14
Project 6	NR	NR	4	3	5	4	NR	2	18
Project 7	5	NR	3	4	NR	5	NR	4	21

Working Groups

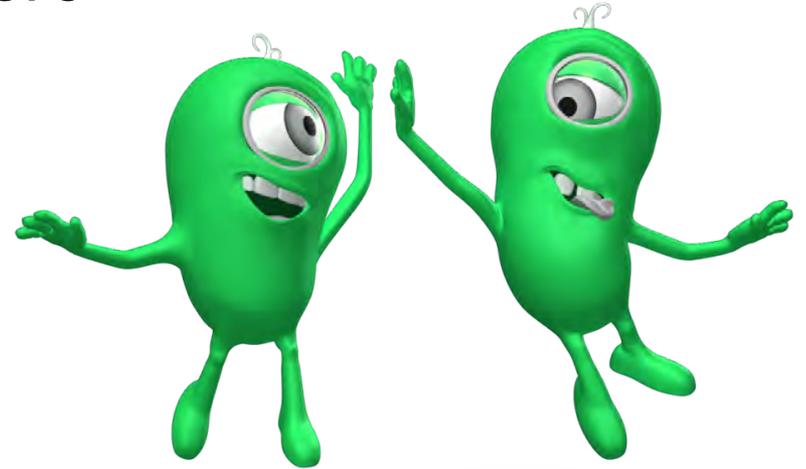


Working Groups



Acknowledgements

- 93 working group members
- 12 steering committee members
- 8 pilot application reviewers
- staff and trainees
- collaborators



a mouse model for data science

