biomedical and healthCAre Data Discovery Index Ecosystem

Lucila Ohno-Machado, MD, PhD

NIH BD2K grant U24 AI117966

BD2K all-hands meeting
Bethesda, 11/29/16
find all data about astrocytomias with TP53 mutations
Data Discovery Index Consortium

a BD2K program

1. Help users find shared data
2. Build a prototype data discovery index
3. Interoperate in the NIH Commons

FAIR: Findability, Accessibility, Interoperability, Reusability

Do for Data what PubMed did for papers

bioCADDIE started on 3/8/15
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
Activity areas

1. **Data Shop**
   - Indexing, Linking data to articles
   - Software dev
   - New ideas via Pilot Projects, Supplements

2. **Data Showcase**
   - Search Engine Prototype
   - User Feedback

3. **Data Market**
   - Incentives to Share, Data Citation
   - Outreach

Illustration by pixabay
Data Shop and Showcase

Repositories

Data Producers

Indexing, Outreach

Researchers

Illustration by pixabay

Supported by the NIH grant 1U24 AI117966-01 to the University of California, San Diego
DataMed Users
Metadata Adoption: 2 fronts

**Annotating existing data**
- Continue to work with data repositories to map into a minimal standard
- Incentives for data producers/repositories to facilitate mapping
- Incentives for data reuse/citation

**Annotating new data**
- Could be done at the source, like publishers do for JATS
- Additional resources need to be provided for data producers/repositories to prepare data for sharing (e.g., after grant funding period ends)

Resources for data producers and/or repositories to maintain data and their annotations are needed

Leveraging resources from various paid projects, consolidation of efforts, and incentivizing data producers/keepers saves time and money
Components

1. Working Groups
2. Metadata Specifications
3. Software Development
   1. Search engine
   2. Indexing processes
4. Collaborations - supplements, pilots
5. Governance
Working groups

- Evaluation
- Use Cases
- Data Citation
- Link to Publications
- Criteria for Inclusion
- Identifiers
- Graphical Interface
- Results Ranking
- DAta Tag Suite (DATS)
- Accessibility Metadata
- Metadata Harvester Pilots
- Application Programming Interfaces
We work together

All materials on biocaddie.org are open to comments, all open source code in GitHub
Integrating indexing efforts across BD2K

Organizing framework and portal for data

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

Data
Digital objects

Core development team

Pilot projects

Data Discovery Index

aggregator

aggregator

B

C

data

Integrating indexing efforts across BD2K

CEDAR
OmicsDI and bioCADDIE

Administrative supplement 2016

- PIs
  - Peipei Ping, UCLA
  - Susanna Assunta-Sansone, U Oxford
  - Eric Deutsch, ISB
  - Henning Hermjakob, EBI

- WPs
  - Map OmicsDI, bioCADDIE data model
  - Re-usable visualization widgets
  - Access metrics

Collaboration

- Mapping from OmicsDI data model to DATS model
- OmicsDI provides metadata from “its” repositories to bioCADDIE
- OmicsDI goes more into the “depth” for omics
- bioCADDIE focuses on breadth

by Hermjakob
Data Citation Implementation Pilot

Provide basic coordination between publishers, repositories and identifier / metadata services for early adopters of data citation according to the Joint Declaration of Data Citation Principles.
Secure aggregating counts of relevant patients

Select count(*) from MD2K A, i2b2 B, GA4GH C where 
A.blood Glucose>110 and B.smoker=true and C.r123140=C
BioCADDIE Harvester Project:
Distributed Data Discovery using GitHub, YAML and Markdown

BioCaddie All-Hands
Denver, Sep 2016

Chris Mungall
Lawrence Berkeley National Laboratory

Metadata Discovery and Integration to Support Repurposing of Heterogeneous Data using the OpenFurther Platform

bioCADDIE All Hands Meeting
September 11th, 2016

Ram Gouripeddi
&
Julio Facelli

Pilot Projects

Feasibility Study of Indexing Clinical Research Data Using HL7 FHIR
A Pilot Project on Harvester for DDI Schema

Harold R. Solbrig, MS, Guoqian Jiang, MD, PhD

Mayo Clinic
September 11, 2016

Supported by the NIH grant 1U24 AI117966-01 to the University of California, San Diego
2016 BIOCADDIE DATASET RETRIEVAL CHALLENGE

Timeline

- Registration: Begins September 9, 2016
- Datasets and Sample Queries Release: September 16, 2016
- Test Queries Release: October 21, 2016
- System Outputs Due: October 28, 2016
- Workshop: TBD

The primary objective of this challenge is to create innovative ways for biomedical researchers to search and discover biomedical research data. Biomedical research's increasing dependence on digital data has led to a significant increase in the number of datasets available to researchers. Finding relevant datasets amid the massive quantity available requires new methods of information retrieval. Dataset searches can involve specific and complex queries that are not typically answered by the metadata associated with these datasets (such as organism and assay type). An example is a user who wishes to know what datasets are available that have genome data about IDH1 and IDH2 in humans for glioma.

bioCADDIE 2016 Dataset Retrieval Challenge Sign Up

Sign up to participate in the bioCADDIE 2016 Dataset Retrieval Challenge. More information will be provided after registration.

Registration begins: September 9, 2016
DataSets and Sample Queries Release: September 16, 2016
Test Queries Release: October 21, 2016
System Outputs Due: October 28, 2016
Workshop: TBD
Identifiers Working Group Timeline

Jun 2015
Working Group Launched

Aug - Oct 2015
Develop Draft Recommendations

Jan - Feb 2016
Finalize Draft Recommendations

• we will not mint identifiers for datasets, rather it will rely on the identifiers provided by the source
• all data sets referenced must be able to be resolved to the full URI for the data set.
• we will maintain a set of landing pages for each of the datasets indexed
• we will maintain landing pages for datasets that may no longer exist
Metadata Group 3 Representation

Members

Institutions

ICPSR, University of Michigan, University of Texas Health Sciences Center, NCBI, Rutgers, Roche, Monarch Initiative, NIH, Center for Expanded Data Annotation and Retrieval, University of Manchester, University of California San Diego, University of Oxford, University of Utah, NLM, Yale School of Medicine, NYU School of Medicine, California Digital Library, NCTR FDA, ELIXIR-NL, W3C ScheMed WG, Johns Hopkins University
**WG4: Use Cases and Testing Benchmarks**

**Usability Analysis Recommendations**

**PHASE I**
Collect feedback from online users that use the DDI interface

**PHASE II**
Gather information from beta (test) users about researchers’ dataset seeking needs, strategies, and challenges

**PHASE III**
Determine if integrating several data sources in one place improves discoverability by comparing bioCADDIE to researchers’ standard approach to finding data

Supported by the NIH grant 1U24 AI117966-01 to the University of California, San Diego
Executive Summary:
Repositories considered for inclusion in the BioCADDIE Prototype (DataMed) will be evaluated on five dimensions. Key aspects of those dimensions are:

• **Prominence**: Priority should be given to types of data that testers of the prototype will expect to find. Less prominent repositories should be considered if they introduce diversity.

• **Sustainability**: Repositories should actively maintain their holdings and have sustainable funding.

• **Scope**: Additional repository included in bioCADDIE should introduce diversity and new data.

• **Quality**: Data sets should be accompanied by sufficient metadata to make them findable, usable, and interoperable. Data formats should follow community standards.

• **Access**: Data should be accessible to the scientific community. Metadata must be in a format usable by the bioCADDIE team for ingest to the DataMed index. Objects must be uniquely identified and web-resolvable by persistent identifiers.
How do I get access to a data set?

**Authorization**
- Obtain permission
  - none
  - clickLicense
  - registration
  - dualIndividual
  - dualInstitution

**Authentication**
- Validate your identity
  - none
  - simpleLogin
  - multiLogin

**Access**
- Use the object
  - download
  - remoteAccess
  - remoteService
  - enclave
  - notAvailable
Acknowledgements

• bioCADDIE consortium

• >100 working group members

• 12 steering committee members

• 8 pilot application reviewers

• staff and trainees

• collaborators
Our journey

If you want to go fast, go alone
If you want to go far, go together