



biomedical and healthCAre
Data Discovery Index Ecosystem

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

5th National Data Service Consortium WorkshopBarcelona

4/5/16

bioCADDIE Overview

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*



SCIENTIFIC DATA

The FAIR Guiding Principles for scientific data management and stewardship

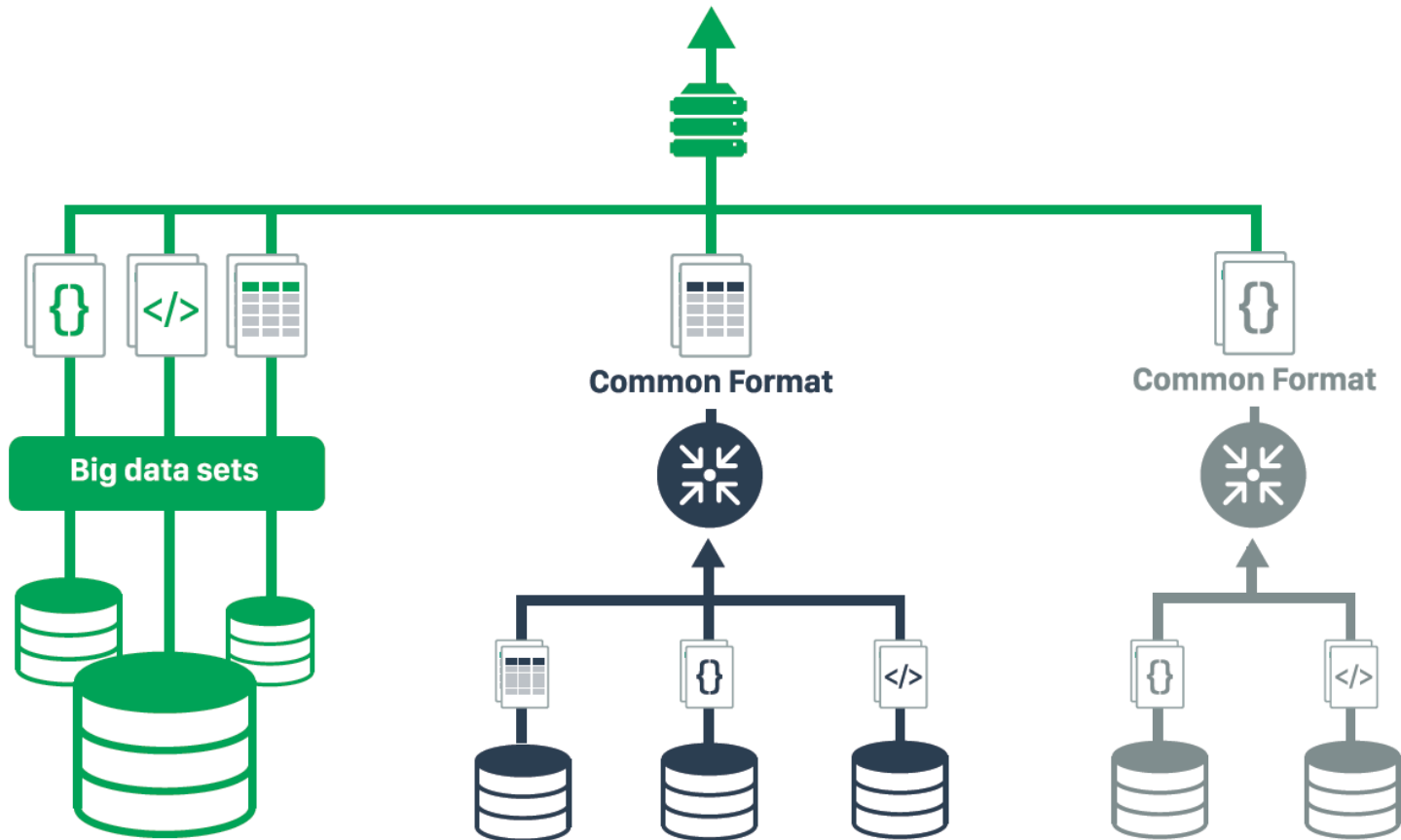
Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Boume, Jildau Bouwman, Anthony J Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J G Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons

Open
data
is about
MORE
THAN
DISCLOSURE
it must be
Fair

- Findable
- Accessible
- Interoperable
- Reusable

<http://www.nature.com/sdata/> 

Data Discovery Index



Big data sets of particular interest to NIH and not covered by aggregators.

e.g. NIH Commons

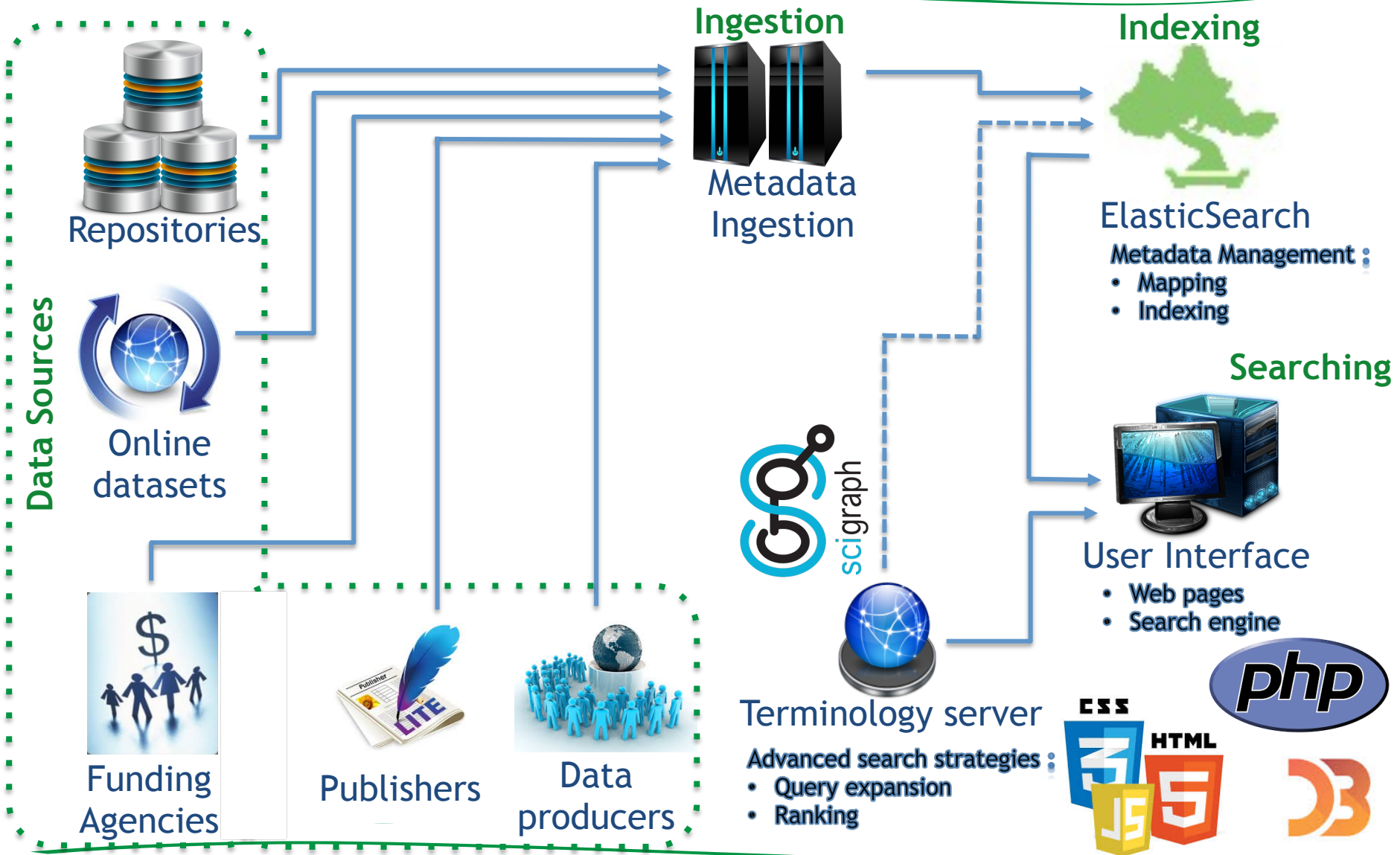
Major aggregator services (i.e., indices or repositories that use a common metadata format)



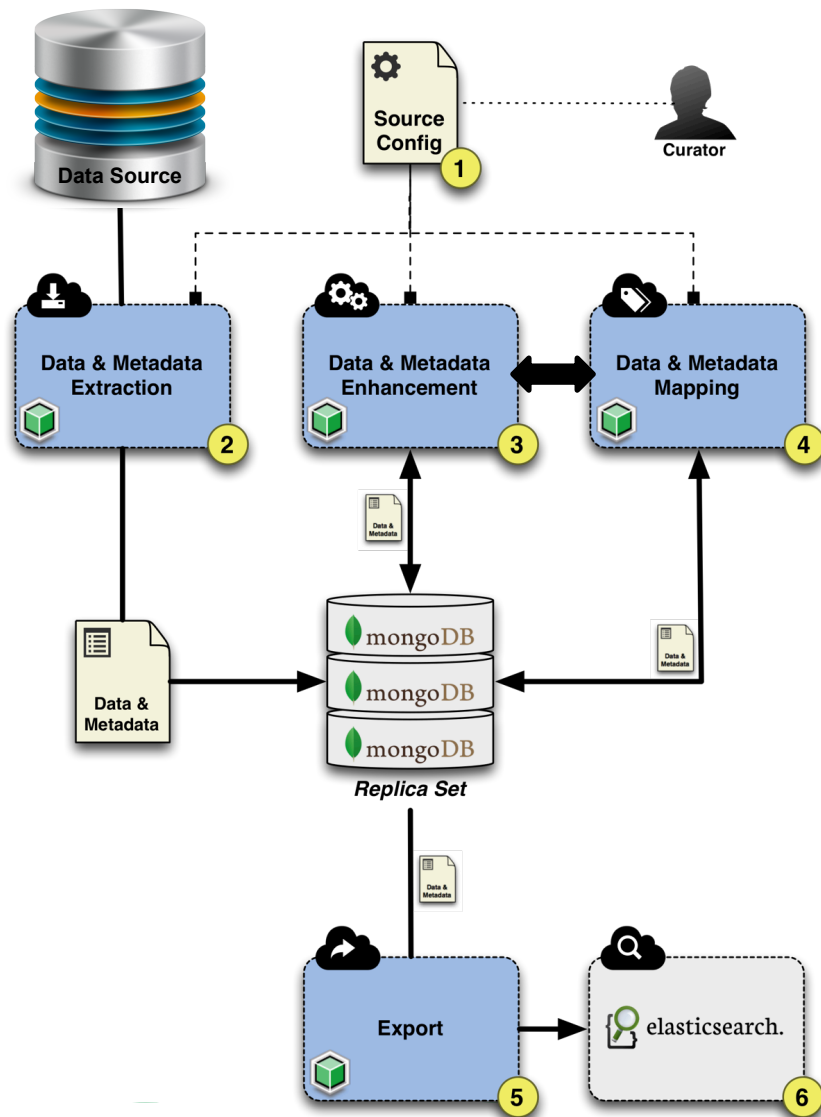
Data Repositories...



bioCADDIE Prototype Architecture



Data Indexing Pipeline



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

WG3: Metadata Specifications

Metadata specification v1, future-proofed for progressive extensions, to support intended capability of the DDI prototype

PHASE 1 OUTPUT:

- NIH BD2K bioCADDIE Data Discovery Index WG3 Metadata Specification v1. 2015. Zenodo. [10.5281/zenodo.28019](https://zenodo.org/record/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 2 complementary approaches

top-down: analyzing use cases

bottom-up: mapping existing standards/schemas

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

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

[biosharing.org](#)
Information Resources

<p>Investigation Study Assay Tabular MODEL/FORMAT</p> <p>Systems: 1 Publications: 1 In Collections: 1</p> <p>No taxa defined.</p> <p>9 Data types, including: REPORT, MATRIX, EXPERIMENT, READOUT, DEVICE, ASSAY</p>	<p>MicroArray Gene Expression Tabular Format MODEL/FORMAT</p> <p>Systems: 1 Publications: 1 In Collections: 1</p> <p>No taxa defined.</p> <p>4 Data types, including: REPORT, EXPERIMENT, DATA MICROARRAY, FILE</p>	<p>MIAME Notation in Markup Language MODEL/FORMAT</p> <p>Systems: 1 Publications: 1 In Collections: 1</p> <p>No taxa defined.</p> <p>4 Data types, including: REPORT, EXPERIMENT, FUNCTIONAL GENOMICS, FILE</p>	<p>Ontology for Biomedical Investigations TERMINOLOGY ARTIFACT</p> <p>Systems: 1 Publications: 1 In Collections: 1</p> <p>No taxa defined.</p> <p>9 Data types, including: DATA TRANSFORMATION, REPORT, DATA TRANSFORMATION, MATRIX, EXPERIMENT</p>
<p>PRIDE XML Format MODEL/FORMAT</p> <p>Systems: 1</p>	<p>RIF-CS MODEL/FORMAT</p> <p>Systems: 1</p>	<p>Schema.org TERMINOLOGY ARTIFACT</p> <p>Systems: 1</p>	<p>Semanticscience Integrated Ontology TERMINOLOGY ARTIFACT</p> <p>Systems: 1</p>

Data Citation Implementation Pilot




The Future of Research Communications and e-Scholarship

ABOUT
COMMUNITY
GROUPS
RESOURCES
NEWS + EVENTS
CONFERENCES
PUBLICATIONS
MEDIA
DONATE

GROUP MENU

Group Home
Members
Expert Groups
Links & Files
Google Forum
Calendar
Boston Workshop

GROUP LEADER


Tim Clark

FORCE11 » Groups » Data Citation Implementation Pilot (DCIP)

DATA CITATION IMPLEMENTATION PILOT (DCIP)

DESCRIPTION

The FORCE11 has been awarded supplemental funding as part of the [NIH BD2K bioCADDIE](#) to extend the work of the Data Citation Implementation Group by organizing a Data Citation Pilot (DCIP).

Members of this FORCE11 community have been participating in NIH meetings and contributed substantial materials to the bioCADDIE white paper outlining the v Discovery Index produced by bioCADDIE. Concrete plans were formulated by members to conduct a pilot project on data citation with international partners based on the data citation workshop and the [joint Elixir-BD2K workshop](#) held in January. At the significant support was expressed for testing the proposed implementation of the [Joint Data Citation Principles \(JDDCP\)](#), developed by the FORCE11 Data Citation Implementation joint Elixir-BD2K workshop recommended a data citation pilot project as one of two meeting.

This program is run by an Executive Committee (see below) and funded by the NIH.

SUB GROUPS

[EG1 DCIP: FAQs](#)

[EG2 DCIP: Identifiers](#)

[EG3 DCIP: Publisher Early Adopters](#)

[EG4 DCIP: Repository Early Adopters](#)

[EG5 DCIP: JATS](#)

[EG6 DCIP: Landing Pages](#)


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

4/5/16

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego

8

bioCADDIE Prototype


 biomedical and healthCare Data Discovery Index Ecosystem

☒ Search for data set
 ☐ Search for repository

Search Examples: (Breast Cancer, Genetic Analysis Software, Gene EGFR, Lung[title] AND Cancer, Cancer AND (Lung[Title] OR Skin[Title]))

Repositories

- ☐ ClinicalTrials (11975)
- ☐ SRA (3745)
- ☐ BioProject (2275)
- ☐ ArrayExpress (2005)
- ☐ GEO (1774)
- ☐ GEMMA (77)
- ☐ PDB (60)
- ☐ Dataverse (32)
- ☐ Proteomexchange (32)
- ☐ Dryad (30)

More...

Data Types

- ☐ Clinical Trials (11975)
- ☐ Gene Expression (3875)
- ☐ Nucleotide Sequence (3745)
- ☐ Unspecified (2337)
- ☐ Protein Structure (60)
- ☐ Proteomics Data (32)
- ☐ Phenotype (14)

Feedback?

If you are having problems using our tools, or if you would just like to send us some feedback, please post your questions on [GitHub](#).

Displaying 10 of 22038 results for "Breast Cancer" Sorted By: Relevance

First ◀ 1 2 3 4 5 6 7 8 9 10 ▶ Last

☐ **Structural consequences of a cancer-causing BRCA1-BRCT missense mutation** [PDB](#)

ID: 1N50

Description: Breast cancer type 1 susceptibility protein

☐ **The Estrogen Receptor Alpha Ligand Binding Domain D538G Mutant in Complex with 4-hydroxytamoxifen** [PDB](#)

ID: 4Q50

Description: Estrogen receptor

☐ **Crystal Structure of the BRCT Domains of Human BRCA1 in Complex with a Phosphorylated Peptide from Human Acetyl-CoA Carboxylase 1** [PDB](#)

ID: 3COJ

Description: Breast cancer type 1 susceptibility protein, Acetyl-CoA carboxylase 1

☐ **Breast Cancer Data** [Dryad](#)

DateIssued: 01-25-2013

Description: Breast cancer data. This R-workspace contains the objects: dat, dat.st, event, st, and time.

☐ **Washington Post Poll: Breast Cancer** [Dataverse](#)

Description: This survey covers the following issues and topics: Concerns about breast cancer (9); cancer as a national health problem (2).


Publication: Washington Post

Release Date: 11-23-2009

☐ **CRYSTAL STRUCTURE OF THE BRMS1 N-TERMINAL REGION** [PDB](#)

ID: 2XUS

Description: BREAST CANCER METASTASIS-SUPPRESSOR 1


 biomedical and healthCare Data Discovery Index Ecosystem

[About Us](#)
[Feedback](#)
[Login](#)


Engaging The Community Toward a Data Discovery Index (v0.5)

☒ Search for data set
 ☐ Search for repository


Search Examples: (Breast Cancer, Genetic Analysis Software, Gene EGFR, Lung[title] AND Cancer, Cancer AND (Lung[Title] OR Skin[Title]))

[Advanced Search](#) [help](#)


Statistics




23 REPOSITORIES



10 DATA TYPES

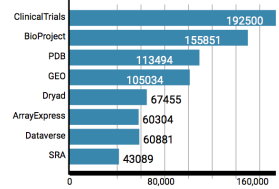


838,137 DATASETS



4 PILOT PROJECTS

Repositories




New Features

March 1, 2016


- Create a user account
- Save/Send search results
- Boolean/Advanced search
- Query expansion
- Link data sets to grant
- Display results from multiple repositories
- Integrate new repositories
- Improve faceted search
- Search for repository
- Autocomplete

Pilot Projects




GWAS Finder

Search literatures for "Genome-Wide Association Studies".




ISEE-DELVE

Search Visualization project for Big Data.



DataRank

Find most suitable datasets for you.



Data Citation Discovery

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

bioCADDIE is supported by the National Institutes of Health through the Big Data to Knowledge, Grant 1U24AI117966-01. NIH | UCSD | DBMI

Related Search Coming Soon!

- Lung cancer
- Breast cancer
- Prostate cancer
- Colorectal cancer

<https://biocaddie.org/sign-access-datamed-biocaddie-prototype>

bioCADDIE Acknowledgements

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

