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

THE FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier, Ush Brand Jan Aalbersberg, Gabrielle Appleton, Myles Antin, Ante Baak, Nicolas Blomberg, Jan-Willem Botstein, Luiz Bernardino da Silva Santos, Philip E Bourne, Jilaa Bouwmeester, Anthony J Brookes, Tim Clarke, Meere Crossas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T Evelo, Richard Ebbers, Alejandro Gonzalez-Bolívar, Alasdair J G Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Haen van der Heurn, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Maitone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultze, Thierry Sengstag, Ted Slater, George Strawn, Monis A. Swartz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andrea Waagmeester, Peter Wittenburg, Katherine Wilencroft, Jun Zhao, and Barend Mons

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

4/5/16

Supported by the NIH grant 1U24 AI117966-01 to the University of California, San Diego
bioCADDIE

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...

- ArrayExpress
- NCBI
- NDAR
- PDB
- Gemma
- GEO
- Gene Expression Omnibus
- BioProject
- NIH LINC Program
- CardioVascular Research Grid
- THE CELL
- DRYAD
- ICPSR
- Sequence Read Archive
- NeuroMorpho.Org
- ClinicalTrials.gov
- Mouse Phenome Database
- EuPathDB
- GeneNetwork
- NURSA
- dbGaP
- The Dataverse Project

Supported by the NIH grant 1U24 AI117966-01 to the University of California, San Diego
bioCADDIE Prototype Architecture

Repositories

Data Sources

Ingestion

Metadata Ingestion

ElasticSearch

Metadata Management:
- Mapping
- Indexing

Searching

User Interface

- Web pages
- Search engine

Terminology server

Advanced search strategies:
- Query expansion
- Ranking

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

Supported by the NIH grant #1-U24-AI117966 to the University of California, San Diego
WG3: Metadata Specifications

_Metadata specification v1_, future-proofed for progressive extensions, to support intended capability of the DDI prototype

**PHASE 1 OUTPUT:**
  - 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.

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**.
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.

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

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