DataMed:
Finding useful data across multiple biomedical data repositories

L Ohno-Machado[1,2], SA Sansone[3*], G Alter[4*], I Fore[5*], J Grethe[1*], H Xu[6*], Alejandra Gonzalez-Beltran[3], Philippe Rocca-Serra[3], Ergin Soysal[6], Nansu Zong[1], H Kim[1]*
on behalf of the NIH BD2K bioCADDIE Consortium

[1]University of California San Diego, [2]Veterans Administration San Diego Healthcare System, [3]University of Oxford, [4]University of Michigan Ann Arbor, [5]National Institutes of Health, [6]The University of Texas Health Science Center at Houston
[*]These authors contributed equally

Abstract
The value of broadening searches for data across multiple repositories has been identified by the biomedical research community. As part of the NIH Big Data to Knowledge initiative, we work with an international community of researchers, service providers and knowledge experts to develop and test a data index and search engine, which are based on metadata extracted from various datasets in a range of repositories. DataMed is designed to be, for data, what PubMed has been for the scientific literature. DataMed supports Findability and Accessibility of datasets. These characteristics - along with Interoperability and Reusability - compose the four FAIR principles to facilitate knowledge discovery in today's big data-intensive science landscape.

Biomedical research has always been a data intensive endeavor, but the amount of information was much more manageable just a decade ago. Today's researchers not only have to stay abreast of the latest publications in their fields, but they also increasingly need to use existing data to help generate or test hypotheses *in-silico*, compare their data against reference or benchmark data, and contribute their own data to various "commons" to help health sciences move faster and be more easily reproducible[1,2].

Data, software, and systems (e.g., analytical pipelines) are essential components of the ecosystem of contemporary biomedical and behavioral research. There are numerous databases focused, for example, on different communities, types of data, or types of research. While these may be indexed and searchable, these indexes usually do not interconnect, making it difficult to search for data across different research communities. Enabling broader focused searches for biomedical data was a key recommendation to the Director of the NIH by the Data and Informatics Working Group[3]. The work described here was funded to provide NIH with practical experience in fulfilling that recommendation.

Starting the data discovery journey

The biomedical and healthCAre Data Discovery Index Ecosystem (bioCADDIE)[4], is a Data Discovery Index Consortium funded by the NIH Big Data to Knowledge (BD2K) program[5]. The goal is to help users find data from sources which they would be unlikely to otherwise encounter, as PubMed does with the medical literature[6]. For example, biomedical researchers and clinicians don't know the names of all journals that may have articles of interest. Even if they knew the journal names, it would be time consuming to search each resource separately. While searches within a certain journal web site could potentially be more detailed than those allowed in PubMed, they are not as useful if users do not get to those sites. PubMed, among other things, allows users to find articles in unfamiliar journals, and it makes sure that these journals meet certain quality criteria. Similarly, the bioCADDIE consortium is developed the search engine DataMed to help researchers find data of interest in a broad spectrum of high quality repositories.

A first prototype of DataMed is available at https://datamed.org. It performs searches on a shallow generic index that includes an initial set of 23 repositories and over 600k data sets, covering 10 data types. DataMed stores metadata generic enough to describe any dataset using a model we have called the DAta Tag Suite (DATS) (see working group 3(1) in Supplementary Box 1).

It is important to highlight some important differences between DataMed and broad harvesting of web-based data sets. DataMed is not the equivalent of Google Scholar[7] or Microsoft Academic for data[8].

a. bioCADDIE has developed criteria for data repository inclusion in DataMed (akin to journal inclusion in PubMed[9]) based on standards, interoperability, sustainability, overall quality, and user demand.
b. The DATS model, which is akin to the Journal Article Tag Suite (JATS) used in PubMed (NISO JATS Draft Version 1.1d3 April, 2015 http://jats.nlm.nih.gov/archiving/tag-library/1.1d3)[10] enables submission of data for "ingestion" by DataMed (Figure 1). DATS has "core" metadata requirements that every data repository is expected to supply.
c. In addition to free text queries, DataMed provides a means to compose a structured query so that metadata specific to the life sciences can be better utilized.

DATS also facilitates the implementation of biomedical data discovery on other platforms in addition to DataMed. Building on prior attempts the use of structured data markup for web content is now reaching some maturity particularly in schema.org[11]. The bioCADDIE consortium worked with the scientific community on use cases for biomedical data search.

DATS represents a platform independent model of the metadata necessary to support those use cases. The DATS model has been designed to be easily translated to and from *schema.org*[12] (See working group 3(2) Supplementary Box 1). The mapping between DATS and schema.org and its biomedical extensions is expected to be seamless, with the potential benefit that data producers can expose their data assets to any web-based search tool. This opens the way for the emergence of new kinds of data-related web applications limited only by the ingenuity of their developers.

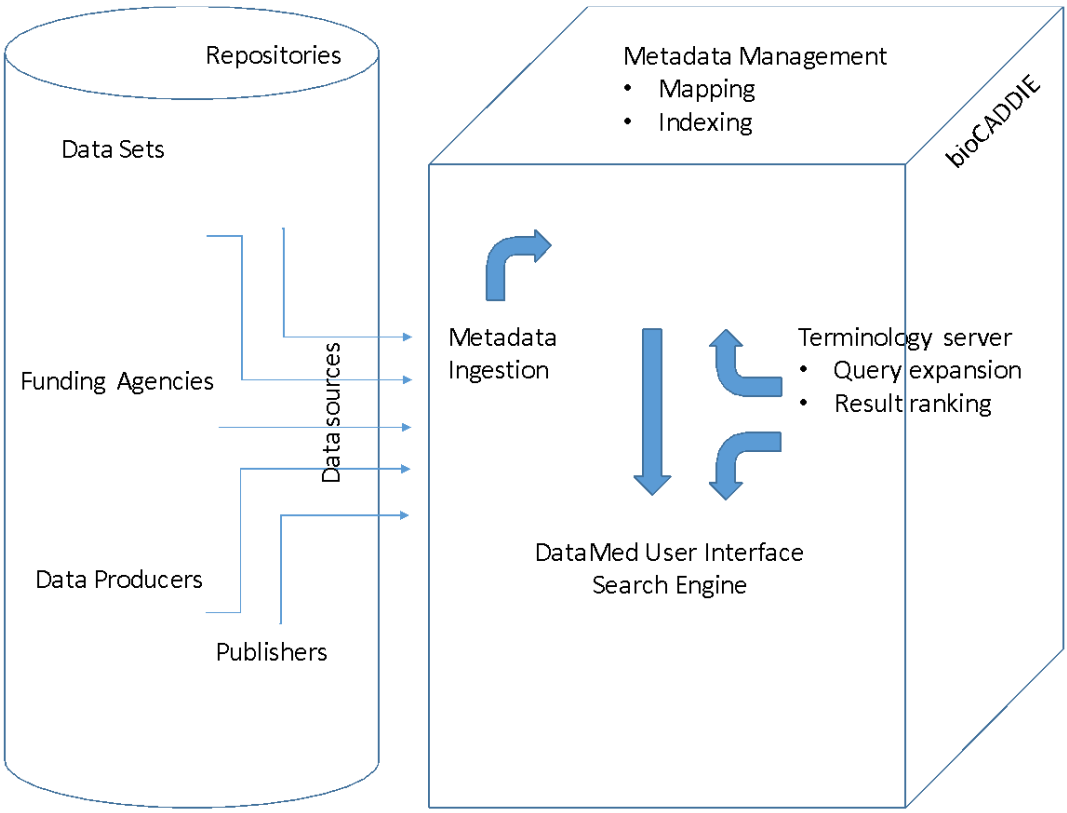

Figure 1. *Data sources have various metadata specifications, which undergo "ingestion" into the common DATS model, whose metadata elements are used for indexing and DataMed searches. A terminology server is used to expand, transform, and standardize concepts used in metadata descriptions and in user queries.*

Getting from point A to point Z: What is in between?

The specialized repositories that serve the needs of their specific communities, with their high level of specialization and more detailed metadata, are important components of the data discovery ecosystem. Although fine-grained metadata that are specific to certain areas

are not yet indexed in DataMed, this search engine helps users find important data assets they are not aware of according to generic factors. We believe the amount of work to make the specialized data indexable by DataMed is very low - the better the quality of metadata in a repository, the easier it is to produce DATS metadata.

Looking more broadly at the ecosystem for data discoverability, a large community is participating in various aspects of bioCADDIE (Figure 2). Although the DataMed search engine interface is what users will interact with, we engaged community input on various components that operate behind the scenes to ensure that searches return the desired outputs. We formed Working Groups (WGs) and invited the research community to help us scope the project, select repositories, define indexing processes, develop a search engine and evaluate its results. The multiple WGs have included to date over 86 members from 56 institutions in the USA and the EU (see list of bioCADDIE collaborators and their affiliations). We expect to further broaden national and international participation as we get feedback from the community, develop new WGs, and continue the work on existing ones.

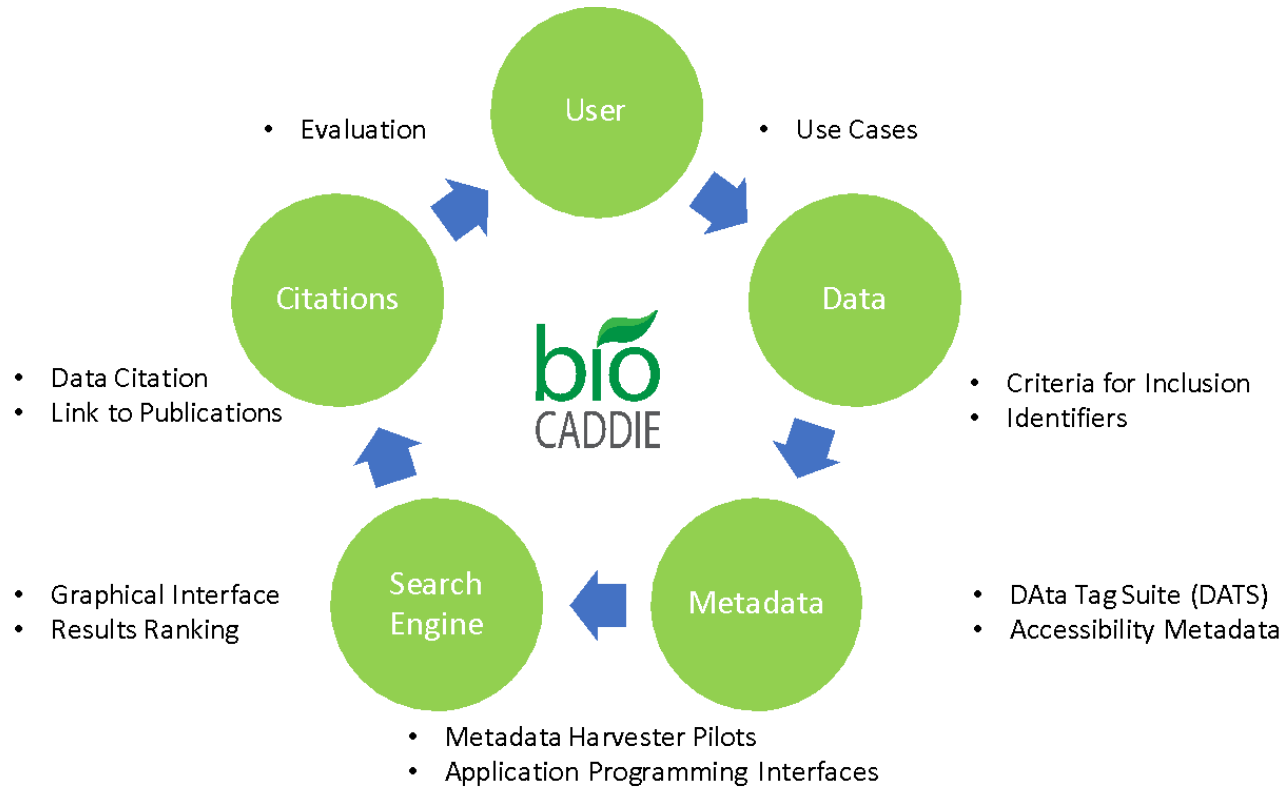

Figure 2. *Community input to the Data Discovery Index Consortium. Working groups involved over 80 people from multiple institutions to scope the project via use cases, develop core metadata specifications, recommend identifier strategies, develop and test the search engine prototype, and discuss issues in data citation. Additionally, bioCADDIE funded*

external pilot projects for development of software that will be incorporated into the DataMed prototype.

Organization

bioCADDIE's activities leading to DataMed can be grouped into the following general areas:

1. **Scope and Use Cases**

Use cases helped define the appropriate boundaries and level of granularity for DataMed: which queries will be answered in full, which ones only partially, and which ones are out of scope. A workshop at the start of the project assembled researchers to discuss use cases. The current DataMed prototype is intended to point to data of interest by indicating the repository in which it is found, and providing some minimal information about the data so users can elect to follow the links to those repositories or move on to the next entry. Most use cases derived from community input (see working group 4 in Supplementary Box 1) are currently focused on finding data for a particular diagnosis and/or health condition (e.g., asthma) or data modality (e.g., fMRI).

2. **Criteria for repository inclusion**: standards, interoperability, sustainability

We established an initial set of criteria for a repository to be indexed by bioCADDIE (Supplementary Box 2). These criteria were inspired by those used by the National Library of Medicine in considering the indexing of a journal by PubMed[9]. An important consideration was to select data repositories that would help us evaluate search results (see working group 6 in Supplementary Box 1). We prioritized highly utilized repositories, because it would be hard to evaluate recall and precision if all repositories were unknown to users. Several new repositories are being added to the initial set currently available in DataMed.

3. **The DATS** model

The Metadata WG (see working group 3(1) in Supplementary Box 1) is a joint activity with the BD2K centers of excellence Metadata WG, and closely connected to other NIH initiatives (e.g., the BD2K Center for Data Annotation and Retrieval) and ELIXIR activities in Europe[13]. This group produced the DATS model, and its serialization describes the metadata needed for datasets to populate DataMed. DATS has a core and extended set of elements, to progressively accommodate more specialized data types. Like the JATS, the core DATS elements are generic and applicable to any type of datasets. The extended DATS includes an initial set of elements, some of which are specific for life, environmental and biomedical science domains and can be further extended as needed. The general concept fits with the idea of annotation profiles which allow a continuum from generic to specific annotation[14]. As with NISO-JATS the DATS model resides at the generic core of such a progression.

DATS has been designed to cover both (i) experimental datasets, which do not change after deposit in a repository, and (ii) datasets in reference knowledge bases describing dynamic concepts, such as "genes", whose definition morphs over time. The core and the extended DATS entities are made available as machine readable JSON schemata, which will be annotated with schema.org elements.

The DATS model was developed given the following considerations:

- A variety of data discovery initiatives exists or are being developed; although they have different scope, use cases and approaches, the analysis of their metadata schemas is valuable. Several meta-models for representing metadata were reviewed (Supplementary Box 3) to determine essential items.
- Identification of the initial set of metadata elements was based on: (i) analyses of use cases (a *top-down* approach); and (ii) mapping of existing metadata schemas (a *bottom-up* approach). From the use cases, a set of 'competency questions' were derived; these defined the questions that we want DataMed to answer. The questions were abstracted, key concepts were highlighted, color-coded and binned in entities, attributes and values categories, to be easily matched with the results of the *bottom-up* approach.

The metadata schemas and models used in the mapping have been described in the BioSharing Collection for bioCADDIE[15], which will be enriched progressively. Provided information includes: creators and maintainers; documentation, including URL where this is located; when the data set became available; version; source of metadata elements (e.g. XSD), including the URL where the model or schema was sourced.

Accessibility metadata

One of the strongest messages in the use cases was the need for information about data access[16]. Many types of biomedical data are restricted to protect confidential information about human subjects. Researchers want to know which data sets are readily available on the Internet and which ones require prior approval and other security clearances. We incorporated three dimensions in the core metadata (see working group 7 in Supplementary Box 1): authorization (Supplementary Note 1), authentication, and access type. Researchers also want to know which data can be accessed directly by machines through an application programming interface (API).

4. **Identifiers and the data ingestion pipeline**

Identifiers

We developed recommendations for the appropriate handling of identifiers for datasets and data repositories within DataMed (see working group 2 in Supplementary Box 1). At the most basic level, all data sets must be uniquely identified and web-resolvable. We will not mint identifiers for datasets, rather we will rely on the identifiers provided by the source. If the source does not have the capability to mint identifiers, we can suggest options for obtaining identifiers (e.g. issuing a DOI via Datacite[17]). In order to reuse identifiers provided by the data repository, a key set of features for a data repository supplied identifier are required. A identifier must be, (1) stable, (2) persistent for the life of the data repository, (3) unique within the data repository, and (4) resolvable (Supplementary Note 2) – i.e., a landing page must exist at the data repository that can accessed via embedding the identifier in a stable URL structure.

Indexing pipeline

The backend indexing pipeline for DataMed consists of a scalable architecture for ingesting, processing and indexing data. Metadata related to datasets are extracted and further enhanced by a document processing pipeline utilizing ApacheMQ[18] and MongoDB[19]. Finalized metadata for a dataset is exported to an ElasticSearch endpoint that is then used by DataMed and can be used by external developers via its native RESTful API[20] (Supplementary Note 3).

Operationally, to bring in a new data resource, a curator initially configures how the ingestion pipeline retrieves the metadata from a source (e.g. rsync[21], OAI-PMH[22], RESTful web service[23]) and provides the appropriate parameters. The ingestion pipeline then samples a number of records from the source for the curator so that an appropriate mapping can be made to the core metadata model described above. If the source has adopted a metadata standard that already exists within the pipeline (e.g. PDB XML[24], GEO MINiML[25], DATS (see working group 3(1) in Supplementary Box 1)) that mapping can be applied to the new source. However, if the source utilizes a different standard or a native API, a new mapping must be developed. After the initial mapping process a number of enhancement modules may be run to insert additional metadata into the dataset description document. Modules are being developed to enhance the pipeline, including an enhancer for semantic annotation (e.g. synonyms and super-classes)[26] and one for citation altmetrics (i.e. the number of times a dataset has been cited)[27]. When a document has completed all processing steps it can be exported to a number of target endpoints via export modules. Currently, metadata records are exported to ElasticSearch whose native APIs are utilized by DataMed's user interface.

5. **Search engine prototype and usability testing**

The Core technology Development Team (CDT)[28] is developing the functioning prototype of DataMed with input from members of the community. The CDT has designed a modular architecture for the prototype backbone into which various projects can be integrated, creating a repository ingestion and indexing pipeline that maps to specifications provided by various working groups in the bioCADDIE community. The search engine interface provides all the functionality expected from a modern search application such as grouping the returning result via facets and saving/downloading search results. During search process, query input from user is processed to identify requested entities, and expanded with synonyms retrieved from a terminology server to enhance search results. This helps to prevent missing of datasets due to wording and terminological differences. Another important role for this it to host pilot projects, and integrate them into search interface seamlessly.

The CDT acquired user-specific input through the preliminary evaluation of DataMed by focus groups and by a limited number of end users (see working group 9 in Supplementary Box 1). Continuous evaluation will iteratively inform DataMed development. We designed and implemented the DataMed web application to provide a user-friendly interface that enables users to browse, search, obtain ranked results (see working group 8 in Supplementary Box 1) and in the near future get recommendations of related datasets tailored to their specific interests, preferences, and needs. A user-rating strategy is being discussed for future versions so that the application can learn from these ratings as well as from analyses of users behavior. The community may report issues on any aspect of DataMed through GitHub[29] which provides a mechanism where they can be discussed and followed up in an open manner.

Current status and next steps

The bioCADDIE consortium is continuing to engage stakeholders in the development and evaluation of metadata and tools. The quality of indexing is being optimized, as is the DataMed prototype search engine. Although highly utilized data repositories were the first to be included, there continues to be a need for indexing the "long tail" of science (i.e., smaller data sets that are being produced daily all over the world and may not be in repositories). Many new challenges will emerge, since it is unclear which data are going to be highly valued and how much assistance data producers will need in exposing their metadata for discovery. (Remember that PubMed relies on publishers, not individual authors.) Where to strike the right balance between quality and cost of maintaining the index from the viewpoints of data producers, data disseminators, and data consumers is still an open question. Nevertheless, we have to start somewhere, and exposing a resource early

in its development (work on DataMed started in April 2015) helps us obtain valuable feedback from the user community to direct our next steps.

bioCADDIE and its products (core metadata specifications, criteria for repository inclusion, the DataMed search engine prototype) are intended to promote engagement and discussion around concepts that will last far beyond a particular grant or program. We invite everyone to join us in this journey to propel health sciences into a future where data are widely shared and easily discoverable and where discoveries relevant to human health are greatly accelerated.

Acknowledgements

This project is funded by grant U24AI117966 from NIAID, NIH as part of the BD2K program.

The co-authors, which are the lead investigators and chairs/co-chairs of the core activities, thank all contributors to the bioCADDIE Consortium and list them here in alphabetical order within each activity group (each name appears only once even though many people participated in different activities)

Philip E. Bourne - NIH Associate Director for Data Science

Steering Committee

Alison Yao – NIAID, NIH
Dawei Lin – NIAID, NIH
Dianne Babski - National Library of Medicine, NIH
George Komatsoulis - National Library of Medicine, NIH
Heidi Sofia – NHGRI, NIH
Jennie Larkin – ADDS Office, NIH
Ron Margolis – NIDDK, NIH

Metadata Working Group (WG)

Allen Dearry – NIEHS, NIH
Carole Goble - The University of Manchester
Helen Berman - Rutgers
Jared Lyle - ICPSR, University of Michigan
John Westbrook - Rutgers
Kevin Read - NYU School of Medicine
Marc Twagirumukiza - W3C ScheMed WG
Marcelline Harris - University of Michigan
Mary Vardigan - ICPSR, University of Michigan
Matthew Brush - Oregon Health and Science University
Melissa Haendel - Monarch Initiative
Michael Braxenthaler - Roche
Michael Huerta - National Library of Medicine, NIH
Morris Swertz - CORBEL, BBMRI, ELXIR-NL and University Medical Center Groningen
Rai Winslow - John Hopkins University
Ram Gouripeddi - University of Utah
Shraddha Thakkar - NCTR FDA
Weida Tong - NCTR FDA

Accessibility Metadata WG
Alex Kanous - University of Michigan
Anne-Marie Tassé - McGill University
Damon Davis - HealthData.gov
Frank Manion - University of Michigan
Jessica Scott - GlaxoSmithKline
Kendall Roark - Purdue University
Mark Phillips - McGill University
Reagan Moore - University of North Carolina

Identifiers WG
Jo McEntyre - EMBL-EBI, ELIXIR EBI Node, Pub Med Central
Joan Starr - California Digital Library, DataCite
John Kunze - California Digital Library
Julie McMurry - Monarch Initiative
Merce Crosas - Data Science
Michel Dumontier - Center for Expanded Data Annotation and Retrieval, W3C HCLSIG
Stian Soiland-Reyes - University of Manchester
Tim Clark - Harvard Medical School, FORCE11 Data Citation Implementation Group

Use Cases and Testing Benchmarks WG
Dave Kaufman - Arizona State University
Dina Demner-Fushman - National Library of Medicine, NIH
Ratna Rajesh Thangudu - BD2K Standards Coordinating Center
Steven Kleinstein - Yale School of Medicine
Thomas Radman - NIH
Todd Johnson - UTH
Trevor Cohen - UTH
Zhiyong Lu - National Library of Medicine, NIH
March 2015 workshop use case contributors

Ranking Search Results WG
Chelsea Ju - University of California Los Angeles
Christina Kendziorski - University of Wisconsin Madison
Chun-Nan Hsu - UCSD
Elmer Bernstam - UTH
Gregory Gundersen - Icahn School of Medicine at Mount Sinai
Griffin Weber - Harvard Medical School
Hongfang Liu - Mayo Clinic
Jim Zheng - UTH

Sanda Harabagiu - University of Texas at Dallas
Vincent Kyi - University of California Los Angeles

Criteria for Repository Inclusion WG
Amy Pienta - University of Michigan, ICPSR
Elizabeth Bell – UCSD
John Marcotte - ICPSR / University of Michigan
John Yates - The Scripps Research Institute
Kei Cheung - Yale University
Larry Clarke - National Cancer Institute (NCI)
Marek Grabowski - University of Virginia
Matthew McAuliffe - Center for Information Technology NIH
Neil McKenna - Baylor College of Medicine
Tanya Barrett - NCBI (GEO, BioSample, BioProject), GA4GH
Tim Clark - Harvard Medical School, FORCE11 Data Citation Implementation Group

Dataset Citation Metrics WG
Daniel Mietchen - National Library of Medicine, NIH
Jennifer Lin - PLOS
Kristi Holmes - Northwestern University
Martin Fenner - DataCite
Michael Taylor - Elsevier
Rebecca Lawrence - F1000
Sunje Dallmeier-Tiessen - CERN
Trisha Cruse - DataONE, CDL

Core technology Development Team (CDT)
Anupama Gururaj - UTH
Burak Ozyurt - UCSD
Claudiu Farcas - UCSD
Cui Tao - UTH
Deevakar Rogith - UTH
Ergin Soysal - UTH
Larry Lui - UCSD
Mandana “Nina” Salimi - UTH
Min Jiang - UTH
Muhammad Amith - UTH
Nansu Zong - UCSD
Ngan T Nguyen-Le - UTH
Pratik Kumar Chaudhary – UTH

Ruiling Liu - UTH
Saeid Pournejati - UTH
Vidya Narayana - UTH
Xiao Dong - UTH
Xiaoling Chen - UTH
Yaoyun Zhang - UTH
Yueling Li - UCSD

Pilot project PIs and collaborators

Aditya Menon – National ICT Australia
Cathy Wu – University of Delaware
Cecilia Arighi – University of Delaware
Chris Mungall – Lawrence Berkeley National Laboratory
Dmitriy Dligach – Boston Children's Hospital
Guergana Savova – Harvard University
Guoqian Jiang – Mayo Clinic
Harold Solbrig – Mayo Clinic
Hwanjo Yu – POSTECH, South Korea
Jaideep Vaidya – Rutgers
Jeeyae Choi – University of Wisconsin Milwaukee
Jina Huh – UCSD
Julio Facelli – University of Utah
Peter Rose – UCSD
Ricky Taira – University of California Los Angeles
Xiaoqian Jiang – UCSD
Zhaohui Qin – Emory University

Supplements to bioCADDIE

OmicsDI
Eric Deutsch, ISB
Henning Hermjakob EMBL-EBI
Peipei Ping, UCLA
CountEverything
David Haussler, UC Santa Cruz
Ida Sim, UC San Francisco
Isaac Kohane, Harvard Medical School

FORCE11
Tim Clark, Harvard Medical School

Maryann Martone, UC San Diego

Administrative Supplements

Bert O'Malley, Baylor College of Medicine

Carolyn Mattingly, North Carolina State University Raleigh

George Hripcsak, Columbia University

Hongfang Lu, Mayo Clinic

Raymond Winslow, Johns Hopkins

Steven Kleinstein, Yale

Tor Wagner, University of Colorado

Trevor Cohen, UTH

References

1 Wilkinson, M. D. *et al. Scientific data* **3** (2016).
2 Collins, F. S. & Tabak, L. A. *Nature* **505**, 612 (2014).
3 Group, D. a. I. W. (National Institutes of Health 2012).
4 Ohno-machado, L. *et al. figshare. http://dx.doi.org/10.6084/m9.figshare.1362572 Retrieved 8/8/2016.*
5 Bourne, P. E. *et al. Journal of the American Medical Informatics Association* **22**, 1114-1114 (2015).
6 Lu, Z. *Database : the journal of biological databases and curation* **2011**, baq036, doi:10.1093/database/baq036 (2011).
7 Noruzi, A. *Libri* **55**, 170-180 (2005).
8 Hands, A. *Technical Services Quarterly* **29**, 251-252 (2012).
9 Kejariwal, D. & Mahawar, K. K. (2012).
10 Huh, S. *Science Editing* **1**, 99-104 (2014).
11 Guha, R. V., Brickley, D. & MacBeth, S. *Queue* **13**, 10-37, doi:10.1145/2857274.2857276 (2015).
12 Schema.org Consortium. <http://schema.org/> (2016).
13 ELIXIR Consortium. <https://www.elixir-europe.org> (2016).
14 Goble, C., Stevens, R., Hull, D., Wolstencroft, K. & Lopez, R. *Brief Bioinform* **9**, 506-517, doi:10.1093/bib/bbn034 (2008).
15 McQuilton, P. *et al. Database : the journal of biological databases and curation* **2016**, doi:10.1093/database/baw075 (2016).
16 Wang, R. Y. & Strong, D. M. *Journal of management information systems* **12**, 5-33 (1996).
17 Brase, J. in *Cooperation and Promotion of Information Resources in Science and Technology, 2009. COINFO'09. Fourth International Conference on.* 257-261 (IEEE).
18 Apache Foundation. <http://activemq.apache.org/> (2016).
19 Chodorow, K. (" O'Reilly Media, Inc.", 2013).
20 Kuć, R. & Rogozinski, M. (Packt Publishing Ltd, 2016).
21 *rsync*, <https://rsync.samba.org/> (2016).
22 Coll, I. S. & Cruz, J. M. B. *El profesional de la información* **12**, 99-106 (2003).
23 Richardson, L. & Ruby, S. (" O'Reilly Media, Inc.", 2008).
24 Westbrook, J., Ito, N., Nakamura, H., Henrick, K. & Berman, H. M. *Bioinformatics* **21**, 988-992 (2005).
25 National Center for Biotechnology Information. <http://www.ncbi.nlm.nih.gov/pubmed/> (2016).
26 Kiryakov, A., Popov, B., Terziev, I., Manov, D. & Ognyanoff, D. *Web Semantics: Science, Services and Agents on the World Wide Web* **2** (2011).
27 Haustein, S., Peters, I., Sugimoto, C. R., Thelwall, M. & Larivière, V. *Journal of the Association for Information Science and Technology* **65**, 656-669 (2014).
28 bioCADDIE. <https://biocaddie.org/core-development-team> (2016).
29 Core Development Team. <https://github.com/biocaddie/prototype_issues> (2016).