HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2019 April 12.

Published in final edited form as:

Nat Genet. 2017 May 26; 49(6): 816–819. doi:10.1038/ng.3864.

Finding useful data across multiple biomedical data repositories using DataMed

Lucila Ohno-Machado[1,2,8], **Susanna-Assunta Sansone**[3,8], **George Alter**[4,8], **Ian Fore**[5,8], **Jeffrey Grethe**[6,8], **Hua Xu**[7,8], **Alejandra Gonzalez-Beltran**[3], **Philippe Rocca-Serra**[3], **Anupama E Gururaj**[7], **Elizabeth Bell**[1], **Ergin Soysal**[7], **Nansu Zong**[1], and **Hyeon-eui Kim**[1,8]

[1]Health System Department of Biomedical Informatics, University of California, San Diego, La Jolla, California, USA.

[2]Veterans Administration San Diego Healthcare System, San Diego, California, USA.

[3]e-Research Centre, University of Oxford, Oxford, UK.

[4]Department of History and Inter-University Consortium for Political and Social Research (ICPSR), Institute for Social Research, University of Michigan, Ann Arbor, Michigan, USA.

[5]US National Institutes of Health, Bethesda, Maryland, USA.

[6]Department of Neurosciences, University of California, San Diego, La Jolla, California, USA.

[7]School of Biomedical Informatics,University of Texas Health Science Center at Houston, Houston, Texas, USA.

[8]These authors contributed equally to this work.

Abstract

The value of broadening searches for data across multiple repositories has been identified by the biomedical research community. As part of the US National Institutes of Health (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 data sets in a range of repositories. DataMed is designed to be, for data, what PubMed has been for the scientific literature. DataMed supports the findability and accessibility of data sets. These characteristics—along with interoperability and reusability—compose the four FAIR principles to facilitate knowledge discovery in today's big data–intensive science landscape.

Correspondence should be addressed to L.O.-M. (machado@ucsd.edu).

AUTHOR CONTRIBUTIONS

L.O.-M., G.A., S.S., I.F., J.G., H.X., E.B. and H.K. supervised the research. L.O.-M., G.A., S.S., I.F., J.G., H.X. and H.K. wrote the manuscript. A.G.-B., P.R.-S., E.S., N.Z. and A.E.G. performed the experiments. A.G.-B., P.R.-S. and N.Z. analyzed the data. A.G.-B., P.R.-S., E.S., N.Z., A.E.G., E.B. and H.K. contributed analysis tools.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

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 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 (for example, 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 databases may be indexed and searchable, the indexes usually do not interconnect, making it difficult to search for data across different research communities. Enabling more broadly focused searches for biomedical data was a key recommendation to the Director of the NIH, according to the report of the Data and Informatics Working Group (see URLs). The work described here was funded to provide the NIH with practical experience in fulfilling that recommendation.

Starting the data discovery journey

The biomedical and healthCAre Data Discovery Index Ecosystem (bioCADDIE) (see URLs), is a Data Discovery Index Consortium funded by the NIH Big Data to Knowledge (BD2K) program[3]. The goal is to help users find data from sources that they would be unlikely to encounter otherwise, as PubMed does with the medical literature[4]. For example, biomedical researchers and clinicians do not 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 website 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 developing the search engine DataMed to help researchers find data of interest in a broad spectrum of high-quality repositories.

A first prototype of DataMed (see URLs) performs searches on a shallow generic index that, as of today (10 May 2017), includes an initial set of 66 repositories and over 1.3 million data sets, covering 15 data types. DataMed stores metadata generic enough to describe any data set using a model we have called the DAtaset Tag Suite (DATS)[5].

URLs.

Data and Informatics Working Group Report to The Advisory Committee to the Director, https://acd.od.nih.gov/Data%20and%20Informatics%20Working%20 Group%20Report.pdf; bioCADDIE white paper—Data Discovery Index, http://dx.doi.org/10.6084/m9.figshare.1362572; prototype of DataMed, https://datamed.org/; NISO JATS Draft Version 1.1d3, http://jats.nlm.nih.gov/archiving/tag-library/1.1d3; use cases and testing benchmarks of bioCADDIE, https://biocaddie.org/group/working-group/working-group-4-use-cases-and-testing-benchmarks; ELIXIR, https://www.elixir-europe.org/; BD2K Aztec, https://aztec.bio; BioSharing, https://biosharing.org/; Schema.org, http://schema.org/; DataCite, https://www.datacite.org/; ICPSR, https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html; ApacheMQ, http://activemq.apache.org/; rsync, https://rsync.samba.org/; GEO MINiML, https://www.ncbi.nlm.nih.gov/geo/info/MINiML.html; bioCADDIE core development team, https://biocaddie.org/core-development-team; bioCADDIE GitHub repository, https://github.com/biocaddie/WG3-MetadataSpecifications; bioCADDIE participate, https://biocaddie.org/participate.

It is important to highlight some significant differences between DataMed and broad harvesting of web-based data sets. DataMed is not the equivalent of Google Scholar[6] or Microsoft Academic[7] for data.

1. bioCADDIE has developed criteria for data repository inclusion in DataMed (akin to the criteria for journal inclusion in PubMed[8]) based on standards, interoperability, sustainability, overall quality and user demand.
2. The DATS model, which is a model inspired by the Journal Article Tag Suite (JATS) used in PubMed (NISO JATS Draft Version 1.1d3, April 2015 (see URLs))[9], enables submission of data for 'ingestion' by DataMed (Fig. 1). DATS has 'core' metadata requirements that every data repository is expected to supply. The National Library of Medicine is exploring DATS and its possible role in ongoing efforts to make a broader range of biomedical data more readily discoverable.
3. 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 used.

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 of which they are not aware according to generic factors. The amount of work to make the specialized data indexable by DataMed is relatively low—the better the quality of the 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 (Fig. 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 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 working groups have thus far included over 86 members from 56 institutions in the United States and the European Union (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 working groups and continue the work on existing ones.

Organization

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

Scope and use cases.

Use cases helped define the appropriate boundaries and level of granularity for DataMed to determine which queries will be answered in full, which ones will only partially be answered 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 locating the repository in which they are found and providing some minimal information about the data so that users can elect to follow the links to those data sets or move on to the next entry. Most use cases derived from community input (see Working Group 4 in the Supplementary Note) are currently focused on finding data for a particular diagnosis and/or health condition (for example, asthma) or a particular data modality (for example, fMRI). All of the use cases are available via the bioCADDIE website under Working Group 4 “Use Cases and Testing Benchmarks” (see URLs).

Criteria for repository inclusion: standards, interoperability and sustainability.

We established an initial set of criteria for a repository to be indexed by bioCADDIE (Supplementary Note). These criteria were inspired by those used by the National Library of Medicine in considering the indexing of a journal by PubMed[8]. An important consideration was to select data repositories that would help us evaluate search results (see Working Group 6 in the Supplementary Note). We prioritized highly used 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.

The DATS model.

The Descriptive Metadata Working Group is closely connected to other NIH initiatives (for example, the BD2K Center for Data Annotation and Retrieval) and ELIXIR activities in Europe (see URLs). The Descriptive Metadata Working Group, along with Accessibility Metadata Working Group 7, produced the DATS model and its serialization that describes the metadata needed for data sets to populate DataMed. Developed iteratively, the DATS model is the result of three complementary approaches: review of existing metamodels, analyses of use cases and mapping of existing metadata schemas to find convergences and common metadata elements.

DATS has been designed around the Dataset element, to ensure the discoverability of both experimental data sets and data sets in reference knowledge bases. The Dataset element is linked to other digital objects, which are the focus of other indexed resources—specifically, Publication (for example, PubMed), Software (for example, BD2K Aztec (see URLs)), DataRepository and DataStandard (for example, BioSharing (see URLs))—implementing the concept that DataMed is built to be part of an interlinked ecosystem of resources. Key information about the Dataset element is about its accessibility, which is represented by the Access metadata element that encompasses information on authorization, authentication and access type. This is important because researchers typically want to know which data sets are readily available on the Internet and which ones require prior approval and other security clearances, as well as which data can be accessed directly by machines through an application programming interface (API).

As in JATS, the core DATS elements are generic and applicable to any type of data set. 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. DATS is a platform-independent model also available as a Schema.org (see URLs) annotated JSON-LD serialization. DATS is being implemented by DataCite (see URLs) and other data discovery indices, such as the Inter-university Consortium for Political and Social Research (ICPSR) (see URLs) and the NIH BD2K OmicsDI[10]. A description of the development process, the application of use cases and the adoption progress is detailed in a complementary DATS article[5]; the specifications, serializations and examples are freely available from the bioCADDIE GitHub repository.

Identifiers and the data ingestion pipeline.

We developed recommendations for the appropriate handling of identifiers for data sets and data repositories within DataMed (see Working Group 2 in the Supplementary Note). At the most basic level, all data sets must be uniquely identified and web resolvable. We do not mint identifiers for data sets; rather, we 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 (for example, issuing a DOI via DataCite[11]). To reuse identifiers provided by the data repository, a key set of features for a data repository–supplied identifier are required. An identifier must be (i) stable, (ii) persistent for the life of the data repository, (iii) unique within the data repository and (iv) resolvable (Supplementary Note)—that is, a landing page must exist at the data repository that can be accessed via embedding the identifier in a stable URL structure

The back-end indexing pipeline for DataMed consists of a scalable architecture for ingesting, processing and indexing data. Metadata related to data sets are extracted and further enhanced by a document-processing pipeline using ApacheMQ (see URLs) and MongoDB[12]. Finalized metadata for a data set are exported to an ElasticSearch endpoint that is then used by DataMed and can be used by external developers via its native RESTful API[13] (Supplementary Note).

Operationally, to bring in a new data resource, a curator initially configures how the ingestion pipeline retrieves the metadata from the source (for example, via rsync (see URLs), OAI-PMH[14] or the RESTful web service[15]) 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 (for example, PDB XML[16], GEO MINiML (see URLs) or DATS (see Working Group 3(1) in the Supplementary Note)), 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 selected to insert additional metadata into the data set description document. Modules are being developed to enhance the pipeline, including an enhancer for semantic annotation (to add, for example, synonyms and superclasses)[17] and one for citation altmetrics (to give the number of times a data set has been cited)[18]. 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 used by DataMed's user interface.

Search engine prototype and usability testing.

The Core technology Development Team (CDT) (see URLs) 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 of the returning results via facets and saving or downloading of search results. During the search process, the query input from the user is processed to identify requested entities and expanded with synonyms retrieved from a terminology server to enhance the search results. This helps prevent non-retrieval of data sets due to wording and terminological differences. Another important role of DataMed is to host pilot projects and integrate them into the 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 the Supplementary Note). 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 the Supplementary Note) and, in the near future, get recommendations of related data sets 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 has been reporting issues on any aspect of DataMed through GitHub (see URLs), which provides a mechanism whereby these issues can be discussed and followed up in an open manner.

To illustrate a simple preliminary of the DataMed prototype and existing search engines, we used an instantiated query from a use case collected from biomedical researchers. The complete, natural language query was to search for "all data for the HTT gene related to Huntington's disease across all databases," from which the relevant keyword query was automatically extracted in DataMed as "HTT gene Huntington's disease." We searched both the natural language query and the keyword query in DataMed V1.5, OmicsDI, Google and Bing. For each system, the top 50 returned results were extracted and manually reviewed by a domain expert to determine relevancy on a categorical scale—relevant, partially relevant and not relevant—as well as the number of data sets. Owing to the large number of data sets retrieved by Google and Bing, only the first 50 were considered in this analysis.

Supplementary Table 1 shows the results returned from each system, as searched on 2 November 2016. As expected, most of the returned results from Google and Bing were related to publications and general web pages and not to data sets. In this specific case, only Google returned three data sets when the natural language query was used for the search. As OmicsDI was developed with similar goals as DataMed, but with a focus on omics data sets, we also included it in the comparison. When the natural language query was used, OmicsDI

did not return any results. For the keyword query, OmicsDI returned 192 results, which is a higher number than that for DataMed (94 results). However, when we evaluated the precision of the two systems at different cutoffs (10, 20 and 50 results), we noticed that DataMed had higher precision for retrieving both relevant and partially relevant results. The precisions for the first 10, 20 and 50 results were 100%, 100% and 100% for DataMed and 40%, 45% and 64% for OmicsDI, respectively.

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 used data repositories were the first to be included, there continues to be a need for indexing the 'long tail' of science (smaller data sets that are being produced daily all over the world and may not be in repositories). Many new challenges will emerge, as 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 relatively 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 and 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.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

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

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

Reference:

1. Wilkinson MD et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018 (2016). [PubMed: 26978244]
2. Collins FS & Tabak LA Policy: NIH plans to enhance reproducibility. Nature 505, 612–613 (2014). [PubMed: 24482835]

3. Bourne PE et al. The NIH Big Data to Knowledge (BD2K) initiative. J. Am. Med. Inform. Assoc 22, 1114 (2015). [PubMed: 26555016]
4. Lu Z PubMed and beyond: a survey of web tools for searching biomedical literature. Database (Oxford) 2011, baq036 (2011).
5. Sansone S-A et al. DATS: the data tag suite to enable discoverability of datasets. Sci. Data 4, 170059 (2017). [PubMed: 28585923]
6. Noruzi A Google Scholar: the new generation of citation indexes. Libri 55, 170–180 (2005).
7. Hands A Microsoft Academic Search—http://academic.research.microsoft.com. Tech. Serv. Q 29, 251–252 (2012).
8. Kejariwal D & Mahawar KK Is your journal indexed in PubMed? Relevance of PubMed in biomedical scientific literature today. WebmedCentral MISCELLANEOUS 3, WMC003159 (2012).
9. Huh S Journal Article Tag Suite 1.0: National Information Standards Organization standard of journal extensible markup language. Sci. Ed 1, 99–104 (2014).
10. Perez-Riverol Y et al. Nat. Biotechnol 35, 406–409 (2017) [PubMed: 28486464]
11. Brase J. IEEE. in 2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology (COINFO 2009); 2009. 257–261.
12. Chodorow K MongoDB: The Definitive Guide (O'Reilly Media, 2013).
13. Kuć R & Rogozinski M ElasticSearch Server (Packt Publishing, 2016).
14. Coll IS & Cruz JMB Open archives initiative. Protocol for metadata harvesting (OAI-PMH): descripción, funciones y aplicaciones de un protocolo. Prof. Inf 12, 99–106 (2003).
15. Richardson L & Ruby S RESTful Web Services (O'Reilly Media, 2008).
16. Westbrook J, Ito N, Nakamura H, Henrick K & Berman HM PDBML: the representation of archival macromolecular structure data in XML. Bioinformatics 21, 988–992 (2005). [PubMed: 15509603]
17. Kiryakov A, Popov B, Terziev I Manov D & Ognyanoff D. Semantic annotation, indexing, and retrieval. Web Semantics 2, 49–79 (2004).
18. Haustein S, Peters I, Sugimoto CR, Thelwall M & Larivière V Tweeting biomedicine: an analysis of tweets and citations in the biomedical literature. J. Assoc. Inf. Sci. Technol 65, 656–669 (2014).

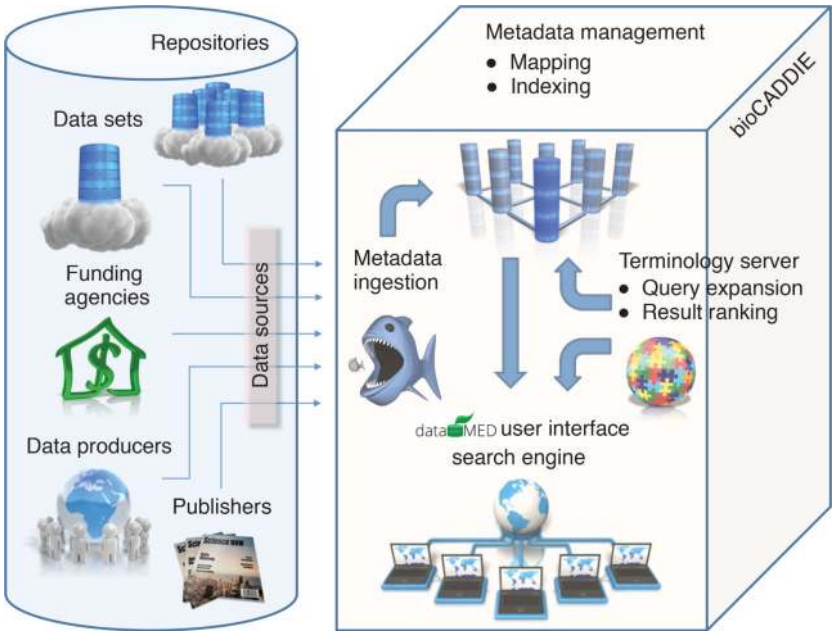

Figure 1.

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. The user is only responsible for submitting a query in natural language to the DataMed search engine, such as “astrocytoma and IDH1.” (The figure uses illustrations from PresenterMedia.com.)

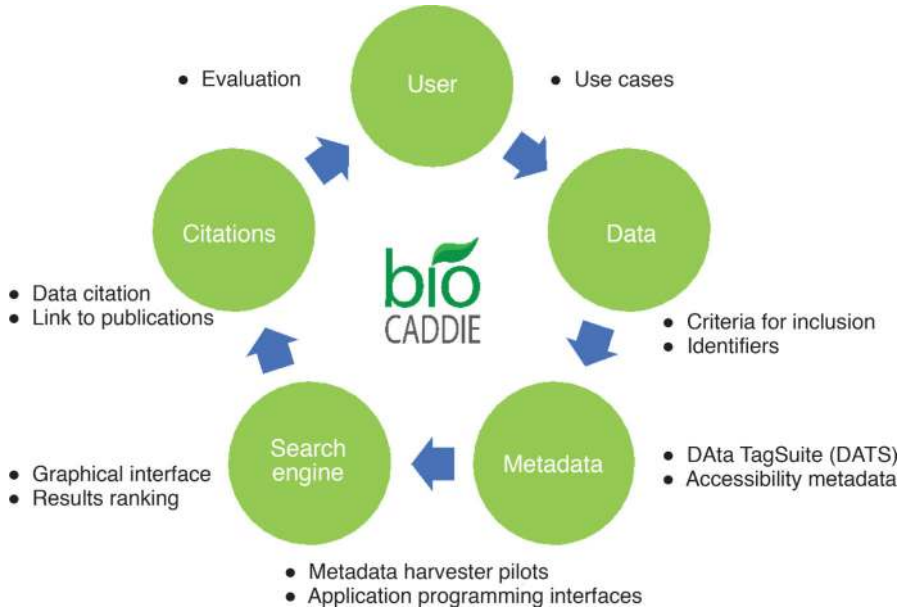

Figure 2.

Community input to the Data Discovery Index Consortium. Working groups involved over 86 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.