

Building a search engine for finding biomedical datasets across repositories – the DataMed system

Xiaoling Chen¹, Ruiling Liu¹, Burak Ozyurt², Anupama E.Gururaj¹, Ergin Soysal¹, Trevor Cohen¹, Firat Tiriyaki¹, Yueling Li², Nansu Zong², Min Jiang¹, Deevakar Rogith¹, Mandana Salimi¹, Hyeon-eui Kim², Philippe Rocca-Serra³, Alejandra Gonzalez-Beltran³, Claudiu Farcas², Todd Johnson¹, Ron Margolis⁵, George Alter⁴, Susanna-Assunta Sansone³, Ian M. Fore⁵, Lucila Ohno-Machado², Jeffrey S. Grethe^{2,*}, Hua Xu^{1,*}

¹The University of Texas Health Science Center at Houston, Houston, TX, USA

²University of California San Diego, La Jolla, CA, USA

³University of Oxford, Oxford, UK

⁴University of Michigan, Ann Arbor, MI, USA

⁵National Institutes of Health, USA

*Corresponding authors:

Jeffrey S. Grethe, PhD

University of California, San Diego, Center for Research in Biological Systems

9500 Gilman Drive #0608, La Jolla, CA, USA, 92093

Phone: 858-822-0703

Email: jgrethe@ncmir.ucsd.edu

Hua Xu, PhD

The University of Texas School of Biomedical Informatics at Houston

7000 Fannin St., Suite 870 Houston, Texas, USA, 77030

Phone: 713-500-3924

E-mail: hua.xu@uth.tmc.edu

Keywords: Data Discovery Index, Metadata, Dataset, Information Storage and Retrieval, Information Dissemination

Word count: 3798

Abstract

Objective: Finding relevant datasets is an important step for promoting data reuse in the biomedical domain; but it is very challenging given the volume and complexity of biomedical data. DataMed (<https://datamed.org>) is a biomedical data discovery index search engine developed by the NIH-funded biomedical and healthCARE Data Discovery Index Ecosystem (bioCADDIE) consortium.

Materials and Methods: DataMed enables users to efficiently find existing heterogeneous datasets that are distributed across a wide range of biomedical repositories. We describe the technical details of developing DataMed, including the data ingestion and indexing processes, as well as the search engine functionality and implementation.

Results & Conclusion: As of May 2017, DataMed has released its version 2.0 and ingested 66 repositories comprising of 1,375,977 datasets across 15 datatypes. Within a year, DataMed has attracted 9452 users from 121 countries and areas, and 15 organizations have requested DataMed to index their repositories.

INTRODUCTION

With the advances of recent high-throughput technologies, the bottleneck of biomedical research has shifted from digital data generation to data management and analysis. Large, complex and diverse data are continually generated and have accumulated exponentially, becoming valuable sources for biomedical discovery. Recently, the value of data sharing and the ease of data reuse have amplified tremendously. To take full advantage of existing data, facilitate knowledge discovery and make scientific discoveries more productive and reproducible, it is critical to follow the widely-endorsed FAIR (Findable, Accessibility, Interoperability and Reusability) data principles[1] for existing research data. However, there are unprecedented challenges in collecting and normalizing pre-existing data from disparate sources for different purposes.

The biomedical and healthCare Data Discovery Index Ecosystem (bioCADDIE) project.[2], funded by the National Institutes of Health (NIH) via the Big Data to Knowledge (BD2K) program, is focused on the discovery of biomedical datasets. Since its start, researchers, service providers and knowledge experts around the globe have participated in various aspects of bioCADDIE, such as working groups, pilot projects and dataset retrieval challenges (K. Robert, 2017, T. Cohen, 2017). To instantiate the concepts and recommendations developed by this large community, bioCADDIE developed a data discovery index prototype named DataMed. The ultimate goal of DataMed is to be an one-stop shop for users to find biomedical datasets of interest from heterogeneous sources[3]. It is similar to PubMed, a one-stop resource for the biomedical literature. This article describes technical details about developing DataMed including its metadata ingestion and indexing pipeline and search engine functionalities. For other aspects of the bioCADDIE work, please refer to complementary articles.[3 4],(K. Robert, 2017, T. Cohen, 2017).

BACKGROUND

To embrace the Big Data era, the biomedical research community has devoted substantial effort and resources to the goal of enabling biomedical research as a digital enterprise. To promote the reuse of existing datasets and facilitate the finding of datasets of interest easily by researchers, major initiatives have been devoted to building repositories and knowledge-bases for specific types of data and domains.[5] For example, GEO (Gene Expression Omnibus) is a public functional genomics data repository for gene expression data.[6] PDB (Protein Data Bank) serves as an information portal for biological macromolecular structures.[7] ImmPort is a data repository for public data sharing of immunological studies.[8]

Such data repositories have greatly improved the discoverability and reuse of datasets in the biomedical domain, since researchers can easily find a particular dataset from a familiar repository. However, searching capabilities for different types of data across multiple repositories is needed as the number of repositories in biomedical research continues to grow. To achieve this, researchers currently need to search individual repositories one by one, which is not only time consuming, but could also limit the ideas that researchers have when they know what types of datasets they might have access to. An integrated biomedical data retrieval and discovery system across different repositories is a first step towards removing this limitation. A successful integrated search engine will provide a one-stop shop, in which data seekers can quickly access all these resources, thus improving the community's capability to utilize existing databases for data query, knowledge dissemination and integrative analyses.

Some data discovery index systems have been built to help users find datasets across multiple repositories. For example, the Omics Discovery Index (OmicsDI), aggregates and indexes

“Omics” datasets comprising of Transcriptomics, Genomics, Proteomics and Metabolomics data resources including 84,937 datasets from 15 repositories.[9] Other resources provide some foundations for data discovery. For example, Datacite is an organization that provides persistent identifiers (DOIs) for general research data with the goal of helping researchers locate, identify and cite research data. As of today, Datacite includes 8,086,054 works and 1,237 data centers globally from many different domains, including biomedical, engineering, etc.[10] The Neuroscience Information Framework[11] and NIDDK Information Network[12] are community aggregators of biomedical data and information focused on specific biomedical domains. These community discovery sites assist researchers in finding data and information (e.g. information on organisms, reagents, etc.). The underlying framework for these projects aggregates information from more than 230 resources and supports initiatives such as the Resource Identification Initiative[13] which supports reproducible research by providing the ability to uniquely identify research resources in the literature. Dataverse is an open source web application for researchers, data authors, publishers, etc. to share, preserve, cite, explore, and analyze research data. Dataverse includes 48,828 datasets.[14] However, in the biomedical domain, so far there is no comprehensive search engine that can cover a broad spectrum of high quality repositories and provide efficient search functions to the users. Several technical challenges exist when building an integrated search engine across different biomedical repository resources, including extracting and normalizing metadata from different repositories to a unified metadata model, as well as finding highly relevant datasets for users in a huge search space.

The mission of DataMed is to provide a centralized data discovery index to help users efficiently find and access existing datasets that are distributed across a wide range of repositories in the biomedical domain. To address the challenges discussed above, we developed a metadata ingestion pipeline which extracts, maps and indexes the metadata of datasets from different

repositories by following the Data Tag Suite(DATS), a unified metadata model for biomedical datasets developed by the bioCADDIE consortium[4], based on input from the community and a thorough analysis of existing metadata from popular repositories. We implemented a fully functional search engine for retrieving relevant datasets, with a user-friendly interface and other advanced technologies such as Elasticsearch,[15] natural language processing (NLP), and terminology services. Currently, the DataMed (available at www.datamed.org) prototype has ingested 1,375,977 datasets across 15 different data types from 66 repositories and has been accessed by over 9000 users worldwide.

METHODS

DataMed consists of two major components: the ingestion and indexing pipeline and the search engine (Figure 1). The ingestion and indexing pipeline collects metadata from different repositories, maps them to the DATS model, and then indexes them to Elasticsearch Endpoint. The search engine is a web-based application that consists of a user interface and various search functionalities, including the core Elasticsearch-based ranking algorithm, the query expansion module utilizing NLP and ontology, and other advanced services such as a dataset similarity calculator.

Ingestion and Indexing Pipeline

System Architecture

For data ingestion, transformation and enhancement, DataMed uses a horizontally-scalable message oriented Extract-Transform-Load (ETL) system. As shown in Figure 2, the pipeline is a loosely-coupled distributed system consisting of a dispatcher component and one or more data processing components called consumer heads, together with a command line management interface. The dispatcher component acts as an event hub, orchestrating a configured data ingestion and processing pipeline through the use of persistent queues. The consumer head

component manages the lifecycle of processing components called consumers. The consumer head has a plugin architecture for extensibility and for separating data specific processing functionality from the generic life cycle management and service functionality. Each consumer is a stateless component that receives a data record wrapper document from the consumer head, does an operation such as transformation, cleanup and/or enhancement on the document and returns the document back to the consumer head. Depending on the processing status returned from the consumer, the consumer head first saves the updated document to the underlying data store (MongoDB) and places a message in the message queue of the dispatcher. The dispatcher receives the message and using the configured pipeline specification decides what is the next step is in the ingestion and indexing pipeline in order to place a message in the corresponding consumer's input message queue. All running consumer heads listen to the message queues configured for the consumers they are managing. Whenever a new message is received in any of the input message queues, the corresponding consumer is provided with the data record wrapper that it needs to process in an asynchronous manner.

Data ingestion

Scientific data is provided by multiple institutions and laboratories in many different ways and formats, making the ingestion process a challenge. To handle heterogeneities in the data distributions, the pipeline abstracts out retrieval modes (e.g. REST API, FTP), data formats (e.g. XML, CSV, JSON) and data traversal functionality. Ingestion is considered the first step of the processing pipeline. Different ingestors for retrieval mode and data format combinations are developed as specific consumers using a specialized plugin interface. Data traversing is generalized using the iterator design pattern. The iterators are designed to allow streaming as much as possible to retrieve data only when needed allowing the system to process data sets much larger than the system memory. Each specific ingestor uses one or more iterators internally to retrieve, parse and break the data into records and traverse the records. Each

traversed raw data record is converted to JSON format by the ingestor, wrapped in a JSON document with additional information used for the pipeline management and provenance and stored in a MongoDB database to be further processed by the pipeline.

Data transformation

For indexing, ingested raw data needs to be transformed into the DATS format. The pipeline has a domain specific language called JSNTL to transform one JSON format to another JSON format. The language uses JSONPath similar to XPath to specify a branch(es) in a JSON object tree. A matching branch from the source JSON document is mapped to a branch in the destination document. This mapping can be one-to-one, many-to-one, one-to-many or many-to-many. To allow all forms of mapping and arbitrary field value manipulation and combination, the language allows embedding scripts written in the Python language to be included in the transformation rules. The language also allows conditional transformations. The transformation engine is integrated into a consumer and run as a part of the processing pipeline by the consumer head(s).

Data enhancement

Once the original metadata has been retrieved from a source, additional enhancements to the metadata records may need to be performed. For example, performing natural language processing (NLP) on a metadata record, particularly on longer text descriptions and abstracts, can provide a detailed list of semantic concepts contained within the dataset description. The current DataMed ingestion and indexing pipeline includes an ingestion consumer followed by the transformation enhancer, data citation enhancer (which attaches information on citations of the dataset from other resources) and an NLP-based biomedical named entity enhancer. After each data source is run through this pipeline, the enhanced DATS transformation for each record stored in the MongoDB is indexed to the Elasticsearch cluster.

Search Engine

The DataMed search engine is a PHP-based web application following the MVC (Model, View, and Control) architecture. The user interface provides various ways for users to interactively refine their search queries and navigate among returned results. In addition to the default Elasticsearch ranking algorithm, different search functionalities that are based on advanced NLP and ontology technologies are implemented to improve the search performance and user experience.

User Interface (UI)

The development of the UI is an iterative process that involved users' feedback during each cycle. DataMed provides a google-like search box for users to enter queries. In addition to the basic search mode, we also provide the advanced search option that allows expert users to define the fields for searching as well as build specialized queries using Boolean operators. Results are a ranked list of relevant biomedical datasets and facets by different categories (e.g., data types, repository names), which can help users further filter and refine the results. Once a user clicks a specific dataset record, it will display general information about the dataset, as defined in DATS, such as title, description, released data, ID, taxonomic information, etc. It also provides a link to the original data resources for users to access the dataset.

In addition to the general view that applies to all data types across all repositories, we also provide a detailed view for a single data repository that provides additional repository-specific information. For example, if a user selects to limit datasets to PDB only, they can switch to the detailed view specific to PDB, which contains PDB-specific facets such as genes and keywords, which can help user to find the datasets of interest easily.

The DataMed UI also provides many other functions to meet user needs during their search. For example, it allows users to share selected dataset information with others via email, download them or manage them as collections using their DataMed accounts. Furthermore, we provide additional information related to biomedical datasets such as similar datasets in DataMed via similar dataset service, and publications and grants, by linking external resources such as PubMed[16] and NIH RePORT.[17]

Search functionalities

Figure 3 shows the workflow of the DataMed search engine. After a user enters a query, it is processed by the NLP service to extract biomedical concepts. Then these concepts are sent to the terminology service to generate synonyms. The search query is then expanded by combining the original query and the expanded synonyms and sent to Elasticsearch to retrieve ranked results and facets. The user can further refine their search results by interacting with the UI. We describe each of the services in the following sections.

NLP service: We implemented two different NLP solutions to identify biomedical concepts from queries: (1) extraction of general MeSH (Medical Subject Headings) concepts using the existing MetaMapLite system;[18] and (2) identification of specific types of biomedical concepts such as diseases, chemicals, genes, biological processes and cell lines using locally developed customized NLP programs. Both rule-based and machine learning-based NLP pipelines were developed for diverse types of biomedical entities using the CLAMP NLP toolkit (clamp.uth.edu). After biomedical entities are identified, we further map them to the UMLS Concept Unique Identifiers (CUIs).[19-21] Our NLP service is also implemented in two ways, a web service, which is used for real time query expansion, and a java program, which is used as a NLP enhancer in the ingestion pipeline.

Terminology service: The domain knowledge and use of biomedical concepts with their relationships play important roles in retrieving the most relevant results from biomedical datasets for advanced functionality. A terminology server based on SciGraph (<https://github.com/SciGraph/SciGraph>) and Neo4j was implemented, adopting major ontologies such as Mesh,[22] SNOMED CT,[23] Gene Ontology,[24] Foundational Model of Anatomy,[25] NCBI Taxonomy[26] and Hugo Gene Nomenclature.[27] These different terminologies are integrated in the context of the UMLS Metathesaurus to obtain a unified ontology of related terms. A web service was then implemented to support real-time concept and relationship (e.g., synonym and parent-child) identification and it is used in conjunction with the NLP service for query expansion and metadata enrichment and as a stand-alone component for spelling correction and auto-completion.

Elasticsearch-based ranking algorithm: After query expansion, an Elasticsearch query is constructed and searched against all the metadata fields in the index. Currently, DataMed 2.0 uses Elasticsearch's default ranking algorithm (cosine similarity based on vector space model using TF-IDF weighting[28]) to retrieve and rank datasets from the entire collection. We are in the process of integrating novel search algorithms from the bioCADDIE dataset retrieval challenge (K. Robert, 2017) and this should be reflected in future releases.

Similar Dataset: A similar dataset service was developed by adopting an iterative variant of the Random Indexing paradigm,[29 30] in which dataset vectors are composed from word vectors using methods of distributional semantics[31] such that words that occur in similar abstracts will have similar vectors. However, before adding them to the dataset vector, these vectors undergo transformations that indicate the field in which they occur such that datasets with similar (but not necessarily identical) words in the same fields will have similar vectors. In addition, an approach to encoding continuous values into semantic vector representations[32 ,33] was used to encode

numerical fields such as date – such that vectors for datasets published at a similar (but not identical) time will be related to one another.[34]

RESULTS

As of May, 2017, DataMed has ingested 1,375,977 datasets from 66 repositories across 15 datatypes. Fifteen organizations have submitted requests to DataMed to index their repositories. We are in the process of ingesting more repositories. Regular updates to the index is planned to include recently added datasets and changes of the metadata of datasets (in case of any modifications) in a timely manner. Table 1 summarizes the ingestors developed and used for DataMed.

Table 1. Ingestors used for DataMed.

Ingestor Type	Sample DataMed Sources	# of DataMed sources using this ingestor
Web Ingestor	Clinical Trials, Uniprot	42
Database Ingestor	NeuroMorpho, PeptideAtlas, Clinical Trials Network	14
OAI Ingestor	Dryad, CVRG	2
Two-stage Web Ingestor	ICPSR, Dataverse Native	2
Rsync ingestor	PDB, dbGAP	2
FTP Ingestor	BioProject, Biological Magnetic Resonance Data	2

	Bank (BMRB)	
Aspera Ingestor	GEO Datasets	1
CSV Ingestor	Gemma	1
XML Ingestor	ArrayExpress	1

Figure 4 shows the screenshot of the search page in DataMed. In the left column, facets are provided for users to refine search results. The middle column displays the search results. Visualization of dataset release dates via a timeline graph, synonyms and search query details are shown on the right. Figure 5 shows the screenshots of the information page for a selected dataset. Besides the metadata of the dataset, similar datasets, related publication and grant information are also provided if available.

To improve DataMed and monitor usage, a user-activity tracking module is implemented that logs all queries as well as keystrokes and clicks. To further engage users and solicit user input during the development process, DataMed collects feedback (<https://datamed.org/feedback.php>) via multiple modes:

- 1) a “Contact Us” form that results in an email to the development team and automatic reporting of an issue on GitHub (<https://datamed.org/about.php>)
- 2) a System Usability Scale (SUS)-style questionnaire that allows users to indicate the extent of their satisfaction with DataMed and suggest new features (<https://datamed.org/questionnaire.php>), and

3) an issue reporting repository in GitHub for reporting of all bugs and issues

(https://github.com/biocaddie/prototype_issues/issues). Feedback from all routes is logged into GitHub which serves as a central node to track user-reported issues for the development team. To better understand users' needs, usability studies were conducted, providing guidance for the iterative development of DataMed (Dixit R, 2017). We also provide a submission form for users to suggest potential repositories for indexing in DataMed.

To evaluate the NLP service, we generated a corpus of 700 randomly selected datasets from 21 repositories that are included in DataMed. The datasets were annotated by domain experts with the concepts of interest. Our NLP service achieved average precision, recall, and F-measures of 91.06%, 71.34% and 79.66% respectively, on this evaluation dataset. To evaluate the query expansion scheme, we extracted the most frequent search terms in DataMed and manually examined the synonyms we obtained from the terminology service. Of the total 156 search terms examined, 124 of them returned the correct synonyms. We also ran the overall ranking algorithm against the benchmark dataset generated for the bioCADDIE dataset retrieval challenge (K. Robert, 2017, T. Cohen, 2017) and obtained reasonable results, with an infNDCG of 0.2948 and a precision at 10 (P@10) of 0.46 for 15 comprehensive natural language queries. The inferred normalized discounted cumulative gain (infNDCG) is a ranking measure that accounts for the fact that not all of the results actually have a manually assigned grade and uses sampling techniques to estimate the true NDCG.

DataMed was first launched on June 30, 2016. Since then we tracked the traffic to DataMed using Google Analytics. Within a year, and still in a prototype stage, DataMed has attracted 9452 users from 121 countries. The top five countries that users are coming from are United States, United Kingdom, India, Canada and Germany. The average session time spent by a user on DataMed is about 4 minutes and the average number of pages that users visit each

session is 4.29, indicating that these are potentially legitimate users. 37.3% of users are returning visitors and visit DataMed more than once.

DISCUSSION

Increased availability of digital data and growing multi-domain research areas in the biomedical field has created a need for users from differing disciplines to find and retrieve datasets that are not from their direct area of expertise. Therefore, toolsets that provide a coherent presentation of metadata information across repositories housing biomedical datasets are important for data search and access. DataMed is one of the first data discovery indexes that harvest metadata from a broad range of data providers and makes it available through a single integrated search system.

DataMed is a completely open-source prototype search system that uses a modular architecture to capture, manage and present metadata of heterogeneous types of biomedical data. DataMed supports use cases from the general, biological and translational communities with a number of common needs for metadata searches. This makes DataMed broad (rather than deep) in its searching capabilities, allowing it to easily span a range of diverse domains and types of data. There are other centralized repositories such as OmicDI[9] developed by the biomedical community that address the needs that are specific to one or few of the various data types of biomedicine. DataMed ingests such aggregators (i.e., other indices) in addition to the repositories that house the datasets themselves in an effort to provide a comprehensive overview of all the available information for a particular dataset. However, such data aggregators have not yet been developed for the majority of the biomedical areas, so future efforts should focus on developing deeper indices for each of the specialized domains.

The DataMed team is working towards exposing the harvested metadata to other general search applications (e.g. Google, Yahoo, Bing etc.) by providing schema.org (schema.org) markups. Additionally, we are developing a RESTful API to provide programmatic access to all of DataMeds' harvested metadata and additional user-interface features. Such sharing capabilities make DataMed easily accessible not only to the biomedical community but also to those outside of the biomedical sphere.

An important and useful approach to reuse of existing tools is exemplified by our approach to allow seamless integration with external tools for data visualization. We are developing a proof-of-concept module to link DataMed to external resources. An example may help clarify what this means: DataMed will use a plugin to access remotely located database and tools from the Library of Integrated Network-based Signatures (LINCS) Data Coordination and Integration Center (<http://lincs-dcic.org/#/>) as a visualization tool for the results of user requested data. Reusing components like these not only provides additional data usability options, but also significantly reduces duplication of programming effort and development costs.

The consortium approach to development has allowed DataMed to receive inputs from various community members to produce a reusable, modular, portable, robust, feature-rich application. The DataMed team is continuously improving the quality of the DDI and expanding the scope of ingested repositories.

CONCLUSIONS

DataMed leverages scalable technologies to ingest, index, and search diverse biomedical datasets across repositories. It demonstrates a successful prototype for building an integrated dataset search engine for the biomedical domain. Its flexible service oriented architecture and the open source nature make it valuable for building other data discovery indexes in the biomedical domain.

Acknowledgements

This project is funded by grant U24AI117966 from NIAID, NIH as part of the BD2K program. We thank all members of the bioCADDIE community for their valuable input in the overall project.

Author contributions

L. Ohno-Machado, I. Fore, J. Grethe, H. Xu, H. Kim, S. Sansone, R. Margolis, G. Alter and T. Johnson supervised the research. X. Chen, B. Ozyurt, E. Soysal, A. Gururaj, T. Cohen, L. Ohno-Machado, I. Fore, J. Grethe and H. Xu, wrote the manuscript. X. Chen, B.Ozyurt, E.Soyas, T.Cohen, R. Liu, F.Tiryaki, Y. Li, N.Zong, A.Gururaj, P.Pocca-Serra, A. Gonzalez-Beltran, D. Rogith, M.Jiang, C.Farcas performed the research and analyzed the data.

Competing interest

The authors declare no competing financial interests.

Figure Legends

Figure 1. Architecture of DataMed

Figure 2. The workflow for the ingestion and indexing pipelines

Figure 3. Workflow of the DataMed web application

Figure 4. Screenshot of search page in DataMed

Figure 5. Screenshot of a single item page in DataMed

References

1. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 2016;3:160018 doi: 10.1038/sdata.2016.18[published Online First: Epub Date]].
2. Lucila O-m, George A, Ian F, et al. *bioCADDIE white paper - Data Discovery Index*, 2015.
3. Ohno-Machado L, Sansone S-A, Alter G, et al. Finding useful data across multiple biomedical data repositories using DataMed. *Nat Genet* 2017;49(6):816-19 doi:

- 10.1038/ng.3864 [http://www.nature.com/ng/journal/v49/n6/abs/ng.3864.html - supplementary-information\[published Online First: Epub Date\]\]](http://www.nature.com/ng/journal/v49/n6/abs/ng.3864.html - supplementary-information[published Online First: Epub Date]]).
4. Sansone S-A, Gonzalez-Beltran A, Rocca-Serra P, et al. DATS: the data tag suite to enable discoverability of datasets. bioRxiv 2017 doi: 10.1101/103143[published Online First: Epub Date]].
5. NIH Data sharing repositories. Secondary NIH Data sharing repositories. https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html
6. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic acids research 2002;30(1):207-10
7. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic acids research 2000;28(1):235-42
8. Bhattacharya S, Andorf S, Gomes L, et al. ImmPort: disseminating data to the public for the future of immunology. Immunologic research 2014;58(2-3):234-9 doi: 10.1007/s12026-014-8516-1[published Online First: Epub Date]].
9. Perez-Riverol Y, Bai M, Leprevost F, et al. Omics Discovery Index - Discovering and Linking Public Omics Datasets. bioRxiv 2016 doi: 10.1101/049205[published Online First: Epub Date]].
10. DataCite - A Global Registration Agency for Research Data. 2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology; 2009 21-23 Nov. 2009.
11. Bandrowski AE, Cachat J, Li Y, et al. A hybrid human and machine resource curation pipeline for the Neuroscience Information Framework. Database 2012;2012:bas005-bas05 doi: 10.1093/database/bas005[published Online First: Epub Date]].
12. Whetzel PL, Grethe JS, Banks DE, et al. The NIDDK Information Network: A Community Portal for Finding Data, Materials, and Tools for Researchers Studying Diabetes, Digestive, and Kidney Diseases. PLOS ONE 2015;10(9):e0136206 doi: 10.1371/journal.pone.0136206[published Online First: Epub Date]].
13. Bandrowski A, Brush M, Grethe JS, et al. The Resource Identification Initiative: A cultural shift in publishing. F1000Research 2015;4:134 doi: 10.12688/f1000research.6555.2[published Online First: Epub Date]].
14. King G. An Introduction to the Dataverse Network as an Infrastructure for Data Sharing. Sociological Methods and Research 2007;36:173–99
15. Rafał Kuć MRs. *ElasticSearch Server*. Packt Publishing Ltd, 2013.
16. PubMed Entrez Programming Utilities. Secondary PubMed Entrez Programming Utilities 2017. <https://www.ncbi.nlm.nih.gov/home/develop/api/>.
17. Research Portfolio Online Reporting Tools (RePORT). 2015
18. Demner-Fushman D, Rogers WJ, Aronson AR. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. Journal of the American Medical Informatics Association : JAMIA 2017 doi: 10.1093/jamia/ocw177[published Online First: Epub Date]].
19. Xu J, Wu Y, Zhang Y, et al. UTH-CCB@BioCreative V CDR Task: Identifying Chemical-induced Disease Relations in Biomedical Text fifth BioCreative challenge evaluation workshop; 2015.
20. Binns D, Dimmer E, Huntley R, et al. QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics 2009;25(22):3045-46 doi: 10.1093/bioinformatics/btp536[published Online First: Epub Date]].
21. Kaewphan S, Van Landeghem S, Ohta T, et al. Cell line name recognition in support of the identification of synthetic lethality in cancer from text. Bioinformatics 2016;32(2):276-82 doi: 10.1093/bioinformatics/btv570[published Online First: Epub Date]].
22. Rogers, F B. "Medical subject headings". Bull Med Libr Assoc. 1963; 51: 114–6.

23. ["History Of SNOMED CT"](#). International Health Terminology Standards Development Organisation. Retrieved 30 May 2017.
24. Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic acids research*. 2004 Jan 1;32(suppl 1):D258-61.
25. [About FMA – Contents](#) – The Foundational Model of Anatomy. Retrieved 30 May 2017.
26. Federhen S. [The NCBI Taxonomy database](#). *Nucleic Acids Res*. 2012 Jan;40(Database issue):D136-43. doi: 10.1093/nar/gkr1178. Epub 2011 Dec 1.
27. Gray, KA; Yates, B; Seal, RL; Wright, MW; Bruford, EA (Jan 2015). ["Genenames.org: the HGNC resources in 2015."](#). *Nucleic Acids Research*. **43** (Database issue): D1079-82. doi:10.1093/nar/gku1071
28. Theory Behind Relevance Scoring. Secondary Theory Behind Relevance Scoring. <https://www.elastic.co/guide/en/elasticsearch/guide/current/scoring-theory.html>.
29. Cohen T, Schvaneveldt R, Widdows D. Reflective Random Indexing and indirect inference: a scalable method for discovery of implicit connections. *J Biomed Inform* 2010;43(2):240-56 doi: 10.1016/j.jbi.2009.09.003[published Online First: Epub Date]].
30. Kanerva P, Kristoferson J, Holst A. Random Indexing of Text Samples for Latent Semantic Analysis. *Proceedings of the Cognitive Science Society* 2000;22(22)
31. Cohen T, Widdows D. Empirical distributional semantics: methods and biomedical applications. *J Biomed Inform*, 2009:390-405.
32. Cohen T, Widdows D, Wahle M, et al. Orthogonality and Orthography: Introducing Measured Distance into Semantic Space. In: Atmanspacher H, Haven E, Kitto K, Raine D, eds. *Quantum Interaction: 7th International Conference, QI 2013, Leicester, UK, July 25-27, 2013. Selected Papers*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014:34-46.
33. Widdows D, Cohen T. Graded Semantic Vectors: An Approach to Representing Graded Quantities in Generalized Quantum Models. In: Atmanspacher H, Filk T, Pothos E, eds. *Quantum Interaction: 9th International Conference, QI 2015, Filzbach, Switzerland, July 15-17, 2015, Revised Selected Papers*. Cham: Springer International Publishing, 2016:231-44.
34. Widdows D, Cohen T. The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics. 2010 IEEE Fourth International Conference on Semantic Computing; 22-24 Sept. 2010.