Working Group 4 (Use Cases and Testing Benchmarks) Recommendation Report

GROUP MEMBERS:

George Alter - ICPSR, University of Michigan
Trevor Cohen - University of Texas Houston
Dina Demner-Fushman - Lister Hill National Center for Biomedical Communications, NLM
Todd Johnson - University of Texas
Dave Kaufman - Arizona State University
Hyeoneui Kim - University of California San Diego
Steven Kleinstein - Yale School of Medicine
Thomas Radman - NIH
Susanna-Assunta Sansone - University of Oxford and Nature Publishing Group
Cui Tao - University of Texas Houston
Ratna Rajesh Thangudu - BD2K Standards Coordinating Center
Hua Xu - University of Texas Houston

i. USER NEEDS ANALYSIS REPORT:

a. PROJECT SCOPE:

To be useful, a data discovery index and any associated user interface for searching that index must consider the needs of its potential users. Unfortunately, there is very little work on assessing needs with respect to data discovery. To help fill this gap, we conducted semi-structured interviews with a variety of stakeholders. The study had two goals:

1) Understand why researchers look for datasets
2) Understand how researchers currently find datasets

During the course of answering these questions we identified factors that stakeholders feel would make data more suitable for sharing and more easily discoverable.

b. METHODOLOGY:

The study was reviewed and declared exempt by the UTHealth IRB. The interviews lasted for a period of thirty minutes each. Semi-structured interviews were conducted with researchers at various levels of experience. We developed the following questions to help structure the interview:

1) What are the datasets that you are currently involved in? Could you please name a few?
2) How do you start searching for the datasets? Do you employ any strategy?
3) In all of these datasets that you work with, has gaining access rights ever been a problem? If yes, could you state an example?

4) What are the popular formats your data is in? Has data formats been a problem to you ever?
5) What data visualization techniques do you employ? Have you faced any problems with visualizing your data?
6) What are the metadata challenges that you come across in your datasets? What would be your best suggestion to improve it?

A total of 13 participants were interviewed. Some of the interviews were conducted in person and some were conducted by phone. A detailed breakdown is given in Table 1. The interview process often started out by understanding the research area of the interviewee. The questions were then modified slightly according to the field of the researcher with the scope of the question remaining the same. The interviewer took hand written notes on each session. The subjects were chosen at the discretion of the senior leadership of bioCADDIE, the criteria involved were that the interviewee must be actively involved in biomedical research and preferably have a good amount of experience in dealing with biomedical data. A written request was sent out to each of the possible participants for interview. If the person agreed to do the interview, a suitable time was decided and the interview was conducted over phone. Before starting the interview, every interviewee was presented with the background and aim of bioCADDIE and the objective of the interview.

Table 1: Background of study participants and interview method

Professor of Neurology	Neuro Images , whole genome data , cognitive batteries, behavioral assessments and measurement of analytes and metabolytes in blood and cerebral spinal fluid	Phone
Postdoctoral research associate	Genome data, Proteomic data.	Phone
Assistant Professor in Biomedical Informatics	Gene expression data, fMRI data,	Phone
Senior Research Scientist	fMRI data , TCGA datasets. sequence data, protein expression, epigenetic data, causes pulmonary fibrotic diseases	Phone
Assistant Professor	SRA, epigenome browser, vizhub, SNPedia,	Phone
Assistant Research Scientist	Comparative Genomics, TCGA, NCBI services EPI services	Phone

	Clinvar Ensemble	
Professor of physiology, medicine and cardiology	Proteomics, Interpro, ClinVar, dbGaP, Sequence data	Phone
Project scientist	Proteomics, Interpro, ClinVar, dbGaP, Sequence data	Phone
Radiation Oncologist	EHR data, PDB data, TCGA datasets	In person
Ph.D student	Hospital compare datasets from a public source	In person
Anesthesiologist	EMR data	Phone
Assistant Professor	PDB datasets	Phone
Assistant Professor and Director of Lab	dbSNP, genomic data, cosmic, dPCG holland	Phone

The qualitative data obtained from interviews were coded into the following major categories, based on the questions above and the themes that emerged from the data:

1. Searching for data and methods
2. Data formats
3. Data visualization
4. Metadata issues
5. Enhancing discoverability

We further divided each of these major into the following subcategories:

1. Current methods
2. Challenges
3. Suggestions
4. Examples

Figure 1 shows an example of the coding process followed in this paper

Figure 1: Example of data coding method

Sl.no	Major code	Subsequent codes	
1	Searching data & Methods	Challenges, current methods, suggestion	Lots of non-clinical data have been made publicly available than the clinical data. The clinical datasets are less accessible due to a number of regulatory factors. And its not something within our role to address it. even if we tried to address it , it would lead to nowhere. At this moment people conduct a google search to start searching their datasets, instead there can be a video on popular media like youtube that guides how to search for datasets. The interviewee gave an example of dbGaP , mentioned that the search engine is largely unspecific. The organization of datasets within the dbGaP is very dated and really does not meet the existing and updated syncing of digital objects. We have to give users a menu. Standard operation protocols not adequately address in dbGaP. In any other things where we have very heavy publications for example protocols at molecular cloning volume 1,2,3,4.... We need to have protocols of data science A.B,C or something along those lines. We need to start publishing along these lines of text to guide users through this specific

c. RESULTS

I. Searching for data and methods involved:

Current Methods:

All of the biomedical researchers that we spoke to initially stated that they had no problem searching for data. Every person knew where and how to find data of interest. While no one reported any sort of problem in searching for datasets, one of the researchers did mention that people usually just start by searching on Google.

Challenges:

The challenges associated with searching for data involved mostly issues with data access. While all researchers felt they were able to successfully search for their data, gaining access to it was a major problem for clinical researchers. In contrast, for non-clinical researchers, data discoverability was not a major issue, because most of the data they need is publicly available. However, there were issues related to being able to access and use the data effectively due to lack of documentation, poor data management, and lack of standards (or adherence to standards) for data formats, coding, and metadata. The following examples help clarify the issues:

Suggestions:

1. Videos that guide users on how to search for data
2. Standards for dataset formats, coding, and metadata.
3. Standards for gaining access particularly for patient (human) sample data. Standard consent forms and process.
4. Link similar studies and sources enhances discoverability
5. Having a centralized repository that would save “number of clicks” to find interested or similar interest data

II. Data formats and Curation of data

Current methods:

1. For protein identification data, one researcher uses a raw text format so that anyone can use the data.
2. Common to need custom scripts to convert downloaded data into the required format for use.
3. Some existing tools can import and export all of the major file formats for the data types used by the tool.

Challenges:

1. Need for custom data wrangling scripts.
2. Difficult to integrate data sets.
3. Lack of granular access to data (such as only accessing a subset of data across data sets). Researchers must write scripts to extract subsets (often from datasets using different formats), then a master script to integrate the data.

Suggestions:

1. An API that returns a researcher specified subset of the data in an integrated standard format.
2. Better standards and adherence to standards for data formats.

III. Data visualization methods and challenges:

As more and more data is collected and analyzed, decision makers at all levels use data visualization software that enables them find patterns, and communicate concepts and hypotheses to others.

Current methods:

1. A chief software architect said that they have done some visualizations in d3 and the output are directed graph , they tried exporting a graph in a standard format so that they could view them in Cytoscape and Gefi for better visualization
2. Another team has developed a tool called the interrogator which talks to databases and allows them to do real time analysis on data that lives in the database. The user can also download the data and use other visualization tools that look primarily at image data or other statistical data.

Challenges:

1. Scaling large amounts of data for visualization.
2. Web based visualization tools are not very robust. There is a need for fast visualization tools that do not need to send a lot of data to the host.
3. One BD2K center reported that they are in the beginning stages of building their own visualization tool. Right now they are using neo4j and he expressed that neo4j is good for a limited amount of data but not good for appreciating data distribution among large sets of data.

Suggestions:

1. One expert expressed that it is best to use online visualization tools so that they don't occupy server space and prevents the hassle of downloading different versions.
2. The concept of a discovery dashboard came up in a discussion with a leader and expert in big data, where they envisioned different categories of users who would want to view the data, they broadly classify people into two groups: The first category are clinicians who want to view the summary/compressed forms of big data in a way that is easily digested and understood, and the second are the data scientists who want to look at the data streams and are not interested in compressed forms.
3. Develop faster methods for visualizing large amounts of data stored in the cloud. For example, one researcher suggested using dimensionality reduction techniques to enable faster visualization.

IV. Metadata Challenges:

The growing amount of data imposes a need to standardize metadata and provide metadata of importance to those seeking to find datasets.

Challenges:

1. One researcher pointed out the process flaw in creating metadata, he stated that every time someone attaches metadata to their system they do it from the perspective of their individual project goal. While he also said that it is very hard for a one fits all metadata to work very well. He suggested reuse of existing datasets when they are in formats that allow mixing. He strongly indicated that we are not going to get rid of this problem because many of the projects are application oriented. But trying to reuse vocabularies that already exist is, publishing in standard formats, having good definition and elucidation of vocabularies is certainly going to lessen the problem.
2. Another researcher mentioned that some of the metadata that he would like, but usually does not get, is a clean description of tissues that the experimental results came from. He also expressed that it would be nice to have datasets coded with ontology of methods, such as experimental methods and analysis methods.
3. Tools for creating metadata are too hard to use.

Suggestions:

1. Seek consensus from journal and domain expert committees to encode necessary information in standard formats.
2. Urge the community to agree upon open standards
3. Better co-ordination between BD2K and NIH to support standardization and commonality across the field.
4. Develop guidance on best practices and standards for metadata.
5. Develop better, easier to use tools for creating metadata.
6. Provide training on creating metadata and designing or refining ontologies.

d. ACKNOWLEDGEMENTS

The authors would like to thank the team bioCADDIE and Dr.Lucila Ohno-Machado for facilitating the interviews with the BD2K group.

ii. USABILITY ANALYSIS PLAN

New computational environments are being designed to allow easier access for researchers ‘s data, software, and computation facilities. An important factor in the utilization of such environments is data discoverability. Finding and effectively using these data requires proper indexing and tools to enable efficient search by a user who may not be an expert in data science. The vision for a data discovery index (DDI) enabled through bioCADDIE (biomedical and healthcare Data Discovery Index Ecosystem) is to define and prototype the main components of a system to index and showcase biomedical and behavioral data originating from highly diverse sources. Our current aim is to develop a prototype of a DDI that provides:

1. A free, user-friendly means for users to locate data sets of interest
2. Standardized, searchable information (metadata) about the contents of a data set

3. Metadata about accessibility of the data and of any existing application programming interfaces (APIs)
4. Privacy protection for the individuals whose data are represented in the data set

As we systematically evaluate the features required in a DDI interface and develop new functionalities for the biomedical domain, usability analyses should be conducted to guide the development of the user interface and to assess how it compares to existing approaches for locating data sets. A highly usable interactive interface is essential for domain experts to quickly search and find datasets for their use. Fundamentally, the task of data searching as it occurs in the context of an interactive system is currently poorly understood. User-centered design of information technology can improve the speed with which tasks are conducted. However, existing data discovery interfaces have not been evaluated from a usability perspective. Therefore, the usability studies should be conducted to facilitate development of an efficient interface via an iterative process.

a. Objectives

Objective 1 – Collect feedback from users that use the DDI interface in the project. The interaction with this group is remote and the information collected should be used to inform the functionality, usability and design of the user interface

Objective 2 – Gather information from beta (test) users about researchers' dataset seeking needs, strategies, and challenges. This data should also refine the design of the bioCADDIE DDI user interface and its functions

Objective 3 – Determine whether integrating several data sources in one place would improve discoverability of biomedical data by comparing bioCADDIE to researchers' standard approach to finding data.

The usability evaluation of the BioCaddie user interface should be performed using standard usability methods, including but not limited to: Performance measures, Thinking-aloud, Interviews, Questionaires, and user feedback. The resulting data should be analyzed using descriptive statistics as well as qualitative analysis as befits the study design. In previous studies, combinations of verbal protocol analysis and granular analysis of video captured data were used to characterize the relationships between actions taken and the thought processes underlying these actions. The analyzed data will assist in evaluating users' cognitive load as well as their overall experience when interacting with the website. In addition, this analysis should reveal novel implicit features that are inferred from the datasets during the process of searching for datasets. These findings can then inform the development of subsequent iterations of the interface, each of which can be evaluated with respect to the relationship between ease of searching and accuracy.

b. Study Population

The usability study consists of three different phases, as specific target audiences are required for each of these phases. The first target audience will be users interested in using the DDI and who wish to assist in developing the dataset query tool. Online user behavior can be collected and feedback solicited from the users to obtain more detailed insight into

the challenges that they face as they interact with the website, and the workflow that they use to find their choice of data sets. The participants for the second and third phases will be representative of the population of biomedical researchers who have the necessary domain expertise to participate in this study. For phase I, soliciting and collecting feedback should be from as many online users as possible. The minimal number of users for this phase is expected to be 30. In phase II, a series of 2-5 small scale formative usability studies (with 5-8 individuals each) should be conducted to iteratively evaluate different versions of bioCADDIE. For Phase III, a controlled usability study is to be done to examine differences between bioCADDIE and a standard tool for finding datasets. This is a 2x2 within-subject design, with Task and Tool as independent variables with two levels each. For this, a minimum of 5 data points per cell (10 subjects altogether) is required, but 20 subjects (10 data points per cell) is the target recruitment for this phase. However, data from Phase II should be used to do a power analysis and then revise the protocol if more subjects are required.

c. Timeline

The protocol for these studies is now under expedited review by the UTHealth Committee for the Protection of Human Subjects. We expect to begin Phase I and Phase II by the end of February 2016. Phase III will be done following the first public release of bioCADDIE.

iii. BENCHMARKING DATASET COLLECTION

In addition to user-centric system, the working group developed a set of recommendations for a system-centric evaluation – a Cranfield-style evaluation in the tradition established by Text REtrieval Conferences (TREC) over the past decade. This will involve the development of a set of reference queries, which will be run using different information retrieval systems. The top results from each system will be pooled, and annotated for relevance by annotators with the prerequisite biomedical expertise, resulting in a reference standard that can be used for comparative evaluation of different algorithmic approaches to dataset search and retrieval.

The search engines recommended to generate the initial pool of results were selected so as to incorporate a range of information retrieval approaches, and mitigate the danger of the reference set being biased toward a particular established approach (as only datasets retrieved in this initial pool are eligible for annotation). These are described in Table 1.

Table 1: Information Retrieval systems for the initial pooling experiments

System	Description and key algorithms
Apache Lucene (https://lucene.apache.org/)	Underlies the ElasticSearch implementation used for the bioCADDIE prototype. Vector Space model (Salton 1975) with capacity for Boolean logic.
Indri (http://www.lemurproject.org/indri/)	Language Models (Ponte and Croft, 1998) and Inference Networks (Turtle and Croft, 1991).
Terrier (http://terrier.org/)	Divergence from Randomness (Amati and van Rijsbergen 2002), BM25 (Robertson et al 1994).
Semantic Vectors	Extends Apache Lucene. Implicit query expansion

(https://github.com/semanticvectors/semanticvectors)	via term similarity - Random Indexing (Kanerva et al 2000), Latent Semantic Indexing (Deerwester *et al* 1990).

In addition, we discussed the idea of providing expanded versions of the queries using the terminology service that is under development for the bioCADDIE prototype, so that results that depend upon terminology-based query expansion are included in the evaluation pool.

iv. TESTING QUERIES

The Working Group also generated a set of constraints for the queries themselves. These were as follows:

(1) Between 30 and 50 queries should be generated.

(2) They should be based on the use cases that emerged from WG3 so they assess the capabilities required to meet the broader goals of the project.

(3) They should relate to data that has already been indexed, so they can be used to evaluate the current prototype.

Therefore, reevaluation of the use cases was a prerequisite to construction of the test queries. We focused our attention on the Competency Questions developed by Working Group 3 X. Typical examples are shown in Table 2, below.

Table 2: examples of competency questions produced by WG3X.

Internal bioCADDIE code	Competency question	Functional requirements/ constraints
BGUC1-2	Search for **data type** x related to **disease** x and **disease** y to compare **behavioral studies** (HD and ADHD)	Identify data types and diseases, normalize terms Is behavioral studies a data type? Taxonomy of data types, Boolean logic
WPUC2	Search for **data types** x and y related to the same **biological process** z	Identify data types and biological process, normalize terms Boolean logic (UMLS for clinical/biological processes, Entrez for gene/)?
BGUC1-1	Search for **disease** x **data** of all **types** across **all databases** (Note: these first three use cases are linked; also there is a **Common Data Element** for the **disease** x [HD])	Disease and data type disease terms and data types to be normalized recognize duplicate entries across databases

Members of the working group reviewed the competency questions to identify which met the proposed constraints. Questions involving datasets containing Protected Health Information (PHI) were excluded upfront, as these do not relate to currently indexed content. Questions that involved returning results other than datasets (such as software packages or drug-drug interactions) were also excluded. The most commonly occurring entity types in the remaining Competency Questions (n=30) were data type, disease type, biological process and organism. In addition, a number of Competency Questions concerned characteristics of the data sets relating to permissions and data format. A smaller number of the questions concerned surrogate indicators of dataset quality, such as funding source or number of referring publications. Based on this review, the working group recommended focusing on queries concerning the most prominently featured entity types (data type, disease type, biological process and organism), with inclusion of a few queries to do with permissions and/or format, as this competency related to a number of the use cases.

Regarding functional requirements, the working group observed that indexing, recognition and normalization of key concepts may be required to respond to many of the queries. For example, many of the questions required identifying datasets that concerned a particular disease or disorder (n=16), and a particular data type (n=39). Data types were often specified using general terms such as ‘omics data’ or ‘imaging data’, suggesting the need for a taxonomy of data types. In addition, many of the questions suggested the need for Boolean connectives, and some concerned the more general notion of ‘similarity’ between concepts. These observations were shared with the development group, and plans are underway to address the identification and normalization of key entity types.

Work to date based on the recommendations

The working group recommended “freezing” an edition of the bioCADDIE prototype index, to provide a consistent evaluation set. Based on this recommendation, we have extracted the data indexed by December 1st, 2015. This includes records for 302,279 data sets, derived from RCSB Protein Data Bank (PDB), Gene Expression Omnibus (GEO), GEMMA, Array Express, Sequence Read Archive (SRA), BioProject, GWAS, LINCS and dbGAP. Data types include protein (sequence and structure), phenotypic data and gene expression data. In addition, our team of annotators has developed a set of queries concerning the entity types identified during review of the competency questions, with the constraint that relevant datasets exist in this set of documents. Some examples are included in the Appendix.

APPENDIX: INSTANTIATED QUERIES FOR SPECIFIC COMPETENCY QUESTIONS

WPUC2 - Search for **data types** x and y related to the same **biological process** z

1. Search for data of all types related to the SHH gene in human embryogenesis across all databases
2. Search for data on hES cells from IVF embryos related to human embryogenesis across all databases
3. Find data on cellular differentiation in mesenchymal stem cells related to human embryogenesis across all databases
4. Find data of all types related to *TGF-β signaling* pathway across all databases
5. Search for data of all types published within the last ten years related to the TGF-β signaling pathway
6. Search for data of all types on *cheA* gene related to bacterial chemotaxis across all databases
7. Search for data of all types related to *Helicobacter pylori* in the stomach and chemotaxis across all databases
8. Search for data of all types on signal transduction related to bacterial chemotaxis across all databases
9. Find protein sequencing data related to bacterial chemotaxis across all databases
10. Search for data of all types related to bacterial chemotaxis with mentions of datasets in literature
11. Search for data of all types for *GLUT1* gene related to glycolysis across all databases
12. Search for data of all types related to CRC tissue in humans related to glycolysis across all databases
13. Find data of all types on synaptic growth and remodeling related to glycolysis in the human brain across all databases
14. Search for gene mutation datasets related to glycolysis across all databases
15. Search for data of all types related to glycolysis in the human brain with more than 10 citations across all databases
16. Search for data of all types related to *MIP-2* gene related to biliary atresia across all databases
17. Search for data of all types related to bile secretions in SR-BI KO mice across all data bases
18. Find all related data on cholesterol storage in mice with biliary atresia across all data bases
19. Search for gene expression data related to bile acid secretions in patients with biliary atresia across all databases

20. Search for all data types related to bile acid secretions in mice with biliary atresia that that require NPG site licensing
21. Search for all data types related to gene *TP53INP1* in relation to p53 activation across all databases
22. Search for all data types related to mouse embryonic fibroblasts related to p53 activation during oxidative stress across all databases
23. Find all data types related to inflammation during oxidative stress in human hepatic cells across all databases
24. Search for protein sequencing data related to p53 activation in human hepatic cells across all databases
25. Search for all data types related to p53 activation during oxidative stress across all databases published within the past 15 years that require a Data Use Certificate (DUC)
26. Search for data of all types for memory augmentation studies across all databases
27. Search for gene expression data regarding memory augmentation studies across all databases
28. Search for gene expression and genetic deletion data regarding memory augmentation across all databases
29. Search for gene expression and genetic deletion data that mention CD69 in memory augmentation studies across all databases
30. Search for gene expression data mentioning CD69, and memory augmentation in a tab-delimited format across all databases
31. Search for data of all types for pigmentation across all databases
32. Search for genomic sequence data regarding pigmentation across all databases
33. Search for gene expression and genomic sequence data regarding pigmentation across all databases
34. Search for gene expression and genomic sequence data that mention MC1R in pigmentation studies across all databases
35. Search for gene expression data for MC1R, and pigmentation in a tab-delimited format across all databases
36. Search for data of all types for aging across all databases
37. Search for protein aggregation data regarding aging across all databases
38. Search for protein aggregation and gene expression data regarding aging across all databases
39. Search for protein aggregation and gene expression data that mention trace elements in aging studies across all databases
40. Search for gene expression data for trace elements and aging in a plain text format across all databases
41. Search for data of all types for cellular differentiation across all databases

42. Search for gene expression data regarding cellular differentiation across all databases
43. Search for gene expression data that mention E2F cell line in cellular differentiation across all databases
44. Search for a gene expression data that mention E2F cell line and Ras activation in cellular differentiation studies across all databases
45. Search for gene expression data for E2F cell line, Ras activation, and cellular differentiation in XML format across all databases
46. Search for data of all types for cellular respiration across all databases
47. Search for genetic sequence data regarding cellular respiration across all databases
48. Search for genetic sequence data and proteomics data regarding cellular respiration across all databases
49. Search for genetic sequence data and proteomics data that mention 5hmC in a cellular respiration studies across all databases
50. Search for data of all types for a genetic sequence, proteomics, 5hmC, and cellular respiration in a RAW format across all databases

WPUC3 - Search for **data types** x (genome data) with **biological process** (mutations) y and z in **species/organism** a for **phenotype** b

1. Search for data on *BRCA* gene mutations and the estrogen signaling pathway in women with stage I breast cancer
2. Search for data of all types for *PTEN* gene related to women with stage I breast cancer
3. Search for genomic data of all types in epithelial cells in women across all databases
4. Find data of all types on the regulation of DNA repair related to the estrogen signaling pathway in breast cancer patients across all databases
5. Find all gene expression data related to the estrogen signaling pathway across all databases
6. Search for data of all types related to the estrogen signaling pathway in women with stage II breast cancer published within the last 10 years
7. Search for gene expression data on fatty acid oxidation and adaptive immune responses in men over 50 with cardiovascular disease
8. Search for data of all types related to the *LDLR* gene related to cardiovascular disease across all databases
9. Search for data of all types on adipocytes in men over 50 with cardiovascular disease across all databases
10. Search for data of all types related to calcium signaling and cardiomyopathy in patients with CD across all databases

11. Search for protein sequencing data related to cardiovascular disease in men over 50 across all databases
12. Find data of all types related to cardiovascular disease in men over fifty across all databases with more than 10 citations
13. Search for gene expression datasets on photo transduction and regulation of calcium in blind *D. melanogaster*
14. Find data of all types related to the *org-1* gene related to *D. melanogaster* across all databases
15. Search for data of all types on pigment cells of ommatidia related to blindness in *D. melanogaster* across all databases
16. Search for data of all types related to circadian sleep/wake cycles in vision impaired/ blind *D. melanogaster* across all databases
17. Search for proteomic data related to regulation of calcium in blind *D. melanogaster* across all databases
18. Find data of all types on blindness in D. melanogaster across all databases redistributed for free under Open Database License (ODC-ODbL)
19. Search for protein sequencing datasets on circadian rhythms and insulin secretion in obese *M. musculus*
20. Search for data of all types related to the *ob* gene in obese *M. musculus* across all databases
21. Find data of all types on adipose tissue related to obese *M. musculus* across all databases
22. Search for data of all types related to energy metabolism in obese *M. musculus* across all databases
23. Search for gene expression datasets related to insulin secretion in obese *M. musculus* across all databases
24. Search for all types of data related to obese *M. musculus* across all databases that require no access requirements
25. Search for proteomic data on iron metabolism and hypoxia signaling in iron deficient C. *elegans*
26. Search for data of all types related to *Tfrc* gene in in iron deficient C. *elegans* across all databases
27. Search for data of all types related to extraintestinal tissues in C. elegans deficient in iron across all databases
28. Search for data of all types on aging in iron deficient C. *elegans* across all databases
29. Search for gene expression data on hypoxia signaling in iron deficient C. *elegans* across all databases
30. Find data of all types related to iron metabolism in iron deficient C. *elegans* that have been published within the last five years across all databases

31. Search for data of all types for different eye colors across all databases
32. Search for data of all types that mention different eye colors in human across all databases
33. Search for data of all types that mention replication and different eye colors in human across all databases
34. Search for mutation data mentioning replication and different eye colors in human across all databases
35. Search for mutation data that mention replication, HERC2, and different eye colors in human across all databases
36. Search for genome data that mention replication, mutation, HERC2, and different eye colors in human in a txt format across all databases
37. Search for data of all types for a human cell line across all databases
38. Search for data of all types for cell differentiation in human cell line across all databases
39. Search for mutation data for cell differentiation and in human cancer cell line across all databases
40. Search for gene expression and mutation data for cell differentiation in human cancer cell across all databases
41. Search for gene expression data for a cell differentiation and mutation in human cancer cell in a tab-limited format across all databases
42. Search for data of all types for baldness in females across all databases
43. Search for data of all types that mention differentiation in female baldness across all databases
44. Search for transcriptional data mentioning differentiation in female baldness across all databases
45. Search for transcriptional data that mention differentiation, EDA2R gene in female baldness across all databases
46. Search for genomic and transcriptional data mentioning differentiation, EDA2R gene in female baldness across all databases
47. Search for genomic and transcriptional data mentioning differentiation, EDA2R gene in female baldness in xml format across all databases
48. Search for data of all types for anxiety-like phenotype in human male across all databases.
49. Search for mutation data from an anxiety-like phenotype in human male across all databases.
50. Search for mutation and genetic variation data in an anxiety-like phenotype in human male across all databases.
51. Search for genetic data in an anxiety-like phenotype in human male across all databases.

52. Search for genetic data in an anxiety-like phenotype in human male across all databases in a FASTQ format.
53. Search for data of all types for developmental phenotype in a fungal species across all databases.
54. Search for sequencing data for a developmental phenotype in a fungal species across all databases.
55. Search for sequencing data that mention transformation in a developmental phenotype in a fungal species across all databases.
56. Search for proteomic and sequencing data that mention transformation in a developmental phenotype in a fungal species across all databases.
57. Search for proteomic and sequencing data that mention transformation in a developmental phenotype in a fungal species across all databases in txt format.

BGUC1-1 - Search for **disease** x **data** of all **types** across **all databases**

1. Search for data of all types on multiple sclerosis of all types across all databases
2. Search for data of all types for *HLA-DRB1* gene in multiple sclerosis across all databases
3. Search for data on oligodendrocytes in female human nervous tissue related to multiple sclerosis across all databases
4. Find data on T-cell homeostasis related to multiple sclerosis across all databases
5. Find gene expression datasets on multiple sclerosis across all databases
6. Search for all data types on multiple sclerosis published within the last 15 years.
7. Search for data of all types on Huntington's disease across all databases
8. Search for all data for the *HTT* gene related to Huntington's disease across all databases
9. Search for data on neural brain tissue in transgenic mice related to Huntington's disease across all databases
10. Search for data on transcriptional regulation related to Huntington's disease across all databases
11. Find protein sequencing data on Huntington's disease across all databases
12. Search for all types of data on Huntington's Disease across all databases that require NPG site licensing
13. Search for data of all types on Parkinson's disease across all databases
14. Search for all data on the *SNCA* gene related to Parkinson's disease across all databases
15. Search for data on nerve cells in the substantia nigra in mice across all databases
16. Search for data on long-term potentiation related to Parkinson's disease across all databases

17. Search for proteomic data on Parkinson's Disease across all databases
18. Search for all data types related to Parkinson's Disease patients over 40 years of age across all databases
19. Search for data of all types on Myasthenia Gravis across all databases
20. Search for data on the *TNIP1* gene related to MG across all databases
21. Search for data on extraocular muscle cells in humans related to MG across all databases
22. Find data on the NF-κB signaling pathway in MG patients across all databases
23. Search for gene expression datasets related to MG across all databases
24. Search for all data types on MG that have been published within the last 5 years across all databases
25. Search for data of all types on Friedreich's ataxia across all databases
26. Search for data on *FXN* gene in relation to Friedreich's ataxia across all databases
27. Search for data on heart cells related to Friedreich's ataxia across all databases
28. Find data on mitochondrial respiration related to Friedreich's ataxia across all databases
29. Search for protein sequencing data related to Friedreich's ataxia across all databases
30. Search for all data types on Friedreich's ataxia that can be accessed free under the Open Data Commons licenses across all databases
31. Search for data of all types for gestational diabetes across all databases
32. Search for data of all types mentioning GLS2 gene in gestational diabetes across all databases
33. Search for data of all types mentioning GLS2 gene and placental tissue in gestational diabetes across all databases
34. Search for gene expression data mentioning GLS2 gene, placental tissue in gestational diabetes across all databases
35. Search for gene expression data mentioning GLS2 gene, placental tissue, in gestational diabetes in chip format across all databases
36. Search for data of all types for osteosarcoma across all databases.
37. Search for data of all types that mention ALP gene in an osteosarcoma across all databases.
38. Search for data of all types that mention ALP gene and Saos-2 cells for an osteosarcoma across all databases.
39. Search for data of all types that mention ALP gene, Saos-2 cells, and cell differentiation for an osteosarcoma across all databases.
40. Search for data of all types for an ALP gene, Saos-2 cells, and cell differentiation for an osteosarcoma in a txt format across all databases.
41. Search all proteomic data available for arthritis across all databases.
42. Search for data of all types for rheumatoid arthritis across all databases.

43. Search for data of all types that mention HLA-DRB1 gene in rheumatoid arthritis across all databases.
44. Search for data of all types for that mention HLA-DRB1 gene and lymphoblastoid cells for rheumatoid arthritis across all databases.
45. Search for sequence polymorphism data that mention HLA-DRB1 gene, lymphoblastoid cells for rheumatoid arthritis across all databases.
46. Search for sequence polymorphism data that mention HLA-DRB1 gene, lymphoblastoid cells for rheumatoid arthritis in FASTA format across all databases.
47. Search for data of all types for Down's syndrome across all databases.
48. Search for data of all types using DSCAM gene in Down's syndrome across all databases.
49. Search for data of all types for DSCAM and BAC genes in Down's syndrome across all databases.
50. Search for data of all types that mention DSCAM and BAC genes and cell differentiation in Down's syndrome across all databases.
51. Search for data of all types that mention DSCAM and BAC genes and cell differentiation in Down's syndrome with associated literature across all databases.
52. Search for data of all types for giardiasis across all databases.
53. Search for data of all types mentioning TPI 1 gene in giardiasis across all databases.
54. Search for data of all types that mention TPI 1 gene and assemblage A in giardiasis across all databases.
55. Search for data of all types that mention TPI 1 gene, assemblage A, and glycolytic process in giardiasis across all databases.