Building an NIH Data Catalog: Bit by Bit

By: Kevin Read, Associate Fellow 2012-13

Date: August 15, 2013

Project Sponsors: Jerry Sheehan OD, Mike Huerta OD

Table of Contents

Abstract

OBJECTIVE The purpose of this project was to a) develop a set of core, minimal metadata elements that would be used to describe datasets, and b) carry out a study to identify datasets in NIH funded articles from PubMed and PubMed Central (PMC) that *do not* provide an indication that their data has been shared in a data repository or registry. These efforts will inform the BD2K initiative and a planned NIH Data Catalog.

METHODS An analysis of the metadata schemas for all NIH data repositories was undertaken. Commonalities from these data repositories were identified, mapped to existing data-specific metadata standards from DataCite and Dryad, and then were integrated into MEDLINE XML metadata to attempt to establish a sustainable and integrated metadata schema.
The second phase of this project identified datasets in articles from PubMed and PMC by searching specifically for NIH funded articles from the year 2011. After excluding articles that contain mention of datasets being deposited in existing repositories, thirty staff members from NLM and B2DK were recruited to analyze a random sample of the results to identify how many, and what types of datasets were created per article.

RESULTS A preliminary set of minimal metadata elements were developed that could sufficiently describe NIH-funded data sets and be integrated within MEDLINE's schema, with minor additions. For the "orphaned" datasets study, a first phase of statistical analysis was completed. While the percentage of difference between annotators for validity came to 43% - a significantly high number – the study still resulted in useful information for BD2K. Based on these findings, we found that for NIH funded research articles from 2011, on average there are 2.92 datasets created per article; 87% of datasets created are completely new data; and over 50% of data created throughout the course of biomedical research are completed using live human or non-human animal subjects.
At present (August 2013), results of the second phase of analysis for PubMed and PMC article datasets are pending once we receive feedback from a biostatistician.

CONCLUSION The efforts to develop a minimal set of metadata elements and identify the amount, and types of datasets that are produced from NIH funded articles will serve to inform the BD2K's initiative to build an NIH Data Catalog going forward.

Introduction

On February 22, 2013 the Executive Office of the President and the Office of Science and Technology Policy (OSTP) created a memorandum to increase access to the results of federally funded scientific research. For the National Institutes of Health (NIH), the memo represents a new step towards enhancing its current public access policy in that it requires each federal agency to share their scientific research in the form of publications and a new directive that will also require the sharing of digital scientific data [1]. In order to meet this new directive, the NIH has developed Big Data to Knowledge (BD2K), an initiative to address how best to manage and utilize the large amounts of biomedical data that new technologies can generate in the course of scientific research.

A major focus of the BD2K initiative is to develop a comprehensive catalog of NIH funded research datasets from all areas of biomedical research. The catalog is meant to be transformative, allowing information about datasets to be discoverable, citable, and linked to the scientific literature with the goal of raising the prominence of data in biomedical research and scholarship. As a result of this initiative, an NIH Data Catalog Working Group was formed to work on addressing these issues. A workshop meeting was scheduled in August 2013 to inform the process of building and supporting an NIH Data Catalog.

To inform the creation of an NIH Data Catalog, the Associate Fellow from the National Library of Medicine (NLM) was asked to complete a project with two separate components; one of the key elements of the envisioned data catalog was the characterization and description of datasets using a set of minimal metadata elements – the goal being to ensure that the description of data is consistent, and that it is described in enough detail that it can be interpreted by a user of the NIH Data Catalog. This effort represented the first phase of the Associate project where an analysis of metadata from existing NIH data repositories was carried out to provide a minimal set of metadata for the NIH Data Catalog.

The second component of this project was designed to provide information about the data landscape at the NIH. Before attempting to construct an NIH Data Catalog of NIH funded datasets, this phase attempted to answer a number of questions that BD2K had about the current state of data created through scientific research: How much data is created in a given year at the NIH? Is data being shared after an article is written? What types of data are being created?

These questions served as motivation for the second phase of this project, which involved searching for datasets that *had not* been shared in an existing data repository – a concept we coined

as “orphaned” datasets. The goal of this effort was to gain a better understanding of what types of “orphaned” datasets exist as well as how many are created in a given year.

This report will outline in detail the two phases of this project: the discovery and recommendation of a minimal set of metadata elements and the analysis of NIH funded “orphaned” datasets. The results of these two phases were instrumental for informing the creation of an NIH Data Catalog.

Methods

Identifying Common Minimal Metadata Elements

To identify a common set of minimal metadata elements that would be used to describe datasets within the NIH Data Catalog, we identified a sample set of NIH data repositories to extract their metadata and search for commonalities. For this project we used the 45 NIH Data Sharing Repositories that are listed on the NIH Office of Extramural Research – NIH Sharing Policies and Related Guidance on NIH funded Research Resources – webpage [2].

The 45 repositories were selected because they represent a complete sample of NIH supported data repositories. Selecting this sample was also an attempt to reduce the burden on researchers; if BD2K can make the NIH Data Catalog interoperable with the 45 existing NIH data repositories, the researcher would only have to provide metadata for the specific repository where they deposited their datasets, and the metadata they submitted could be cross-walked to the NIH Data Catalog.

Following the submission process from each data repository, the metadata descriptors were collected. Each descriptor field was then recorded into a spreadsheet where it was defined in the context of its respective repository. This process was repeated for each repository in order to gather all of the available metadata [Suppl-1].

Analyzing Metadata Schemas for Commonalities

The main goal of this exercise was to identify commonalities in the 45 NIH data sharing repository metadata descriptors. A metadata descriptor was considered ‘common’ if it was identified within the 45 repositories more than twice. The reason for choosing such a small number for commonality was due to the varied range of data types represented in the repositories; the subject and types of data represented in these repositories spanned zebrafish genotypes to

chemical compounds to cancer imaging. This broad scope resulted in few commonalities across the sample of 45; therefore identifying more than two commonalities was considered to be a success.

Once commonalities were identified, descriptors were categorized into broad classifications that best represented that metadata element; this step was taken to account for the amount of variation that was identified from the 45 data repositories metadata descriptors. One example of these broad classifications was Authorship; within this category data repositories used different metadata descriptors to described authorship such as author, principal investigator(s), data author, data submitter, and contributor(s). Because the metadata descriptors varied so widely across each repository, it was important to create a classification system to help identify those commonalities.

Creating a Taxonomy of Common Metadata Elements

A taxonomy was created to help identify the metadata variations used by the 45 data repositories [Suppl-2]. The taxonomy was organized by providing a major classification that represented the common metadata element and then listed underneath was the minor variations of metadata descriptors that refer to that major classification in the hierarchy. The total number of times a major classification was identified in the metadata spreadsheet is located in parentheses next to its heading in bold. The number of repositories that use a particular metadata descriptor is also indicated beside each element in parentheses [Suppl-2].

Because there were so many descriptor variations across the 45 metadata schemas, the taxonomy was instrumental to informing the development of minimal metadata elements for the NIH Data Catalog.

Mapping to Dryad and DataCite

To validate the commonness of the metadata extracted from the 45 NIH data sharing repositories, the most common metadata descriptors were included in a side-by-side comparison with DataCite's metadata schema [3] [Suppl-3] and Dryad's metadata schema [4] [Suppl-4]. Both DataCite and Dryad were selected because their metadata schemas are kept up to date, and they describe a vast range of data ranging from biomedical datasets to social science datasets. This measure was also designed to fill in gaps in the metadata descriptors from the 45 NIH data repositories.

After mapping to both DataCite and Dryad was complete, a more thorough set of common metadata elements were compiled. These common metadata elements were then mapped to the NLM's existing MEDLINE metadata schema [5] for journal articles in PubMed and PubMed Central

(PMC) to test for interoperability and sustainability within an NIH system that is already in place. Mapping to MEDLINE was also carried out because it was thought it could provide a way to link datasets to their associated articles in PubMed, which is one of the main goals of the NIH Data Catalog.

Mapping Metadata to MEDLINE

The same method of mapping that was carried out for DataCite and Dryad were applied to MEDLINE, where the new set of metadata elements derived from our previous mapping was compared side-by-side with MEDLINE's metadata elements for journal articles. The traditional definition used for each MEDLINE metadata element was modified to account for the changes that would be required to describe datasets. Furthermore, allowed values were altered if necessary to address the needs of a dataset [Suppl-5, Fig. 1].

Fig. 1 Mapping to MEDLINE – Repository

Common Metadata Element	MEDLINE Metadata Element	Definition modified for NIH Data Catalog	Allowed Values
Data Location	DataBank	The name of the entity that holds, archives, publishes, distributes, releases, issues or produces the data.	**Values**: DataBankName: Name of repository where data is located. AccessionNumber: accession numbers associated with the dataset. **Generates:** Attribute: DataBankList: Additional repositories where the data could be located.

The above example provides an indication of how the common metadata element 'Data Location' could be mapped to the MEDLINE metadata element DataBank. The element DataBank is traditionally applied to scientific publications that exist within MEDLINE that refer to when data has been shared within a specific, pre-approved NLM data repository [6]. It is believed that this DataBank element could be expanded to incorporate any data repository where NIH funded researchers share their data.

Mapping to MEDLINE proved to be the final step towards creating a minimal set of metadata elements for the NIH Data Catalog. The final set of metadata was finalized based on the

fact that they were supported by strong evidence from the metadata of 45 different NIH funded repositories; mapped consistently to the metadata schemas of existing large-scale data services DataCite and Dryad; and were found to be mostly interoperable and sustainable within MEDLINE's existing metadata schema for journal articles.

Analysis of "Orphaned" Datasets in PubMed and PMC

Exclusion Methodology

To perform an analysis of NIH funded "orphaned" datasets, a number of exclusions had to be made from the literature in both PubMed and PMC to find articles that created datasets that had not been shared in a data repository. For this study, NIH funded articles from 2011 were searched exclusively because they represent the most current complete set of articles for a given year.[1]

Once all NIH funded articles were retrieved in PubMed (*n* =113,089*)*, the following exclusions were made: all non-PMC articles were eliminated; and all non-research articles were removed by excluding articles with the Publication Type (PT) review, editorial, news, letter and introductory journal article.

Next, a series of measures were taken to remove articles that indicated when an author shared their data in a specific repository. First all articles with the MeSH heading (MH) Molecular Sequence Data were excluded; this measure was taken because all articles with this MH were thought to also have deposited their data in the GenBank sequence database. In the second phase, all articles that had a Secondary Source Identifier[2] [SI] were removed because this identifier refers to when an author has shared their data in a specific NLM-approved repository such as ClinicalTrials.gov or Protein Data Bank. Completing these steps amounted to a number of exclusions that reduced our larger set of 113,089 NIH funded articles down to a smaller set of 71,910 that contained "orphaned" datasets [Suppl-6, Fig. 2].

Fig. 2 PubMed Exclusion Strategy

Search **2011 [dp] AND (NIH [gr] OR Research Support, N.I.H., Extramural [pt] OR Research Support, N.I.H., Intramural [pt]) AND medline [sb] NOT (GDB [si] OR GENBANK [si] OR OMIM [si] OR PDB [si] OR PIR [si] OR RefSeq [si] OR SWISSPROT [si] OR ClinicalTrials.gov [si] OR ISRCTN [si] OR GEO [si] OR PubChem-Substance [si] OR PubChem-Compound [si] OR PubChem-BioAssay [si]) NOT molecular sequence data [mh:noexp] AND pubmed pmc all[sb] NOT review**	71910

[1] The year 2012 was not selected because many of the articles are still embargoed, and therefore would not provide a representative set.

[2] The **[SI]** qualifier identifies a secondary source that supplies information, e.g., other data sources, databanks and accession numbers of molecular sequences that are pre-approved by the NLM.

[pt] NOT letter [pt] NOT news [pt] NOT editorial [pt]	

An additional step that was taken to exclude articles that mention the sharing of data was to eliminate all articles that mention a data repository in the Acknowledgments field [7] of PMC. The Acknowledgments field is often used by authors to indicate when they have shared their data in a specific repository. Using the 45 NIH Data Sharing Repositories webpage [2] as our gold standard to gather a list of NIH-specific data repositories, the keyword variations and acronyms for each repository were searched in the Acknowledgments field in PMC for 2011 [Suppl-7]. Additionally, we added the terms "DataCite" and "Dryad" to the search set because we found mention of these data services in articles when preparing the exclusion strategy. The results of this search (814 results) in PMC were subsequently added to the exclusion strategy described above in PubMed.

The final measure taken to eliminate articles that mention data sharing was to search for the same keyword variations and acronyms from the 45 data repositories in the PMC XML [Suppl-8]. This step was taken to fill in any gaps the two previous strategies may have missed, and to search beyond the Acknowledgments field in PMC to find more mentions of data repositories. After the three methods of exclusion were completed, the total number of remaining NIH funded datasets that contained "orphaned" datasets was 69,657. We then performed a random sample with a 95% confidence interval to bring our total number for analysis to 383 articles.

For the study we recruited 30 members of NLM and BD2K staff to be annotators for the 383 articles. Annotators were subject experts working in a variety of disciplines including indexers of biomedical literature, biomedical informaticians, physicians, neuroscientists, molecular biologists, librarians and organizational directors. Each annotator was assigned 25 articles, and two participants were assigned the same 25 articles for inter-rater reliability.

Analysis of 383 Articles with "Orphaned" Datasets

Each annotator was asked to review each article they were assigned in its entirety to answer questions for each dataset collected from the research described in an article. For the purpose of this study, we defined a dataset as any data that was collected to inform the results of an article. We also developed controlled vocabularies for the data types and subjects of study in hopes that they would match well with the datasets created within the 383 articles. Listed below are the series of questions we had annotators answer for each dataset they found:

1. **What category of dataset was used for the research described in the article?**

A. New dataset
   Examples: lab results or blood pressure measurements after administration of a new drug treatment, new survey results, or mutation analysis of a tumor
B. Existing dataset with modifications or added value
   Example: research using previously collected phenotype data combined with newly collected genotype data
C. Existing dataset as-is
   Example: study using pre-existing survey data to answer a new question
D. None
   Example: No indication that data was created, such as an article about another study that has already been completed.

2. Were live human or animal subjects used in the collection of the data?

A. Yes
B. No

3. If new dataset(s) were created, what were the subject(s) of study (from which or whom the data was collected)?

A. Human (e.g. includes human subjects, tissues or cells)
B. Non-human Animal (e.g. includes animal subjects, tissues or cells)
C. Plant (e.g. includes plant subjects, tissues or cells)
D. Immortalized cell lines (e.g. HeLa, HEK)
E. Bacteria
F. Virus
G. Computational model
H. Other (please provide your best judgment as to the subject of study)

4. If new dataset(s) were created, what type(s) of data were collected:

A. Image
   a. Example: two or three dimensions
B. Genetic or genomic
   a. Example: SNP or genetic insertion/deletion
C. Chemical
   a. Example: chemical and crystal structures, spectra, reactions and syntheses
D. Biochemical
   a. Example: structures, functions and interactions of proteins, nucleic acids, carbohydrates, lipids, etc.
E. Electrical (electrophysiological)
   a. Example: EEG or EKG
F. Optical – non-image
   a. Example: fluorescence signals indicating a biochemical event (non-image)
G. Behavioral – non-questionnaire/survey
   a. Example: reaction times during a working memory task

H. Computational simulation or model
    a. Example: computer simulation of fluid mechanical forces in cardiovascular disease development and therapy
I. Magnetic resonance – non-image
    a. Example: nuclear magnetic resonance (NMR), electron spin resonance (ESR), and electron paramagnetic resonance (EPR)
J. Structural or anatomical
    a. Example: measures of shape, size and other spatial features of molecules, organelles, cells, tissues, organs or organisms
K. Physiological
    a. Example: measures of function across interacting parts of the cell, tissue or organism
L. Questionnaire/survey
    a. Example: collected in epidemiology or health services research, self-report by individuals, etc.
M. Clinical measures
    a. Example: data assessing quality of care and patient satisfaction
N. Geospatial
    a. Example: data points related to particular places on the earth
O. Other (please use your best judgment if the type of data is not represented above)

5. **If an existing dataset was used, please specify which one:**
   Examples: a pre-existing survey, MIMIC data, a computational model, previously collected phenotype data, etc.

Annotators were asked to populate a spreadsheet with their answers and add a new row to the spreadsheet each time they found a new dataset within an article. This measure would allow us to count how many datasets were created per article. Once annotators completed their 25 articles, they were returned for review and analysis.

Results

Common Metadata Elements

Based on the combined results of the metadata mapping exercise with DataCite and Dryad, the metadata that was extracted from the 45 NIH Data Sharing Repositories mapped well to MEDLINE's current XML metadata schema for journal articles. The results from these mapping exercises provided a preliminary set of the most minimal level of metadata that could be used to describe datasets in an NIH Data Catalog – with a few modifications and additions [Suppl-9]. This proposed set of metadata for the NIH Data Catalog was mapped to MEDLINE's existing metadata schema in attempts to make the NIH Data Catalog interoperable with PubMed. As mentioned above,

this would require a number of modifications and additions for the datasets to be described appropriately.

The first modification would be to MEDLINE's current Affiliation field. Whereas the current field in PubMed only provides an affiliation for the first author, it was felt that there should be an affiliation for each author of a dataset. Because data is created in different labs at different periods of time, it would be important for a user of the NIH Data Catalog to have contact information for each author of the dataset, therefore increasing the transparency of data-driven research.

The PMID field is another component that would require modification. One of the goals of the NIH Data Catalog is to make datasets discoverable; by assigning a PMID to each dataset in the NIH Data Catalog, a dataset would be searchable both within the NIH Data Catalog *and* within PubMed. Similarly, because it is a goal of BD2K to have datasets linked to their associated articles, each article that used the dataset in the course of its research findings would be added to this field as well.

Related data is an important addition to the metadata schema; data is often created using pre-existing data such as findings from a cohort of the Framingham Heart Study [8]. For the sake of transparency and provenance in scientific research, it is important that this data is accounted for in an NIH Data Catalog record so that users can track the creation of a new dataset back to the pre-existing dataset(s) that were used to create it.

Versioning metadata within the NIH Data Catalog is another addition that could be addressed by allowing each version of a dataset to have its own record that will provide a link back to its original version. This allows for users to view a description of each unique version of a dataset while simultaneously allowing them to trace the research back to the original dataset.

Finally, and perhaps the most challenging aspect of developing minimal metadata elements for the NIH Data Catalog is addressing how datasets will be described both in terms of their subject matter and types of data. For this project, two methods were proposed for dealing with this challenge; one method would be to provide a structured narrative of a dataset, analogous to a structured abstract in a paper. This approach would provide a systematic way for researchers to describe their datasets in detail, and potentially explain how others can reuse their datasets. The second method would achieve a more granular level of description, providing specific data descriptors about the subject of study, type of data, and tools used to create the dataset – similar to the MeSH controlled vocabulary used in MEDLINE to describe journal articles. This area requires

more careful thought and consideration to address the complexities of data description, as developing a controlled vocabulary of all data types for biomedical datasets would be a massive undertaking of great cost and labor. The challenge of describing datasets was further explored in the analysis of articles with "orphaned" datasets in the second phase of the study.

This minimal set of metadata was a first attempt to describe datasets in the NIH Data Catalog. These findings have been presented to the BD2K Working Group and will be discussed at their August 2013 meeting, where decisions will be made as to how datasets will be described.

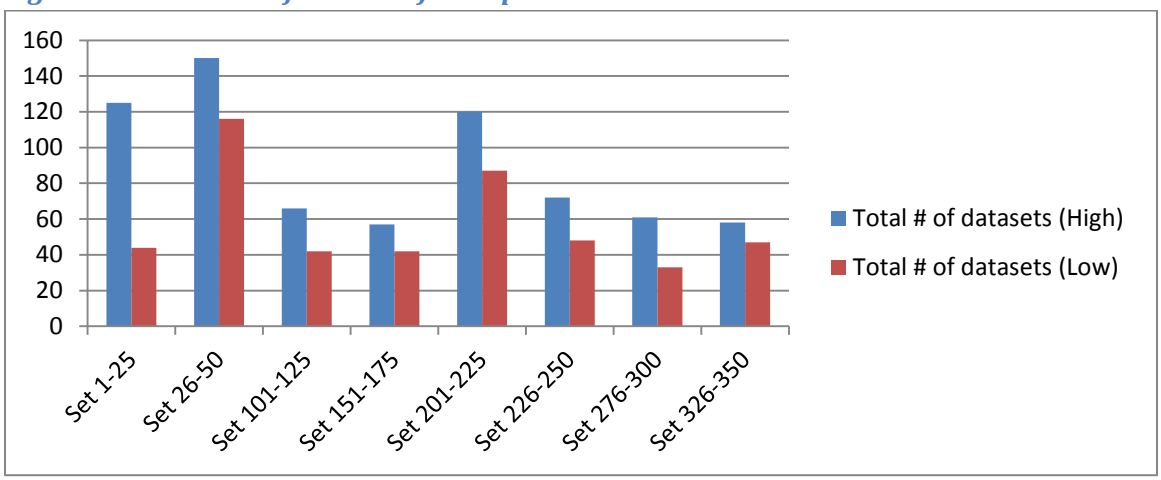
Analysis of "Orphaned" Datasets in NIH funded articles

Responses from seven of the 30 participants did not count more than one dataset per article as instructed. As a result, each set of 25 articles that included annotations only counting one dataset per article were removed from the final statistical analysis. This exclusion left 8 sets of 25 articles for analysis; from these sets we separated annotators who found a higher number of datasets per 25 articles from those annotators who found a lower number of datasets to create a high-low spectrum.

All statistical analyses were performed on both the high and low spectrum annotations to identify the range of difference between the annotators who analyzed the same set of 25 articles. From these 8 sets, we analyzed the percentage of difference between annotators' overall count of datasets; the average number of datasets that were found per article; the percentage of datasets that used live human and non-human animal subjects; and the percentage of datasets created that were considered to be new versus those that used pre-existing data. These measures were completed for each set of 25 articles, and then compiled and evaluated as a whole.

The first goal of the analysis was to determine the percentage of difference between annotators for each set of 25 articles to test for validity [Fig. 3]. The figure below illustrates the total number of datasets found by each annotator for the 8 sets of 25 articles. As mentioned earlier, each set of 25 articles includes a high and low spectrum to differentiate between annotators.

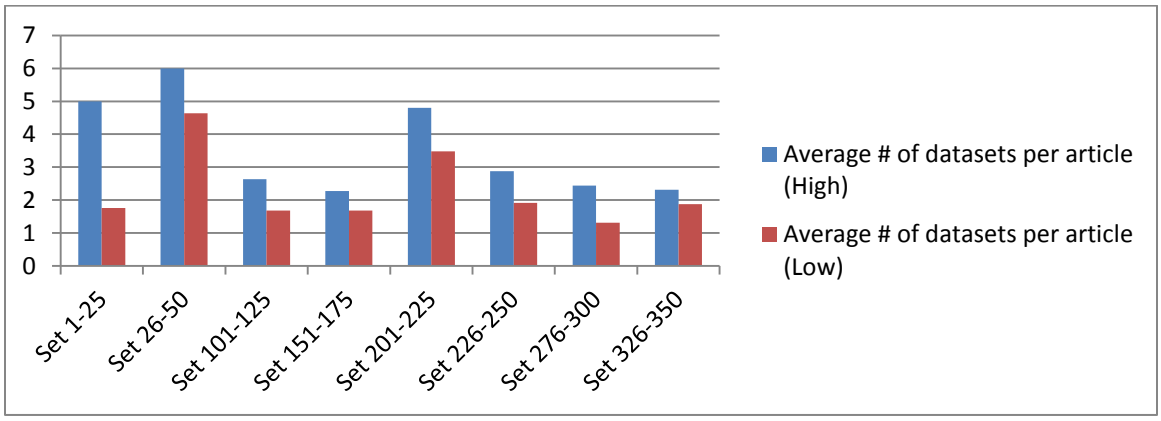
Fig. 3 Total number of datasets found per 25 articles



Based on the difference between the high and low annotators for each set of articles, the total percentage of difference for all 8 sets combined was 43% - a significantly high number with respect to the validation of this exercise. This finding emphasizes the fact that identifying datasets within the biomedical literature is a difficult exercise, despite having skilled annotators performing the analysis.

The next finding was the average amount of datasets found per article by each annotator [Fig. 4]. For this exercise, the total number of datasets found per article by the high spectrum of annotators was 3.54, while the total number of datasets from the low spectrum of annotators came to 2.3 per article. Therefore, the average number of datasets per article found by all the annotators is 2.92 datasets per article.

Fig. 4 Average Number of Datasets per Article



Another finding pertained to how much of the data found within the set of articles used live subjects. This measure was taken to ascertain whether or not the NIH Data Catalog could prioritize

the description of datasets that used live subjects for ethical considerations. This analysis was broken up into two parts: human subjects [Fig. 5] and non-human animal subjects [Fig. 6]. When compiling all of the responses from annotators, the total percentage of datasets from the larger set of 383 articles that used live human or animal subjects was 52% for the high spectrum, and 57% for the low spectrum of annotators. While the separation between the two levels of annotators represents a strong divergence, it is still valuable to understand that over half of the datasets found in NIH funded articles for a given year are created using live subjects.

Fig. 5 Total Number of datasets that used live human subjects

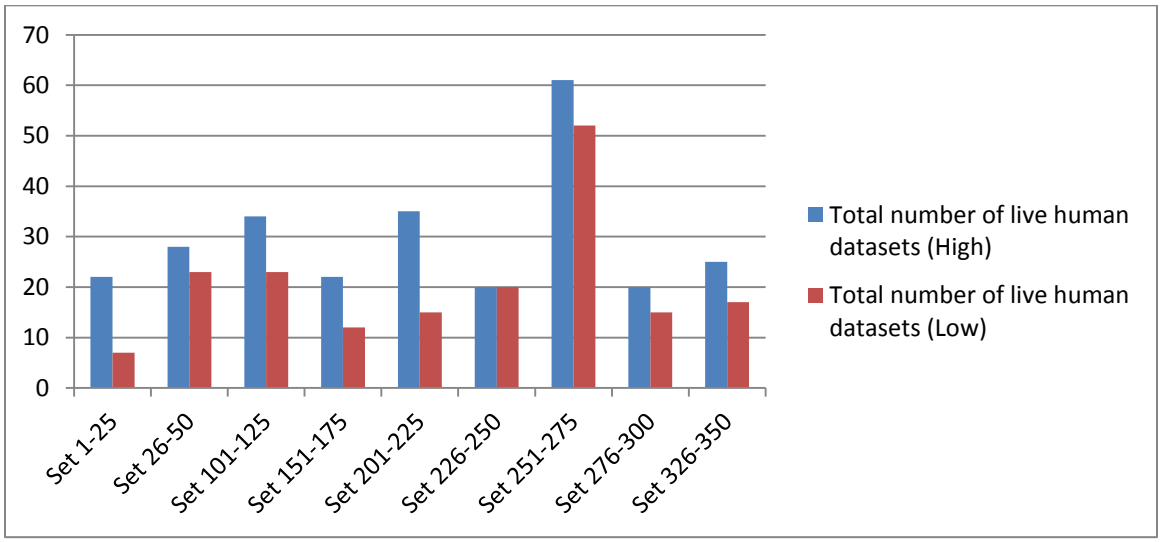
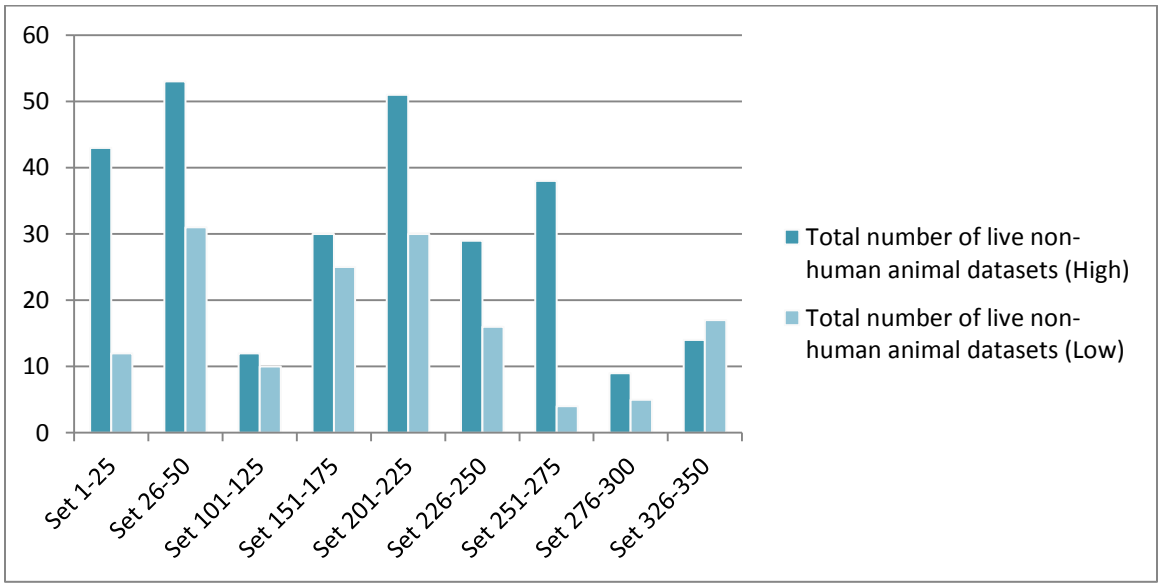


Fig. 6 Total number of datasets that used non-human animal subjects



Finally, the last phase of analysis was completed to determine how much of the data that was created throughout the course of NIH funded research in 2011 was new data, and how much of it was created using pre-existing data. For this phase we counted the percentage of new versus pre-existing data for each set of 25 articles in the study on the high and low spectrum of annotation [Fig. 7], and then calculated the total percentage from all articles [Fig. 8].

Fig. 7 Percentage of new versus pre-existing data created per set of 25 articles

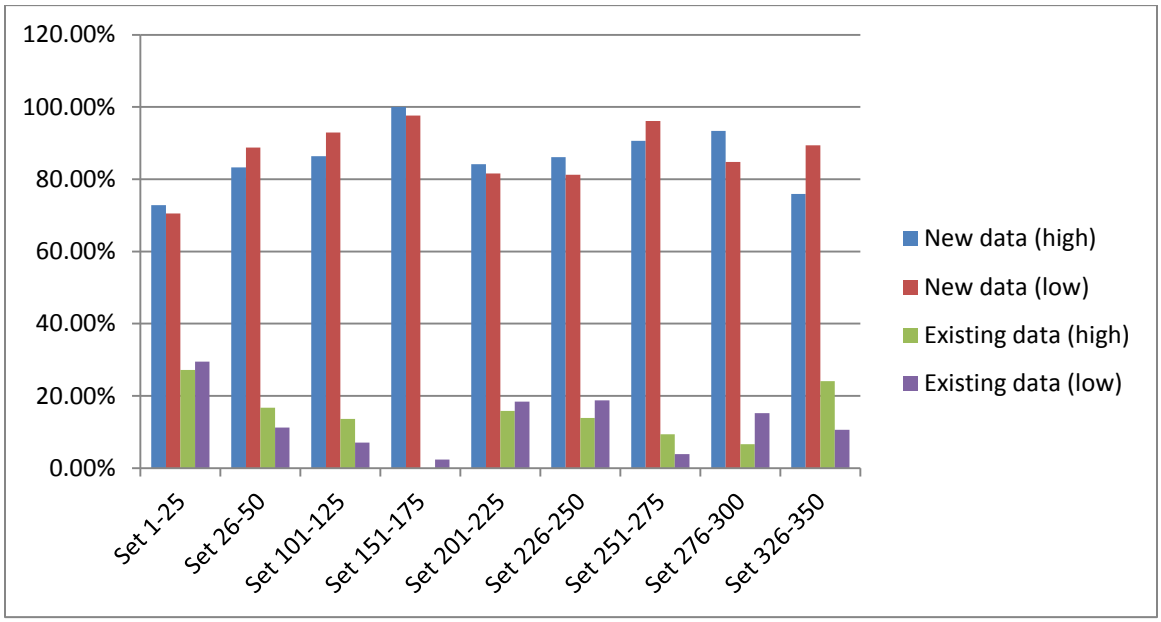
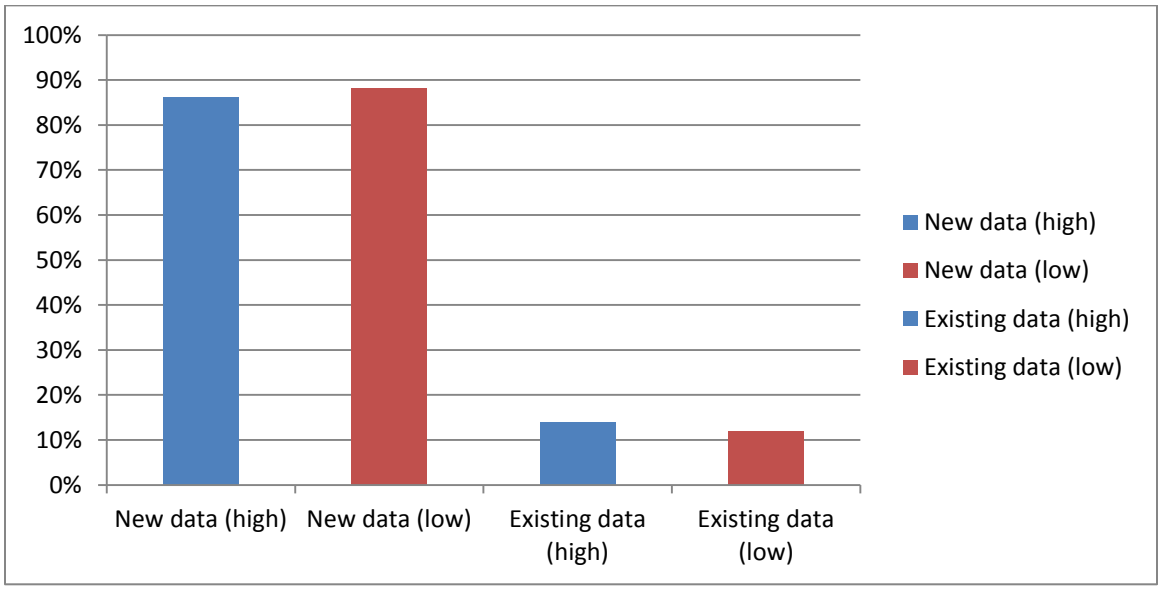


Fig 8. Total percentage of new versus existing data created



The total percentage of new versus existing data was the most significant and statistically valid finding from this study. While all other measures taken throughout this study identified a vast

difference in the way that annotators identified datasets within an article, the range of difference in the case of new versus existing data came to less than 2% (new data = 86% [high] to 88% [low]), (pre-existing data = 14% [high] to 12% [low]). This finding is noteworthy because it provides a strong indication that the majority of data created through NIH funding in a given year is between 86% and 88%; this could represent the types of datasets that would need to be described in the NIH Data Catalog.

Discussion

While the quantification of "orphaned" datasets proved to be difficult, and the validity across annotators' perception of how much data was created within an article was substantially different, this study has raised additional questions that will need to be answered before an NIH Data Catalog can come to fruition.

First, it is important to determine how a dataset will be defined in the NIH Data Catalog. As evidenced by the lack of consistency from annotators with respect to the amount of datasets they found, and the fact that 8 annotators completed the exercise incorrectly by only selecting one dataset per article, it is clear that different people have different perceptions of what constitutes a dataset. Is a dataset all of the data that is created for a study thus only requiring one dataset per article? Does a dataset represent different types of data created at different times throughout a study – therefore producing a number of datasets for each experiment completed within an article? Or is a dataset every individual measurement within a research article, requiring a large number of datasets per article? The NIH Data Catalog must clearly define a dataset to outline expectations of what researchers will be required to share and submit, and develop a common understanding of datasets that is NIH-wide.

The second question that has emerged from this study relates to how data should be described and at what point it should be described. Based on the difficulties from the minimal metadata elements portion of this project, and the uncertainty of annotators assigning data types to datasets despite their strong biomedical backgrounds, it is clear that a strategy must be developed to describe datasets appropriately. The pipeline of data creation can be long and complicated; because it is so complex it can be difficult to ascertain at what point data should be described. In the context of an NIH Data Catalog, is it important to describe a figure showing the analysis of data that can also be viewed in an article? Or is it pertinent to describe the underlying raw numerical data that was used to create that figure? To venture even further down the pipeline, is it useful to

describe an image that was used to derive that raw numerical data? How far down the data creation pipeline will data need to be described? The NIH Data Catalog must decide how it will describe data in a way that will be useful for those researchers, physicians or everyday users who are interested in biomedicine. Because data description was such a difficult issue in both phases of this project, it is recommended that a comprehensive ontology of data types should be developed to accurately represent the types of datasets that are created throughout NIH funded research.

A final question that relates to the assigning value to datasets in the data creation pipeline, is how will the NIH Data Catalog identify the types of data that will be considered useful for others, and describe it appropriately so that it can be reused and validated for accuracy by others. Both phases of this project provided information about how data is described, and the types of data that exist within a scientific paper. While all data can be considered useful, it is clear that some of the data created during biomedical research would be more useful than others. For example, is it more useful for the NIH Data Catalog to describe every single standalone measurement of blood pressure taken from mice *or* new data that was created to cure a particular form of cancer? While it is difficult to compare the two, these are just two examples of the broader data landscape that represent the vast array of data that is created in an NIH funded article or study. To keep the NIH Data Catalog manageable and useable by others, it could describe datasets that show potential for reuse or that point to significant findings that should be available to others to assess its validity and accuracy. How data is described and what types of data are described will be essential for making the NIH Data Catalog a useful resource for biomedical research.

Conclusion

These exploratory studies to analyze NIH funded "orphaned" datasets and identify minimal metadata elements for a planned NIH Data Catalog will inform the next phase of development for BD2K. While the statistical difference between annotators in the "orphaned" datasets phase of the project was very high, the questions raised from annotators and members of BD2K suggest that the creation of an all-encompassing NIH Data Catalog will not be as straightforward as many initially thought. The NIH Data Catalog will require careful consideration in identifying how to describe datasets derived from NIH funded research, and decisions will have to be made as to what data will be selected for description.

With respect to the development of minimal metadata elements to describe datasets in the NIH Data Catalog, strong evidence from a variety of metadata schemas and NIH data repository

metadata indicates that the baseline description of datasets will not vary greatly from traditional journal articles or archival objects. However, just as the data description was a challenge in the "orphaned" dataset phase of the project - more discussion is needed to decide how datasets will be described. The suggestion of a narrative and data descriptor metadata element will be necessary to accurately describe data so that it can be reused and comprehensible to those who will use the NIH Data Catalog.

These findings represent a first look into the data landscape at the NIH. An understanding of the varying types of data that are created throughout the course biomedical research and the knowledge that a substantial amount of new data is created per article in a given year will serve to inform BD2K as they move forward with the creation of an NIH Data Catalog.

References

1. Holdren JP. Increasing Access to the Results of Federally Funded Scientific Research [Internet]. Washington, D.C.: Office of Science and Technology Policy; 2013. Available from: http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

2. Trans-NIH Biomedical Informatics Meeting Committee. NIH Data Sharing Repositories [Internet]. National Center for Biotechnology (US): Bethesda MD; 2012 December [updated 2013 June 19; cited 2013 Aug 2]. Available from: http://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html

3. DataCite Metadata Schema 2.2 XML Schema [Internet]. DataCite – International Data Citation: 2011 July [cited 2013 Aug 4]. Available from: http://schema.datacite.org/meta/kernel-2.2/doc/DataCite-MetadataKernel_v2.2.pdf

4. Dryad Metadata Application Profile Version 3.0 [Internet]. Dryad Development Team: 2010 Aug 2 [cited 2013 Aug 4]. Available from: http://wiki.datadryad.org/wg/dryad/images/8/8b/Dryad3.0.pdf

5. National Library of Medicine. MEDLINE PubMed XML Element Descriptions and their Attributes [Internet]. National Center for Biotechnology (US): Bethesda MD; 2005 Dec 12 [updated 2013 June 19; cited 2013 Aug 4]. Available from: http://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html

6. PubMed Help – Secondary Source ID [Internet]. National Center for Biotechnology (US): Bethesda MD; 2005-. Available from: http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.Secondary_Source_ID_

7. PMC Help [Internet]. Bethesda MD: National Centre for Biotechnology Information (US); 2005-. Available from: http://www.ncbi.nlm.nih.gov/books/NBK3825/#pmchelp.Acknowledgements_ACK

8. ed. Arruda H. Framingham Heart Study [Internet]. National Heart, Lung and Blood Institute, Boston University; 2013 [updated 2013 Aug 1; cited 2013 Aug 4]. Available from: http://www.framinghamheartstudy.org/index.html

Acknowledgements

I would like to thank Jerry Sheehan and Mike Huerta for spearheading this project and for providing a limitless amount of support over the course of this project. I would also like to thank Betsy Humphreys for her guidance and time devoted to this project.

I would also like to thank Lou Knecht and Jim Mork for all of their hard work on the exclusion methodology in PubMed and with the XML in PMC. I would especially like to acknowledge Lou for all of her hard work, recruiting participants for the analysis of "orphaned" datasets, and for her valuable connections inside the NLM.

None of this would be at all possible without the hard work of all the annotators who participated in this study. In no particular order, they are: Preeti Kochar, Helen Ochej, Susan Schmidt, Melissa Yorks, Shari Mohary, Olga Printseva, Janice Ward, Oleg Radionov, Sally Davidson, Jennie Larkin, Peter Lyster, Matt McAuliffe, Greg Farber, Betsy Humphreys, Jerry Sheehan, Mike Huerta, Lou Knecht, Suzy Roy, Swapna Abhyangkar, Olivier Bodenreider, Karen Gutzman, Dina Demner Fushner, Laritza Rodriguez, Sonya Shooshan, Samantha Tate, Matthew Simpson, Tracy Edinger, Olubumi Akiwumi, Marry Ann Hantakas, Corinn Sinnott.

Finally, I would like to thank Library Operations Joyce Backus and Dianne Babski and NLM Leadership Dr. Lindberg and Betsy Humphreys for giving me the opportunity to work on these wonderful projects and sponsoring this excellent program.

Appendices

Supplementary Files

See:

Suppl-1_MetadataElements_NIHRepositories.xlsx

Suppl-2_metadatataxonomy.pdf

Suppl-3_datacite_metadata_mapping.pdf

Suppl-4_dryad_metadata_mapping.pdf

Suppl-5_MEDLINE_metadata_mapping.pdf

Suppl-6_PubMed_exclusions.pdf

Suppl-7_PMC_acknowledgements_exclusions.pdf

Suppl-8_XML_exclusions.pdf

Suppl-9_minimalmetadata_results.pdf

Databank phrases for Compound Word Dictionary.txt

Raw Data

See:

Bigdata_randomsample_1 through 385.xlsx (30 files in total)

Rawdata_orphaneddatasets_dataanalysis.xslx