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

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