A Cross-Institutional Data Discovery Collaboration: Indexing Institutional Research Data

Kevin B. Read, MLIS, MAS[1], Nicole Contaxis, MLIS[1], Ian Lamb[1], Catherine Larson[1], MLIS, Alisa Surkis, PhD, MLS[1]
[1]Health Sciences Library – NYU Langone Health, New York, NY;

Introduction

While there are many biomedical data repositories that are designed to help researchers locate specific kinds of data (e.g., GenBank, dbGaP), many research datasets – even data resulting from NIH and NSF funded research – are not shared in these repositories[1]. The NIH has devoted considerable resources towards the development of systems to aggregate biomedical research data across multiple repositories to improve discoverability of that data[2,3], but these systems do not address the discoverability of datasets that have not already been shared through these repositories. The Data Catalog developed at NYU Langone Health is an open source tool for indexing datasets generated by researchers within an institution[4-6]. This low-barrier approach to data sharing requires only that researchers provide a description of their dataset. The data is made discoverable, while the researchers retain control of the data itself, thus providing an avenue for researchers who are not ready to deposit their datasets in a repository to begin the process of data sharing. Subsequently five other health sciences libraries situated in academic medical centers implemented their own institutional instances of the Data Catalog. The six libraries (from this point on referred to as the Data Catalog Collaboration) meet regularly to share challenges in indexing data from different research domains and meeting researchers' needs in varied institutional environments.

Methods

NYU Langone Health developed a data catalog to meet institutional needs around data discovery and collaboration. Metadata was developed with institutional needs in mind, while taking advantage of the project lead's participation in the NIH bioCADDIE initiative to ensure that it was also in alignment with national efforts around data discovery. Collaborations were established at the earliest stages of catalog development with a number of NYU stakeholders with an interest in data discovery and data sharing initiatives.

Data Catalog metadata was developed to maximize the discoverability of datasets across a range of research disciplines while being sufficiently lightweight to not require unrealistic levels of input by researchers or overly time-consuming curation by indexers. The metadata focuses on describing: the types of data being indexed, characteristics of the data (e.g., format, size, equipment used to collect data), funding awards associated with the data, access instructions and restrictions for the data, and contextual information about the dataset (e.g., associated publications)[6].

To facilitate broader implementation of the Data Catalog across institutions, the code was made open source. The goal of this effort was twofold: 1) increase the discoverability of a larger number of otherwise difficult or impossible to find datasets, and 2) utilize cross-institutional input on Data Catalog metadata, functionality and usability to inform catalog refinements such that the catalog would meet the needs of a broader range of institutions. The code was made available on GitHub and modified to allow for easy implementation and other institutional branding. The availability of the data catalog code was then marketed at professional conferences and through the National Library of Medicine (NLM).

Finally, to align Data Catalog metadata with incoming national data discovery initiatives, the data catalog metadata schema was mapped to the DATS[3] schema.

Results

The Data Catalog has been used to index many different types of data. Specifically, it indexes data pull requests from the electronic health record (EHR) for research purposes, and clinical trial, epidemiological, longitudinal, and geospatial population health datasets. Since May 2015 the Data Catalog has 47,350 page views, with 10,148 unique visitors.

A number of stakeholders at NYU have contributed to the development and continued expansion of the Data Catalog, including our Clinical and Translational Science Institute, Department of Population Health and the institution's clinical data management core. Researcher challenges identified through the Data Catalog's outreach efforts have also led to the development of a workgroup with institutional data policy stakeholders to address concerns researchers have expressed regarding their lack of knowledge of institutional data policies and data use agreements.

Since outreach began to identify external collaborators looking to implement the Data Catalog, the University of Pittsburgh, the University of Maryland, Baltimore, University of North Carolina – Chapel Hill, University of Virginia, and Duke University have installed the Data Catalog at their respective institutions with funding from the NLM. This collaboration has led to ongoing discussions concerning the refinement of metadata to accommodate new research disciplines at other institutions (e.g., basic science datasets), and improvement of the code as more developers contribute to it.

After mapping Data Catalog metadata to DATS, the project lead has begun preliminary discussions with the NLM metadata team about their data discovery initiatives and the metadata mapping[6] has been shared with them.

Discussion

The Data Catalog is an effective means of indexing institutional research datasets that are otherwise not easily discoverable, while providing a low-barrier way to introduce researchers to data sharing. At NYU Langone Health, the Data Catalog has afforded the opportunity to index datasets that are not discoverable elsewhere, establish institutional partnerships to improve data policy workflows, and eliminate redundancies for data requests. As part of a larger collaboration, the Data Catalog is helping to improve data discovery metadata and usability, and streamline the Data Catalog code for future institutional implementations. As more institutions adopt the Data Catalog locally, it can provide an avenue to uncover institutional research data and inform new developments in data discovery across biomedical research. A future potential benefit of the Data Catalog is that, with metadata aligned with national data discovery efforts, it is poised to allow the sharing of the information about these datasets at a national level, thus allowing the community to decide which datasets are of greatest importance based on access requests. This information on dataset usage can serve to inform NIH priorities for future resource allocation.

References

1. Read KB, Sheehan JR, Huerta MF, et al. Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study. *PLOS ONE.* 2015;10(7):e0132735.
2. Ohno-Machado L, Sansone S-A, Alter G, et al. Finding useful data across multiple biomedical data repositories using DataMed. *Nature Genetics.* 2017;49:816.
3. Sansone S-A, Gonzalez-Beltran A, Rocca-Serra P, et al. DATS, the data tag suite to enable discoverability of datasets. *Scientific Data.* 2017;4:170059.
4. Read KB, Contaxis N, Lamb I, Larson C, Surkis A. NYU Data Catalog. 2015; https://datacatalog.med.nyu.edu/.
5. Lamb I, Larson C. Shining a Light on Scientific Data: Building a Data Catalog to Foster Data Sharing and Reuse. *Code {4} Lib.* 2016;32.
6. Read KB, Contaxis N, Lamb I, Larson C, Surkis A. NYU Data Catalog. 2017; https://osf.io/vg7rn/.