IJDC | *General Article*

Promoting Data Reuse and Collaboration at an Academic Medical Center

Kevin Read, Jessica Athens, Ian Lamb, Joey Nicholson,
Sushan Chin, Junchuan Xu, Neil Rambo and Alisa Surkis
NYU School of Medicine

Abstract

A need was identified by the department of population health (DPH) for an academic medical center to facilitate research using large, externally funded datasets. Barriers identified included difficulty in accessing and working with the datasets, and a lack of knowledge about institutional licenses. A need to facilitate sharing and reuse of datasets generated by researchers at the institution (internal datasets) was also recognized. The library partnered with a researcher in the DPH to create a catalog of external datasets, which provided detailed metadata and access instructions. The catalog listed researchers at the medical center and the main campus with expertise in using these external datasets in order to facilitate research and cross-campus collaboration. Data description standards were reviewed to create a set of metadata to facilitate access to both externally generated datasets, as well as the internally generated datasets that would constitute the next phase of development of the catalog. Interviews with a range of investigators at the institution identified DPH researchers as most interested in data sharing, therefore targeted outreach to this group was undertaken. Initial outreach resulted in additional external datasets being described, new local experts volunteering, proposals for additional functionality, and interest from researchers in inclusion of their internal datasets in the catalog. Despite limited outreach, the catalog has had ~250 unique page views in the three months since it went live. The establishment of the catalog also led to partnerships with the medical center's data management core and the main university library. The Data Catalog in its present state serves a direct user need from the Department of Population Health to describe large, externally funded datasets. The library will use this initial strong community of users to expand the catalog and include internally generated research datasets. Future expansion plans will include working with DataCore and the main university library.

Received 08 January 2015 | *Accepted* 10 February 2015

Correspondence should be addressed to Kevin Read, 577 First Avenue, 2nd Floor, New York, NY, 10016. Email: kevin.read@nyumc.org

An earlier version of this paper was presented at the 10th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

A need was identified by the Department of Population Health at NYU School of Medicine to facilitate research using large, externally funded datasets (e.g. U.S. Census data, public health survey data). The barriers identified included a lack of knowledge about the types of datasets that are available, a lack of knowledge about institutional licenses, difficulty in accessing datasets and finally, difficulty in working with these external datasets. The Health Sciences Library subsequently partnered with the Department of Population Health to create a catalogue of these datasets to address the identified barriers. This Data Catalogue provided detailed metadata, access instructions, and information about researchers at the institution with expertise in a given dataset. From the outset, library plans included expansion of the Catalogue to include datasets generated by the institution's researchers (internal datasets), with the goals of facilitating data sharing and reuse; and strengthening collaboration within the medical center, between the medical center and the university's other campuses, and with researchers beyond the institution. This paper will outline the strategy used to create the initial Catalogue for external datasets, and describe next steps for expanding the Catalogue to include internal datasets.

Methods

Data Interviews

To inform the creation and implementation of the Data Catalogue, a series of interviews were conducted with researchers to better understand their research data challenges and needs. Researchers at NYU School of Medicine with active grant funding were identified through the institution's grant management system. A subset of researchers was then selected to reflect a range of research, from the basic sciences to clinical to population health. Interview questions focused on eliciting an overview of the interviewees' research and a description of how they collect, organize, store and share their data. Researchers (n=30) were interviewed until no new insights concerning the Catalogue or their data needs emerged.

Metadata

Data description standards were reviewed to create a set of metadata to facilitate access to and use of both externally and internally generated datasets. Metadata schemas were selected from the U.S. National Institutes of Health (NIH) Big Data to Knowledge (BD2K) minimal metadata elements, DataCite[1], Dryad[2], and W3C Data Catalog Vocabulary[3]. The schemas were analyzed and compared; specific metadata elements were selected based on their relevance and applicability to the external datasets within the Data Catalogue, and their future use in describing internally generated datasets. This work was completed in consultation with the Department of Population Health to

1 DataCite Metadata Schema Repository: http://schema.datacite.org/
2 Dryad Metadata Application Profile (Schema): http://wiki.datadryad.org/Metadata_Profile
3 W3C Data Catalogue Vocabulary (DCAT): http://www.w3.org/TR/vocab-dcat/

 doi:10.2218/ijdc.v10i1.366

identify metadata descriptors needed to provide a clear description of the datasets that fell outside of the more general-purpose schemas examined. Metadata fields for the datasets were populated by the librarians, in consultation with a population health researcher.

Recruiting Local Experts

The library searched publications and grants to identify researchers at the institution who had experience working with the external datasets indexed in the Catalogue. A collaborator from the Department of Population Health then contacted those researchers to request their participation as local experts. Researchers were identified both at the medical center and at the university's main campus in order to both increase the pool of experts and facilitate cross-campus collaboration.

Building the Data Catalogue: Technical Specifications

Apache Solr, a fast key and value data store with advanced search capabilities, was used to store and search the Catalogue metadata. That metadata is inputted into spreadsheets and then uploaded into Apache Solr. The front-end of the Data Catalogue was written using Javascript, which makes use of Backbone.js, jQuery and Underscore.js. PHP script was also used to output the data in a format more easily accessible from Javascript.

Initial Results

The creation of the Data Catalogue has provided a tool for researchers to find information about available datasets, how to access them and who is available locally to guide them. The homepage of the Data Catalogue provides a simple interface where researchers can locate datasets either by using the search bar, browsing and refining results with the filters located along the left side of the page, or using a combination of searching and filtering. The filters on the left side include the Subject Domain, Timeframe, and Access Restrictions of the datasets (see Figure 1). For the Subject Domain, the library chose to use Medical Subject Headings (MeSH)[4] wherever possible to maintain a controlled vocabulary of terms. However, when it was felt that using MeSH would detract from usability, we opted to maintain usability above all else. For example, the MeSH term for "cancer" is "neoplasms"; because users would be more likely to search for the word "cancer" due to its prevalence and familiarity rather than "neoplasms", the library chose to include the more common term. The Timeframe metadata field was chosen for external datasets because many of those relevant to population health research are longitudinal; Timeframe provides an avenue to account for when a dataset was first created, and the subsequent years data was collected. Finally, Access Restrictions provides users of the Catalogue with information about whether an external dataset is free to use, requires registration or an application, is available via an institutional license, or requires a fee in order to access it.

4 Medical Subject Headings (MeSH): http://www.nlm.nih.gov/pubs/factsheets/mesh.html

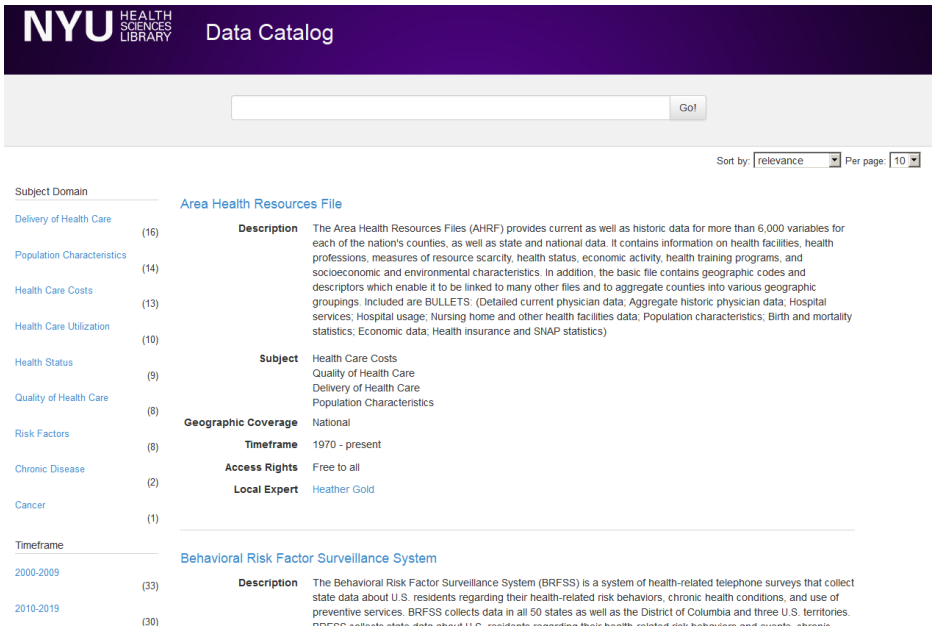

Figure 1. Image of the NYU Health Sciences Library Data Catalogue with search bar and filters.

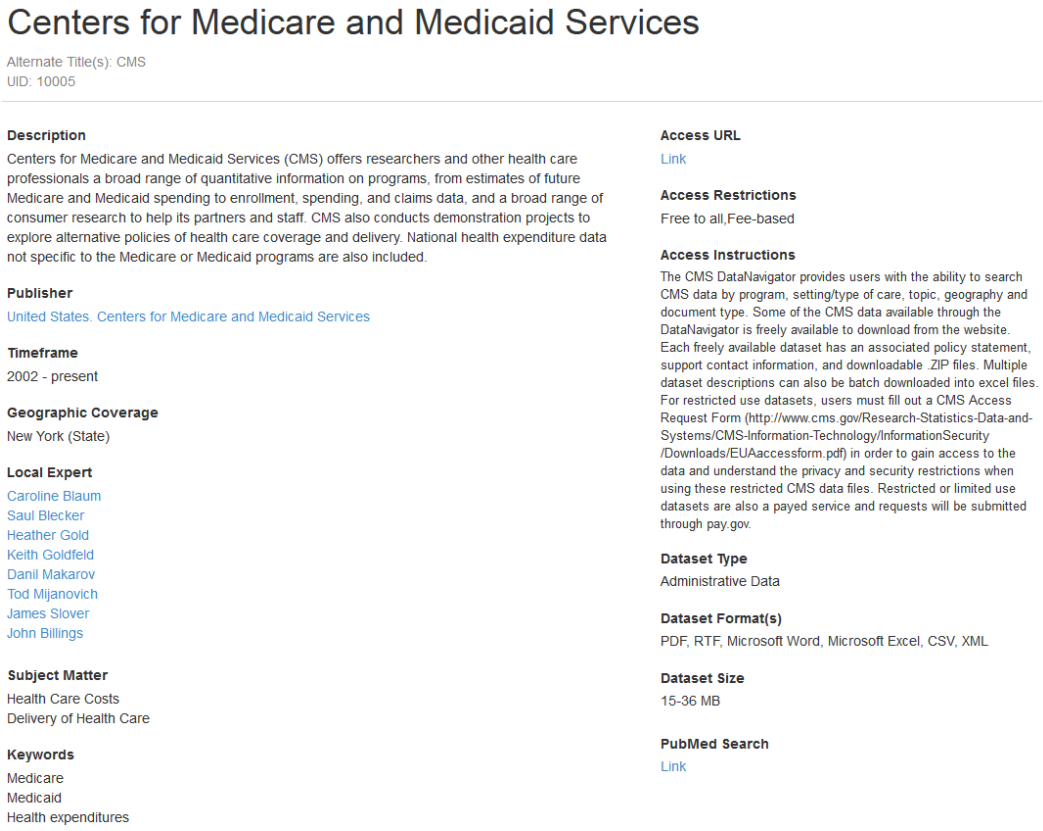

Figure 2. Full metadata record of a dataset within the NYU Health Sciences Library Data Catalogue.

 doi:10.2218/ijdc.v10i1.366

The search result view for a dataset record provides information describing the dataset, the subject(s) of the dataset, the geographic coverage of the dataset, timeframe, access rights and local expert(s) of that dataset (see Figure 1). Clicking on the local expert will take the user directly to an institutional researcher profile that includes contact information. The full metadata record (see Figure 2) provides all the information from the search result view, as well as more detailed information, including key information about how to access the dataset, what software is required, what applications or registrations must be completed, a direct link to the website where the dataset lives, a link to a PubMed search of literature pertaining to the dataset, information about the type of dataset and its format, and links to the publisher of the dataset. The library has also included a "Related datasets" metadata element to link datasets with explicit relationships. These relationships can be of many different types; linkage datasets exist between some datasets, there are datasets created by the same publisher, some datasets have national (USA) and local (New York City) versions. The Catalogue also provides a context note metadata field to describe these dataset relationships.

Next Steps

With an infrastructure already in place where researchers can access useful information about relevant research datasets, the next step in this phased implementation is to begin describing internally generated research datasets. Based on the data interviews, it was determined that clinical and population health researchers were the most receptive to both sharing data and using shared data, so initial outreach for internal datasets to include in the Catalogue will be to researchers in these areas.

Metadata

The metadata currently in use are those elements that are applicable to external datasets. Some of these elements are primarily or exclusively for use with external datasets, such as publisher or local experts. In developing our metadata schema we also included elements that would be primarily used for internal datasets, such as award, funder or associated citation. As much as possible, we tried to define elements so that they would work for both internal and external datasets. However, there were cases for which we created two separate types of metadata. For example, where the Catalogue metadata for an external dataset provides a PubMed search that links to articles that *used* the dataset, internal datasets will link to articles that *created or collected* the dataset.

A second consideration around metadata is that some elements are relevant for the current external datasets, as well as for internal population health and clinical datasets, but will not be relevant for internal basic science experiments. These include elements such as Geographic Coverage and Timeframe. As we expand to other dataset types, we will not eliminate current metadata elements, but will consider the addition of limited additional elements relevant to other broad areas, as warranted by the need to maximize discoverability.

Promotion

The library's approach to promoting the Data Catalogue and generating interest in describing internal datasets includes both broad and targeted outreach. Broad outreach will be accomplished through email blasts to the entire research community. Targeted outreach will include presentations for relevant departments and institutes, such as the Department of Population Health at the medical center and the Global Institute of Public Health at the main university campus. In addition, the Catalogue will be linked to and gain visibility through its association with two cross-cutting entities. The Catalogue is also part of both the Clinical and Translational Science Institute, whose mission is to promote collaboration and data sharing as well as bridge the gaps across different schools within the academic institution, and the medical center's DataCore, whose role centers around clinical research data management, maintaining access to the clinical data warehouse, and providing data curation possibilities for external and internal datasets. These additional stakeholders increase the visibility of the Data Catalogue, and provide multiple avenues for identifying researchers within the institution who want to describe their data.

Issues and Strategy

Mixing Internal and External: Strategy over Purity

Building a Data Catalogue that mixes internal and external datasets raises a number of issues. The metadata schema has to accommodate elements that have little or no relevance for one or the other type of dataset. The internal dataset entries provide information that will often be unavailable elsewhere, while the external datasets are catalogued elsewhere (e.g. ICPSR). Acknowledging the differences between how the two types of datasets are treated within the Catalogue and how they are used, the question arises as to why create a mixed Catalogue, rather than two separate Catalogues, each with clear functionality and clean metadata. The answer is strategy – our primary goal for building this Catalogue is to create a tool that will be broadly used within our institution.

The external datasets that comprise the first version of the Catalogue are of broad interest to researchers – primarily population health researchers, but clinical researchers as well. Any given internal dataset is likely to be of interest to a much smaller group of researchers. By initially adding internal datasets that fall within a similar domain to those external datasets that already exist in the Catalogue, we have set the stage for the serendipitous discovery of relevant internal datasets by the broad range of users for which the initial external datasets are useful. In this way, we seek to build and foster a strong community of users to maximize dataset reuse.

Promotion

In addition to the outreach efforts described above, the library will increase the visibility of the Data Catalogue and the datasets described within by making the data discoverable through multiple systems. In collaboration with the university's main library, the Data Catalogue records will be harvested for the main library catalogue via an API. The library is currently in the process of implementing Profiles Research Networking

 doi:10.2218/ijdc.v10i1.366

Software[5], which is designed to describe research output from individual researchers. A plug-in will be developed for Profiles such that the datasets from the Data Catalogue will be listed as research outputs for each researcher at the medical center who has provided a description of their dataset in the Catalogue.

Scalability

As the library begins to describe internal datasets, scalability will increasingly be an issue. One factor will be the amount of curation time that will be required to describe each dataset. Also, there may be researchers who want a large number of datasets to be described. Plans include building an interface that will allow researchers to upload descriptions of their own datasets where they will then enter a queue to be curated by the library. However, the extent to which the dataset description can be outsourced to the researchers themselves will be limited less by technical issues than by the degree of buy-in that can be generated on the part of the researcher community. Ultimately, the issue of scalability goes hand in hand with promotion, as demonstrated use and utility of the Catalogue will be critical to building a case for funding for more personnel being allocated to the Catalogue.

Maintenance

Another issue that will have to be carefully considered will be the maintenance of the Data Catalogue moving forward. Researchers leave the institution on a regular basis (e.g. new research opportunity, retirement). The library has experience tracking researchers through their work developing a comprehensive faculty bibliography (Vieira et al., 2014), which makes use of a data feed from the institution's enterprise data warehouse to track researcher status. Another method to track researchers will be through the implementation of ORCID[6]; by Spring 2015 all researchers at the academic medical center will have an ORCID, which will provide the library with another opportunity to track a researcher's location. The Catalogue will similarly make use of this information, and establish policies and workflow around the exit of researchers from the institution, recognizing that issues around data ownership and location may mean that the workflow may be different in different cases. In order to accommodate these types of changes, the Catalogue will have an unpublish option so that records can be maintained in an archival state.

Expansion

While the initial step toward adding internal datasets will involve the direct outreach to researchers working in population health, there are two other means of expansion that have yet to be discussed. The first is working with librarians at the main university to ingest datasets from that campus. Because population health research is at the intersection of biomedical research and social sciences research, this would work well with our plan to slowly build out the scope of the Catalogue. Secondly, the DataCore has future plans to provide storage and access to analysis datasets from researchers, and the Catalogue would provide a discovery element to those datasets.

5 Profiles Research Networking Software: http://profiles.catalyst.harvard.edu/
6 ORCID: http://orcid.org/

Conclusion

The Data Catalogue in its present state serves a direct user need from the Department of Population Health to describe large, externally funded datasets. The library will use this initial strong community of users to expand the Catalogue and include internally generated research datasets. Future expansion plans will include working with DataCore and the main university library to expand and streamline the intake of datasets generated by institutional researchers across all campuses.

References

Vieira, D., McGowan, R., McCrillis, A., Lamb, I., Larson, C., Bakker, T., Spore, S. (2014). The faculty bibliography project at the NYU School of Medicine. *Journal of Librarianship and Scholarly Communication, 2*(3), eP1161. doi:10.7710/2162-3309.1161