Summary of Responses to the Request for Information (RFI): Input on Development of a NIH Data Catalog (NOT-HG-13-011)

Key Dates
Release Date: June 6, 2013
Response Date: June 25, 2013

Purpose
This Request for Information (RFI) solicits comments and ideas for the development and implementation of an NIH Data Catalog as part of the overall Big Data to Knowledge (BD2K) Initiative.

Figure 1: Wordle representation of text from all Data Catalog RFI responses, excluding the words "data" and "catalog".

Demographics of RFI Responses:

This RFI received 62 responses. Of these, 2/3 (N = 41) came from academic respondents, 9 from commercial interests (companies active in knowledge management), 7 from not-for-profit institutions, and 5 from other groups (including publishers and government organizations). Some of these responses represented the thoughts of a single individual, while others represented organized groups with interests and expertise in the area. The majority of the respondents were from the United States. There was broad geographical representation in the responses representing 26 states and three countries. Respondents were well distributed across the United States, with 19 responses from the central United States, 17 responses from the northeastern states, 13 responses from the western states, and 7 from the southeastern United States.

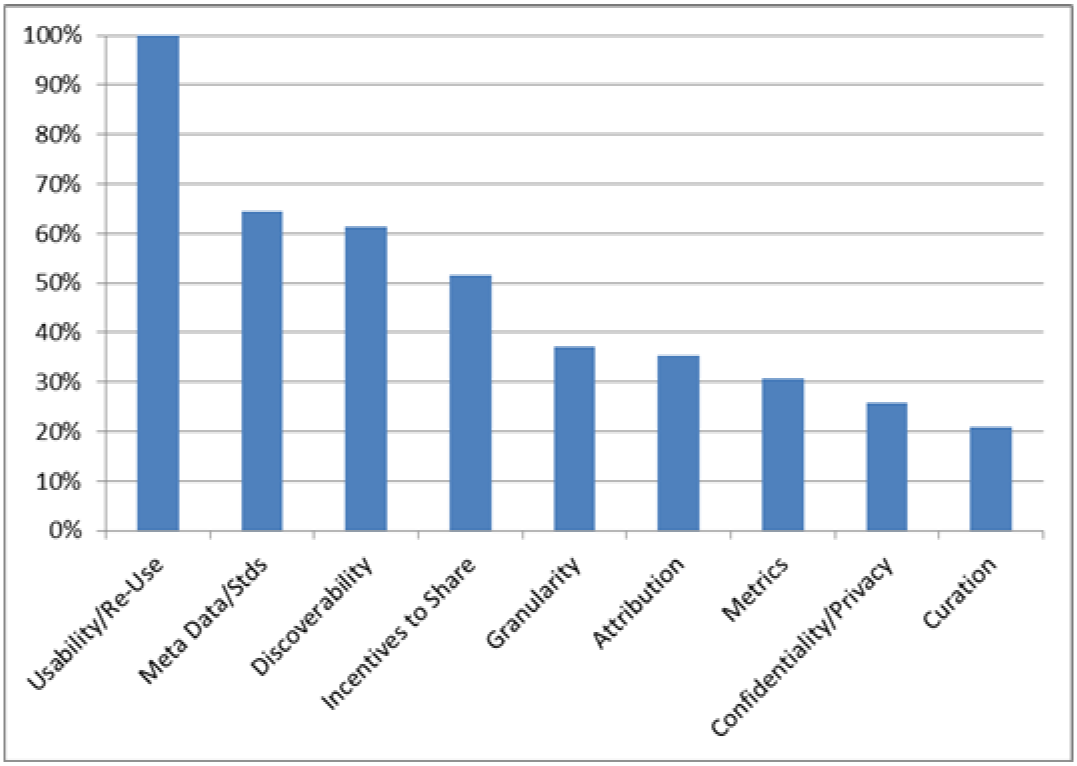

Figure 2. Percentage of responses that addressed each of the nine defined topics.

RFI Response Topics: Access and Discoverability

Key points that came up in several contexts was that accessibility and discoverability of data depends on having a sufficiently robust set of metadata with a controlled vocabulary so that the data entries can have some basic standards. Once a central index (catalog) is then established it will be easier to identify subsets of data. Access should follow so long as universal rules for data sharing are in place to assure that the data are indeed available. Having dataset metadata combined with microdata markup language would also enhance findability. Usability will flow from how easy it is to find and access the data. Having the tools and algorithms to easily search the catalog will also be important to easy of discovery and access.

RFI Response Topics: Usability and Re-Use

The general issues surrounding the concept of reuse generated energetic responses from community. The concept of reuse is fairly broad here: sometimes responses refer to specific issues relating to the overall goal of achieving reuse of data; sometimes responses suggest we need to reuse existing standards or methods. A key issue that has to be dealt with is how to incentivize the community to contribute to the NIH Guide. Its (re)use will not be vibrant unless there is one stop shopping. Key to this is findability and accessibility. Of course, the catalog is only as useful as the underlying repositories, so NIH needs to promote publication of data on the Internet. Technically the Guide will be useful if it can be expanded hierarchically to a broad range of research domains. Some responses suggested particular technologies, for example dealing with URL, ORCID ID, handling PHI, and citations to similar experiments. Other responses emphasized the importance of learning from other programs such as caBIG. What is striking is the long list of projects and sites out there that have addressed some or all of the issues surrounding universal catalogue(s) of resources, data and otherwise: data.gov, clinicaltrials.gov, NIF, biositemaps, BRO, Eagle-i, NCBI, dbGaP, GEO, NIAID PRiME, PubChem, RC3, Databib, ViVo, Monarch, DataDryad, DatagaStar, CTSA Connect.

RFI Response Topics: Attribution

Attribution requires a common and universal identifier, such as a DOI that would allow for citation of the data. That also would enhance the ability to find the data. By more easily tying a publication with its data, one could rapidly begin work on subsequent experiments without the delay in tracking down the data and determining if it is usable.

RFI Response Topics: Metadata and Standards, and Curation

Virtually all respondents made reference to the importance of good metadata in enabling users to discover and make use of data sets listed in the catalog. Several commenters specified particular metadata elements they would want to see in a data catalog, with a few calling for a tiered system of general elements for all data sets (title, author, date, persistent ID) and domain-specific data elements with more detail (type of instrument, type of data, etc.). Additional suggestions were made for metadata to include indicators of methodology and provenance. Respondents generally supported the use of standard metadata and identified particular candidates, including NLM and NCI vocabularies and standards, biositemaps, biomedical resources ontology, ICPSR's metadata, and Data.gov metadata. Respondents expressed differing opinions on the value of an abstract: some respondents view abstracts as an unnecessary addition to structured metadata, but others view them as an important element of a data catalog entry to support natural language processing and improve the tagging of entries. Several commenters indicated that investigators would not do a good job of annotating their own data and suggested paid curators or better tools for data curation. While one commenter supported community-based curation others highlighted NLM's experience in curating and cataloging data. One commenter proposed links to related journal publications as a means of improving data quality.

RFI Response Topics: Incentives to Share

Of the 21 RFI responses that addressed incentives (sometimes inseparable from barriers), there were several common themes. Data sharing should be a requirement associated with NIH funding, and the NIH should support tools to make this as easy as possible for the PI. There was also strong support for having data sharing being accounted for in professional incentives (peer review of grant applications, promotion and tenure), and it was also noted that in order for this to be effective there needs to be a uniform manner to cite data, to help provide metrics for research impact of data sharing. It was noted that expectations/mandate for data sharing should be phased in gradually and should allow for different tiers of data exposure (though all tiers could have the basic metadata made available through the data catalog). Respondents noted the importance of common, consistent metadata and many noted having a tiered approach, with common metadata being supplemented by field-specific metadata being a useful approach.   The challenge of ensuring that all data have permanent unique identifiers DOIs or some other method) was seen as essential for effective data citations. Some additional interesting thoughts included concerns about the social aspects of data ownership, concern about QA/QC, engagement of the community, and consideration of incentives for both the data producer and for the secondary data users. It was noted that simply requiring data catalog entries without addressing the incentives and barriers may result in an ineffective system with unusable data. It was also emphasized that the economics of the cost of building and maintaining the infrastructure to support the data catalog and data citations should be considered and not underfunded.

RFI Response Topics: Granularity

The granularity needed depends upon the purposes for which the catalog is to be used. Several responses indicated the need for each catalog entry to include information such as: authors, where the data are located, whether and how the data are available and what is the nature of the data (or as one put it: who, what, where, when and how). Most responses focused on the data descriptors with the need for consistent descriptors a common theme and a few suggestions for a nested hierarchy of data descriptors (e.g., image > MRI > spectroscopic).

RFI response Topic: Metrics

Of the 20 RFI responses that addressed metrics in some fashion, there was consensus that the Data Catalog should have a set of core metrics to help determine its usefulness. These include number of entries, numbers of researchers and institutions represented, geographic locations of searchers, number of accesses, number of searches, how often people follow a link out, and number of data citations in the literature. A number of the proposed metrics relate to access and download of the datasets themselves, which may be outside the purview of the Data

Catalog, but would be a useful measure of adoption and usefulness of the data catalog. There were also some longer-range or more difficult metrics proposed which may be useful: comparison to other data reuse sources (Data Citation Index or CrossRef), percentage of NIH studies for which data are cataloged, whether data access and reuse are higher for data in the data catalog, and user feedback.

RFI Response Topics: Confidentiality and Privacy

It is recognized that the Data Catalog will require attention to privacy and security of data to prevent leaks of sensitive information. Privacy concerns create obstacles to full data sharing and limit the granularity of data that can be freely distributed. The Data Catalog will provide useful metadata and summarized information on these controlled access data. However, care must be taken to prevent privacy leaks, in file names, participant identifiers, and other features that may be chose as annotations. Finally, any privacy or PHI restrictions in the data should be clearly described in the catalog entries.

Appendix I: RFI Text

Request for Information (RFI): Input on Development of a NIH Data Catalog
Notice Number: NOT-HG-13-011

Key Dates

Release Date: June 6, 2013
Response Date: June 25, 2013

Purpose

This Request for Information (RFI) is to solicit comments and ideas for the development and implementation of an NIH Data Catalog as part of the overall Big Data to Knowledge (BD2K) Initiative.

Background

Biomedical research is becoming more data-intensive as researchers are generating and using increasingly large, complex, and diverse datasets. This era of 'Big Data' in biomedical research taxes the ability of many researchers to release, locate, analyze, and interact with these data and associated software due to the lack of tools, accessibility, and training. In response to these new challenges in biomedical research, and in response to the recommendations of the Data and Informatics Working Group (DIWG) of the Advisory Committee to the NIH Director (http://acd.od.nih.gov/diwg.htm), NIH has launched the trans-NIH Big Data to Knowledge (BD2K) Initiative.

The long-term goal of the BD2K Initiative is to support advances in data science, other quantitative sciences, policy, and training that are needed for the effective use of Big Data in biomedical research. (The term "biomedical" is used here in the broadest sense to include biological, biomedical, behavioral, social, environmental, and clinical studies that relate to understanding health and disease). The term 'Big Data' refers to datasets that are increasingly larger, more complex, and which exceed the abilities of currently used approaches to manage and analyze. "Big Data" is also meant to capture the opportunities and address the challenges facing all biomedical researchers in accessing, managing, analyzing and integrating large datasets of diverse data types. Such data types may include imaging, phenotypic, molecular (including –omics), clinical, environmental, behavioral, and many other types of biological and biomedical data. "Big Data" also includes data generated for other purposes (e.g. social media, search histories, cell phone data) when they are repurposed and applied to address health research questions. Biomedical Big Data primarily emanate from three sources: (1) a small number of groups that produce very large amounts of data, usually as part of projects specifically funded to produce important resources for use by the research community at large, or large collections of electronic health records; (2) individual investigators who produce large datasets for their own project, but which might be broadly useful to the research community at-large; (3) an even greater number of investigators who each produce small datasets whose value can be amplified by aggregating or integrating them with other data.

One of the DIWG recommendations was to promote data sharing through the establishment of central and federated Data Catalogs. Among the issues raised were how to establish minimal and relevant metadata to facilitate data sharing, broad adoption of standards to enhance data retrieval, as well as data citation and adoption of the catalog by the broader biomedical community.

BD2K is now considering the development of a biomedical Data Catalog to make biomedical research data findable and citable, as PubMed does for scientific publications. Such a Data Catalog would make it easier for researchers to find, share, and cite data, as well as the publications and grants that they are associated with. A Data Catalog is distinct from a data repository, but would help make data in such repositories more easily findable and citable in a consistent manner. In addition to supplying core, minimal metadata to ensure a valid data reference, it is envisioned that a Data Catalog would include links out to the location of the data, to the NIH Reporter record of the grant that supported the research, to relevant publications within PubMed or journals, and possibly to associated software or algorithms.

An NIH BD2K Working Group charged with exploring the concept of a Data Catalog has determined that it would be important to query a broad mix of Data Catalog designers, stakeholders, and potential users about their experiences and advice to the NIH as it considers development of a Data Catalog. In order to better appreciate the issues that need to be addressed and the possible solutions that could lead to implementation of a Data Catalog, the NIH thus seeks input from the broader research communities.

Establishing such a Data Catalog could also be part of NIH's response to the White House Office of Science and Technology Policy February 2013 memorandum, "Increasing Access to the Results of Federally Funded Scientific Research."

Information Requested

To maximize the impact of this potentially valuable community resource and facilitate its use by scientists with a broad range of expertise, we seek input on a proposal to develop a Biomedical Data Catalog. Your comments can include but are not limited to the following categories:

- Your area of expertise and interest in a Data Catalog. This may include, biomedical researcher, informatics professional, library sciences expert, publisher, professional society, or participation in another stakeholder community.
- The critical barriers, opportunities, or incentives to making data more easily discoverable and citable, and the possible impact of a Data Catalog.
- Possible Data Catalog linkage to existing data repositories to ensure data within the repository are findable and how to ensure that such linkages remain up to date and accurate.
- If your research field has no existing repositories to store data, comments can include how a Data Catalog might usefully link out to the data and where such data might be located.
- How the lack of a data repository might affect data discoverability, usability, and citability.
- The useful level of granularity for a Data Catalog entry. For instance, a Data Catalog entry may correspond to all the data in a publication, only a particular data type within a given study, or individual dataset from a single experiment.
- Any potential requirements for Data Catalog registration of data by NIH-funded or supported investigators.

- Whether a Data Catalog entry benefits from a scientific abstract that describes the data, including its potential uses and the rationale for its creation.
- The feasibility of the development of a Data Catalog to potentially support future uses.
- The appropriate metrics to use to create a successful Data Catalog.