# BD2K Data Discovery Index Workshop Summary Report

## 1. EXECUTIVE SUMMARY

A data catalog is not a data repository but rather a place where data is described with an index to what is available. As the workshop participants examined the need, justification, and value of a NIH Data Catalog it became clear that central focus should be on data discovery, accessibility, and citation. Successful implementation of a Data Catalog will require collaboration of research funders, biomedical researchers, and practitioners of big data science, data stewards, publishers, and users forming a Big Data ecosystem. This will, in turn, require adoption of broad data sharing and citation behaviors built upon an appropriate balance of incentives and requirements and supported by this diverse ecosystem. As a member of this ecosystem the NIH would play a role in convening the stakeholders, funding start-up activities, and overseeing a working process within which development can proceed.

It was recommended that the NIH consider a modified definition of a data catalog as a **Data Discovery Index (DDI)** that is designed to enhance discovery of data through detailed indexing of shared datasets deposited at many sites but with sufficient metadata to allow for discoverability and accessibility and with identifiers to facilitate citation. This data should be "live" and of value to the community. Thus the DDI would need to work with domain-specific and institutional data repositories, journals, and the private sector to ensure that data in these diverse places can all be indexed in the DDI and to develop a robust culture of data citation. The workshop emphasized the importance of developing a DDI in a framework that supported rapid innovation, testing, failure as a learning tool, and improvement of the concept. Workshop participants recommended several pilot projects to be considered in the short term that addressed different opportunities and challenges. For the longer term, they articulated the goal that the DDI should enable trans-disciplinary data indexing and queries – both simple and complex – from users at all levels along the continuum of informatics sophistication to tackle questions that advance the mission of the NIH. The DDI must combine intuitive searching and visualization tools to enable users to find datasets of interest and relevance to specific scientific questions. Finally effective stewardship of the DDI must be developed to assure that the data remain of value to all stakeholders and users.

## 2. Background

**<u>Data Sharing: Vision and Problem Statement</u>**

The mission of the National Institutes of Health (NIH) is to seek fundamental knowledge about the nature and behavior of living systems and to apply that knowledge to enhance health, lengthen life, and reduce illness and disability. NIH was founded on the enduring premise that public investment in biomedical science yields new knowledge that benefits the public. Two important corollaries of that premise are that publicly-funded science should generate data that is publicly available whenever feasible and that greater sharing of data will accelerate scientific inquiry and discovery. Thus, society receives a great benefit from its investment in research.

Biomedical research is witnessing a very large increase in the amounts of data generated from measurements on living systems. Ever-increasing quantities of data are emerging from NIH funded research. This data impacts at all levels and has had a dramatic effect on the growth of knowledge in the biomedical sciences. However, this data has even more potential to revolutionize next generation health care in the U.S. if it can be accessed and this “big data” can be converted into knowledge. There are three fundamental challenges that impede this dramatic potential. First and foremost there is no easy query or search infrastructure (i.e. a “Google” for research data) that can help identify the presence and availability of such largesse of data sets, given that journals which present primary results from these data measurements rarely contain depth of information in an easily searchable manner to facilitate identification and access of the data. Second, the “linguistic” richness of the biomedical terminology makes it difficult to relate apparent diverse but potentially related data that could serve as a rich source of biomedical knowledge. Third, usefulness of data for conversion into knowledge relies heavily on the nature and availability of metadata (data about data; see https://en.wikipedia.org/wiki/Metadata) and the introduction of combined pointers to data and associated metadata that will help the researcher create new knowledge in biomedicine.

These intuitively appealing and seemingly self-evident principles contrast with the reality of NIH-supported science. As noted in the 2012 report of the Data and Informatics Working Group (DIWG) of the NIH Advisory Committee to the Director:

> *Currently, data sharing among biomedical researchers is lacking, due to multiple factors. Among these is the fact that there is no technical infrastructure for NIH-funded researchers to easily submit datasets associated with their work, nor is there a simple way to make those datasets available to other researchers. Second, there is little motivation to share data, since the most common current unit of academic credit is co-authorship in the peer-reviewed literature. Moreover, promotion and tenure in academic health centers seldom includes specific recognition of data sharing outside of the construct of co-authorship on scientific publications. The NIH has a unique opportunity — as research sponsor, as steward of the peer-review process for awarding research funding, and as the major public library for access to research results. [http://acd.od.nih.gov/06142012_DIWG_ExecSummary.pdf]*

Even with the vast riches of data currently available, life scientists typically approach a new project by determining how to procure their samples, collecting the discrete data around a narrow focus (genome, transcriptome, proteome, or microbiome), and analyzing those data in isolation. Despite the burgeoning number of biomedical databases, as evidenced by the annual Nucleic Acids Research (Journal) update that has grown to over 1500 databases, even well-established databases such as the NCBI’s Gene Expression Omnibus (GEO) that has nearly one million datasets freely available are underutilized. **The obstacles to finding resources and evaluating their utility makes it easier to re-collect data than to use existing data and develop a well-reasoned hypothesis that can be tested with targeted follow-up experiments**.

A long-term goal of a Data Discovery Index (DDI) would be to promote new scientific discoveries, new collaborations, support a more diverse system to acknowledge scientific and academic achievement, and increase transparency and accountability of the results of government funding of science. The DIWG recommended creation of a new technical infrastructure in the form of 'catalogs and tools to facilitate data sharing' (Recommendation 1b):

> *The NIH should create and maintain a centralized catalog for data sharing. The catalog should include data appendices to facilitate searches, be linked to the published literature from NIH-funded research, and include the associated minimal metadata as defined in the metadata framework to be established… [ibid]*

**<u>Catalogs: Then and Now</u>**

The idea of a catalog is familiar and ordinary; the experience of browsing a printed listing of offerings assembled and distributed by vendors has been a universal part of modern society since the advent of the printing press and public literacy. "Multimedia" catalogs with both words and pictures have become an icon of Americana since the Sears Roebuck catalog, which was broadly disseminated and served marketing, educational, and even sanitary purposes in the era preceding indoor plumbing [http://www.searsarchives.com/catalogs/history.htm]. Today's "traditional"-style catalogs look more like Ebay than the Sears Roebuck catalog of the past and incorporate sophisticated searching, rating and other relevant metadata to facilitate their use.

In the era of 'big data', which is essentially always produced and maintained in digital forms and amenable to communication via digital communications networks, the notion of a catalog is no longer bound to a physical object. The rise of Internet search engines as both locators of information and immediate providers (via hyperlinks to the actual sources) has **recast the notion of a catalog from a book or even a persistent database to a collection of dynamic, real-time computing functions**. Some of these functions assist human users to find ranked lists of resources that match or are closely associated with words entered via a search interface, thus emulating the keyword indexes and tables of contents of traditional printed catalogs. But newer 'catalog' functions have no precedent in printed books and are thus not "bound" by their limitations. Functions such as the real time synthesis of disparate data from multiple sources to create query-specific displays of graphical maps with overlaid points of interest near a user's current location represent a strong justification of an index that can speed discovery of data.

What constitutes a NIH Data Catalog? The NIH Data Catalog will serve as the indexed repository of publicly available biomedical data-metadata and will embed a navigation map of the knowledge graph representation of the biomedical data combined with accompanying metadata. Thus we now speak of a NIH Data Discovery Index (DDI). The workshop opined that while the DDI will only provide pointers to the data-metadata, the success of such a Catalog will have to be founded on existence of strong data management infrastructures which contain data and metadata in structured and ontologically-based formats.

For the purpose of this workshop, the most general view of 'data catalogs' was taken, where a catalog could in some cases be a human-viewable database analogous to a traditional printed catalog and in other cases, could be a set of functions to serve both human users and, increasingly, machine interfaces (i.e., 'computers talking to computers') to support the needs of scientific data discovery, exchange, and analysis. In the latter case, the driver for defining a new functionality is the 'use case' – a real world scenario of how the 'catalog resource' would be used to facilitate data sharing. In some cases, the best digital catalogs may not exist as distinct resources known to its users but instead may be invoked by a question or request and provide a real-time, seamless connection directly to the data sets that relate to the query. Unlike the printed catalog of the past, it can be predicted that a new resource that enables locating, characterizing, and accessing NIH-funded data will have to evolve in an agile way to serve both data producers and data consumers to keep pace with the ever-changing, networked world. Technical approaches to describing, finding and providing access to the broad variety of 'data objects' that are the output of contemporary biomedical science will likely continue to evolve and improve.

The wealth of knowledge contained in these data can be mined and integrated with context-specific data to glean novel insights into normal and diseased function in humans. Researchers are increasingly recognizing the power of data in biomedical research and there is increasing scientific literature on integrative analysis of data.

## 3. Gaps and Challenges to Accessing Data.

**A data discovery index should facilitate finding, accessing, and citing data to support a holistic data re-use ecosystem.**

The Data Discovery Index (DDI) is envisioned to be part of a holistic environment (ecosystem) that supports the discovery, re-use, and exchange of data. It is a critical component of the data infrastructure needed to support 21st century research and innovation across all disciplines, particularly within the biomedical sciences. The DDI should encompass more than just a listing of the metadata of datasets funded by NIH but should also support a set of platforms and tools by which biomedical researchers can find, reanalyze, and reuse data and which will support data citation metrics and precise scholarly attribution for data.

Today's researchers seek to answer complex questions by discovering, mining, analyzing, and combining information from trans-disciplinary data sets. Even a conceptually straightforward question like – "Are you more at risk for asthma in Mexico City than Los Angeles?" may require access to a diverse array of different data sources (web sites, data bases, etc.) that may come from disparate communities and may include clinical data, research data, population density data, air quality data, location of factories and hospitals, etc. If the questions become conceptually more difficult, e.g., "What genes correlate with behavioral phenotypes observed in Parkinson's disease and what cell types are they found in?" the task becomes more daunting. We know from efforts to create data catalogs for specific communities that potentially relevant data sets may be distributed over multiple locations, creating a prohibitive barrier to those currently seeking to find and use them. Knowing where the data is and having enough metadata and relevant links to other resources would foster the ultimate goal of enabling and accelerating scientific research.

The DDI should be built with an expectation that it will serve the researcher, the funder, the student, and the lay citizen. An important side-benefit from a funder's perspective is that **the DDI should provide a mechanism for identifying and tracking data produced by federal funding making it possible**

**to link data to individual grant awards and thus track the impact of federal funding decisions.** Researchers today and in the future will benefit from a DDI that supports sophisticated discovery tools that can provide information at all levels with links to relevant related publications. The DDI should provide both human and machine readable interfaces that will support and encourage development of solutions to complex data science challenges associated with data discovery, integration, and citation.

## 4. What are the critical characteristics of a DDI?

### 4a. Key Characteristics

The NIH DDI should build social communities around data, facilitate data interoperability, and link data and associated tools to literature. Thus the DDI should:

1. Support creation of new ways of discovering data across research communities and data types with low or no barrier to access for new users. The DDI should provide an essential framework for data discovery and citation that can support a variety of new applications for data discovery, integration, and citation to meet the evolving needs of the community. Users (including researchers, educators, and the general public) should benefit from such discovery tools built on the DDI platform that facilitate discovery and enriched data mining.
2. Support citation of data. Data citation is an appropriate and important recognition for data producers. By linking subsequent publications to specific data sets, data citation and associated provenance information acknowledges the contributions of data producers to research. Citations are also a metric that can be used by NIH and the academic communities to assess scholarly activity. Data citations and access to data guarantee transparency and reproducibility, which is a publicly visible assurance of the integrity of scientific research.
3. Be extensible and scalable as new needs emerge. The DDI and associated applications should grow in scope, content, and approach through community contributions. The DDI should be a loosely coupled system that encourages third party access so that a rich ecosystem of tools and services become available.
4. Leverage existing activities. The NIH should partner with stakeholders, including community databases, institutional data repositories, domain-specific data catalogs, and the private sector.
5. Be easily accessible for users and data providers. The DDI should be a system that allows both easy data indexing and rapid discovery. Utilization of the DDI will be greater if the system incorporates useful metadata that are common across multiple data sources.
6. Provide services in both human-readable and machine-actionable forms. Some research communities have developed their own standards for machine-actionable, rich metadata. By providing pointers to those resources, the DDI can facilitate the development of tools and services that take advantage of domain-specific metadata for analysis, visualization, data integration, and other applications.
7. Incentivize good data management practices. In addition to using trusted digital repositories, the DDI will seek to index data from a variety of sources throughout the biomedical community.

8. Design the structure to prioritize links to "live" (accessible) datasets and develop a strategy to encourage stable stewardship of DDI-linked data. Develop a structure that can associate "fresh" (newer or enhanced) versions of linked datasets with the original target datasets.
9. Support organic growth by iterative development. Look for early successes and feedback from the user communities for subsequent expansion and increased utility and complexity. Where possible the DDI should use existing standards, for example, using ORCID identifiers or other DOIs to support provenance and identify data authors.

**4b. Concepts to address the DIWG call for an NIH data catalog.**

The discussion at the Data Catalog workshop identified several different conceptions of what a DDI might look like. This section explores several of the possibilities that came up. The strong recommendation from the workshop was to implement something fast and light with as simple a metadata model and registration process as possible, with the understanding that this could be extended and refined in the future. Concepts explored included a Data Catalog per se, a Data Exchange, and a Social Network. However, examples presented by Ramananthan Guha from Google showed how a dynamic index could be created, using evolving microdata formats, metadata, and vocabulary standards to achieve a centrally localized and managed index. The keynote presentation from Dr. Kenneth Casey from NOAA emphasized the importance of community engagement. Building upon existing resources in use by the community, it would be possible to define appropriate initial metadata for biomedical resources. These metadata could then be utilized by community data resources to markup their content to help launch a Data Discovery Index.

## 5. Featured Workshop Discussions

**5a. Roadblocks**

Development of a NIH DDI will have to confront several roadblocks on the way to implementation and broad use. As an index it will need to rely on easy availability of data that can be linked and indexed in ways that foster ease of use. Such easily discoverable data resources would, in turn, foster the development and application of tools through third parties to permit users at all levels of sophistication to make use of the data indexed. Doing so in a community-based manner will both justify the broad community buy-in to the concept of a NIH DDI and create synergies through ease of discoverability of the data. The DDI should support data discovery and citation across multiple field-specific data repositories and catalogs, to make data discovery both easier and more comprehensive. An important and community-wide barrier to data sharing is the concern researcher may have of the perception of being "scooped" on discoveries extracted from their own data or their data being used and analyzed inappropriately. Concerns about data provenance and ensuring credit for data producers exacerbate these concerns. Academic perceptions of data sharing as a positive metric for career path advancement will need to advance if the academic culture of data sharing is to evolve in the direction envisioned in the DIWG report. Agreement on technical details such as universal identifiers, citation standards, and links to other resources will also need to be worked out.

**5b. Metrics for success**

To assess success and determine lessons learned from development and implementation of a NIH DDI, it will be necessary to consider that the DDI will exist as part of a data sharing ecosystem that includes data providers, funders, users, and publishers. This ecosystem will require ready access to the data in ways that can solve the problems of credit for the submitters, ease of discovery and access by users (with firm commitments to appropriate citation), and active participation by journal editors and publishers to ensure that data associated with publications is available for indexing. A unique and permanent identifier for data submissions that can be used in citations will help address concerns regarding data citation and attribution and may facilitate the necessary cultural change to promote a data ecosystem. A successful DDI will result in adoption of scalable and evolving methods to link data to the DDI, will enable easy data discovery and citation and should facilitate the BD2K goals for data sharing, re-use, and support of new research collaborations and discovery.  Quantifiable metrics should include measures for how much data is indexed in the DDI (both existing repositories, from publications, and from other sources), use of the DDI as indicated by numbers of queries or activity of links out to resources, and use of primary and secondary data citations in the scientific literature.

**5c. Linkage with other community "products"**

To overcome the perception that the NIH DDI is competing with other data catalogs, development of the DDI should be through active engagement and collaboration with key stakeholders. Development of the DDI should be an active process of identification and linkage to academic, institutional, and private-sector data catalogs and data resources. The NIH DDI should seek to position itself as an integrated member of a data catalog community, not as the dominant player, and should work collaboratively with other federal agencies to identify, develop, and implement best practices for data discovery and citation. Development of reciprocal links with other data catalogs will signify the maturation of an NIH DDI through scaling and expansion of the content of the index and will support more comprehensive and seamless data discovery that is not constrained to the limits of any one data resource.

**5d. Fit with BD2K**

The DIWG report to the ACD and Dr. Collins recommended establishment of an NIH Data Catalog. Discussion and refinement of the Data Catalog concept arrived at a definition of a Data Discovery Index that would enhance the ease of finding data (discovery) based on a searchable index of available data resources. While the DDI would address key issues of supporting data discovery and citation, it is expected that other activities within BD2K would help ensure that data would be useful and actionable. For instance, the BD2K initiative might partner with software and tool developers to enhance the usefulness of the DDI to the user community or with professional societies and journals to strengthen tools and practices to support data citation.  Just as the DDI would enable discovery of data across multiple biomedical data resources, an NIH DDI may also be interoperable with data catalogs from other government agency data catalogs, as part of a larger data ecosystem consisting of funding agencies, data intensive communities, and the broad public base of users. By supporting such trans-discipline and trans-agency database data discovery, one important mandate of the OSTP and OMB memos will have been achieved.  By working with field-specific data repositories and catalogs, the needs of specific biomedical research communities will also be addressed while also linking such resources out to other and more diverse data resources.

## 6. How do we proceed? Proposed use cases and pilots.

### 6a. Data catalog and publications

We provide a few proposed DDI pilots that could be implemented in the short-term and that illustrate the need for a DDI that point to primary data sources. Additional use cases were generated during the workshop and are included in the appendix.

The workshop strongly recommended that the NIH support a variety of small, experimental pilot projects to test development and implementation of a DDI. Experience with web technologies has shown that design is extremely important and that we rarely get the first version right. Consequently, it is very important to conduct a number of small scale experiments with possible solutions for building the data catalog before we begin large scale implementation. The experiments should be done in a public shared space so that the entire community can both help and critique it. The NIH should be willing to carry out short-term experiments with the understanding that some may fail while others may provide solutions that can be integrated into a final plan for implementation of a DDI.

1. **Link out to supporting data from publications.** This pilot project would work with interested journals (such as PLoS, BMC, or Nature Genetics) to require that every table and figure links out to original data and software. This may be implemented in a focused fashion, such as special issues or focused topics, to make the pilot feasible.

   One challenge in this pilot is that not all data has an existing repository where they could be shared. While some underlying data may have an existing repository, not all data may have an obvious home. Thus this pilot would also pilot making such "homeless" data available through commercial clouds, so as not to burden researchers affected by this pilot with finding or paying for data sharing resources. Google has offered to make their 'cloud' available to the NIH for purposes of experimenting with the development of a DDI at no charge as long as the number of page views does not exceed 5 million per month and as long as there is no liability attributed to unexpected server crashes. Both NIH directly, and indirectly through academic stakeholders, will be permitted within this experimentation sandbox.

   This pilot would show that publications can incorporate links to indexed data, will help inform whether such publications might be more useful and potentially better cited than current practice, and would also help inform potential roadblocks to implementing such linkages between data and publications through publishers and data resources. Data resources may include existing domain-specific repositories, institutional data repositories, or other resources including commercial clouds. It is recommended that this pilot effort start as soon as possible and with as little process overhead as possible, with definitive milestones and timelines, so that NIH and the stakeholders may rapidly and nimbly learn from these efforts to inform subsequent DDI development.

2. **Make data available from NIH-funded clinical trials that demonstrate both positive and negative results.** Clinical trials with negative results may have difficulty in getting published, and represent an untapped and important scientific resource. All clinical trials are expected to be registered on clinical trials.gov, including clinical trial metadata. This pilot could work to make the data available through the DDI and provide a means for researchers to find, access, and potentially re-use all forms of biomedical big data. Success metrics for such a pilot might include new scientific publications (review and analysis articles) that cite these datasets and trials. Incorporation of negative trial data into new research articles by intersection of the actual data with other datasets would be an indication that DDI is serving one of its intended purposes. *While this aim is important, it may best be a mid-term objective after establishment of a core DDI.*

3. **Make data access easier through a "blue button" data download in PubMed.** One successful example of how to make the availability of data more noticeable is the Veteran Administration's Blue Button (http://www.va.gov/bluebutton/). The VA Blue Button provides an easy and obvious way for Veterans to access their personal health information from the VA Electronic Health Record and other key data sources, and thus fosters patient engagement and supports patient-centered care. It accomplishes this by providing a prominent download button for users to access and download their medical history. This pilot would work with a few key data providers to enhance the visibility of dataset availability.

   Currently, PubMed provides visible links to providers of the article itself (see Elsevier link in Figure 1 in the Appendix). But the availability of data is currently "hidden" in the link-out section on the PubMed abstract pages (see Figures 1 and 2 in the Appendix.). In order to make data a first class citizen, a data download button could be provided in the same area on the PubMed abstract page (see Figure 3 in the Appendix for a mock-up). This would immediately notify the user that associated data is available. Since a publication may have multiple data products associated with it, an intermediate landing page may be necessary that will provide a listing of data products available. This page could be a part of the NIH data catalog or of a domain specific data catalog and would provide direct links to the individual data pages (e.g. at the domain specific repository) for the data products available.

   Currently, there are a number of community data catalogs and repositories, e.g. Neuroscience Information Framework (NIF) and the NIDDK Information Network (DKnet), that already provide PubMed with LinkOut information for data housed in their resources. Both NIF and DKnet already provide landing pages for such resources and could serve as an incubator for testing a more visible data download link on the PubMed site. For PubMed to display this information correctly, additional information may need to be captured in the XML files that contain the LinkOut information or use (e.g.) schema.org to mark-up or denote that a specific link is for data associated to a publication. For instance additional SubjectTypes may need to be defined so that data will not be classified as miscellaneous. Such a pilot could begin to define best practices and methods for community resources to provide information on data to PubMed and for PubMed to more visibly display this information.

4. **Ensure that data resources generated through NIH funding can be associated with those specific awards.** This pilot may work with specific NIH programs (either within BD2K or IC-supported programs) and with NIH OER to test how the DDI can be used to associate data sets with those awards. One possible test may be to have DDI links or landing pages (as in the Blue Button example) available on the NIH Reporter page.
5. **Experiment with community building and information APIs used within the private sector.** Consider pilots that support social networking and augmented information interfaces, i.e.
    a. Structure the DDI to interoperate with tools that support voluntary community building "Facebook-style" around specific data collections or within specific areas.
    b. Structure the DDI to provide "yellow pages" or "Ebay-style" collections of relevant datasets within specific areas. Facilitate association of Amazon-style tools with these collections (if you liked data set X, you may be interested in data sets Y, Z, …).
    c. Associate data with both structured keyword sets and user-provided keyword sets. Make sure each can evolve over time to assure that the DDI is not static.

## 7. Towards Development of a Data Discovery Index

In summary, a Data Discovery Index (DDI) **emphasizes development of an adaptable, scalable system through active community engagement** that would serve as an index to large biomedical datasets. Rather than in a traditional "catalog" the DDI concept stresses discoverability, access, and citability. As such the DDI should be subject to rapid development through a series of pilots that seek to demonstrate feasibility for data discovery, accessibility, and citation. An iterative development approach over the short- to medium-term to establish a working concept would be followed by a long-term implementation plan that would also include consideration of sustainability**. It will be important in the early phases to accept failure of individual experiments as a means of learning what works and to help identify the best way forward**. Development through these early phases would require robust community engagement with stakeholders, potential users, and an external scientific panel. Policies will need to be defined and/or changed, including a requirement for data sharing as part of grant submission and funding, provision of policies for data entry to the index, and appropriate tools to facilitate citation. Sufficient resources will be needed for project management and to develop initial pilot projects/experiments leading to implementation of a final functional DDI. Proposed outcomes should be realistic and not promise too much in ways that unduly raise community expectations. The creation of a DDI ecosystem that includes stakeholders, data providers, journals, funders, and NIH policy offices will set in place the means for community development of the index and set the stage for discussions of sustainability.

The community desires rapid and immediate engagement of the NIH in the construction of the NIH DDI. Towards this end, NIH should identify mechanisms that will help launch the creation of the DDI.

**Phase I (FY14-15):** Phase I would establish collaborations with key stakeholders in the community and implement several short-term pilots intended to inform subsequent development. Phase I would emphasize leveraging existing opportunities and focusing on projects that will help inform how a DDI might be built to address needs and be scalable, useful, and sustainable.

- Initiating the process to create the design document for the DDI that lists the criteria, the goals, and the constraints for the development of the DDI. Identify the timeline for completion of this design document.
- Establish a DDI working group to create an initial information model (schema) and test implementation of a DDI. It is expected this group will contain representatives from NIH and representative stakeholders with relevant expertise. This working group will identify effective methods of communication with those interested in development, testing, and application of the DDI.
- Identify a strong and substantive “alpha tester” base offering some incentives for testing and feedback.
- Initiate DDI pilots/experiments (some examples given above), which should rapidly inform iterative development of the DDI by January 2015. Each pilot should identify specific goals relevant to development of the DDI with the core goal of supporting and informing improved data sharing.
- Develop potential options to house the DDI and support its continued development.
- Articulate how the DDI might interface with other components of BD2K and Infrastructure+.

Mechanisms for funding Phase I could include supplements to existing early adopters and/or small awards (e.g. R03/R21/X01). Enabling a rapid review process for these awards would be beneficial. It might be worth considering funding a coordination center that would coordinate activities and possibly fund short-term pilot projects.

In addition to engaging participation with stakeholders outside of NIH (academia, journals, industry), NIH should also coordinate internally to ensure that all NIH partners are engaged, including NIH policy; NLM/NCBI for including NIH repositories in DDI pilots, NLM/PubMed to explore how best to link to publications and data repositories, and OER to test how to link DDI entries to specific NIH awards. Wherever possible, DDI efforts should include relevant NIH Intramural activities, including the Infrastructure+ initiative.

As Phase I pilot efforts are meant to explore possibilities and inform subsequent development, which means there may be more than one pilot active in a given area, so consideration should be given to supporting helpful communication. Cross cutting working groups during Phase I would help these diverse efforts keep informed about progress by relevant pilots, to learn from both successes and failures. For instance, pilot activities exploring data citation enacted through publishers, NLM, and others may wish to communicate about their progress and barriers. In other cases, such as exploring how to associate grant funding information with resulting data may establish another interest- specific working group that may include data repositories, OER, and other funding bodies. It is expected that subsequent funding and development of the DDI will be informed by outcomes of these early-phase activities.

**Phase II:**

- Assemble a team of software engineers-biomedical informatics researchers from academic, as well as the private sector- to implement an alpha version of the NIH Data Catalog based on the design document specifications.
- Develop a "Google"-style query system that can rapidly query the Catalog and display the results in biologically-sensible formats.
- Develop complex query tools (style sheets, API, etc.) that can be used by the community to expand the capability of querying the Catalog.
- Develop navigable graphical tools that display "data-metadata connection maps" for both general case and specific items in the Catalog.
- Build graphical point and click maps that can extract DOI information as well hyperlink to the primary data source, where applicable.
- Develop tools to routinely (daily?) test the concurrency of the hyperlinks from the Catalog to the primary repositories.
- Identify select research community who can test the alpha version and provide feedback.
- Organize a workshop marking the annual release of the alpha version involving researchers who have extensively provided input and develop the design document for the beta version.
- Develop/adapt tutorials for education of the community on the use of the NIH Data Discovery Index
- Most importantly, develop tools for "data generators" and repositories to easily enter information for cataloguing their data.

*Funding could be through supplements and/or regular research projects, either R21 or short-term (3 year) R01s. These activities will require substantial coordination, which may be supplied either by NIH BD2K staff or by an extramural award or contract. If a DDI coordination center is funded, an External Scientific Panel (ESP) could work with the center to manage peer review of supplements. As the DDI is expected to be a community-supported effort, it is expected that internal NIH activities will be represented among these pilot projects. As such, funding will likely be needed internally at NIH to support DDI activities at NLM and OER, to develop implementation of successful pilots to link to publications and grants to datasets via the DDI.*

**Phase III:**

- Release the beta version of the Catalog for public use.
- Develop feedback mechanisms for creating the final "first" publicly available development version of the Catalog.
- Analyze usage and users of the DDI and develop an evolutionary trajectory for DDI improvement. Use DDI user / usage analyses to help inform NIH policy and programs with respect to data-driven research.
- Ensure sustenance of the Catalog through development of funding mechanisms.

*Full implementation may be done through a cooperative agreement mechanism, such as a U54, or through a contract or BAA mechanism. It will be essential to begin immediate discussion of models for achieving sustainability.*

## Summary of recommendations:

The decision for the best implementation will likely be driven by the desired performance and feature set. Nevertheless, a strong case was made that the DDI should not be a monolithic, closed or top-down (i.e. NIH-driven) entity but should take advantage of innovative solutions and modern web-based programming concepts drawn from the user community to ensure the system is flexible, has content exposed to the web, and allows formation of an ecosystem around data, and including involvement of third party applications. The workshop participants strongly endorsed that NIH should proceed quickly with development of the DDI, including a definitive statement of purpose with objectives, use cases, and potential solutions in the form of pilot DDI experiments. Partnerships with academic data producers/users, journals, and the private sector will create a DDI ecosystem within which development, evaluation, and production can be accomplished. In the longer term efforts will be needed to assure both "data literacy" on the part of users and effective stewardship of the data to have it remain current and of continuing value to the community.

## Appendices

Blue Box link for quick downloads

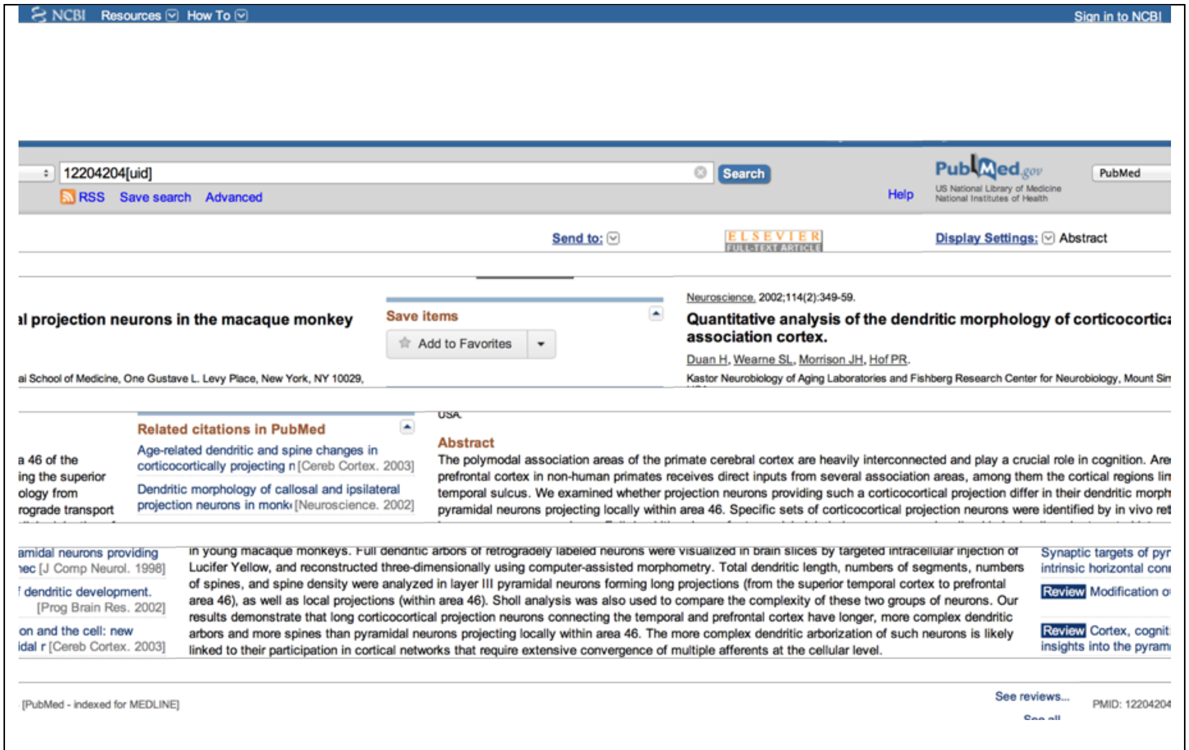



Figure 1: An abstract display page in PubMed for an article that has open data available in a community repository and that is indexed by a community data catalog. Availability of data is “hidden” in the LinkOut section at the bottom of the page.

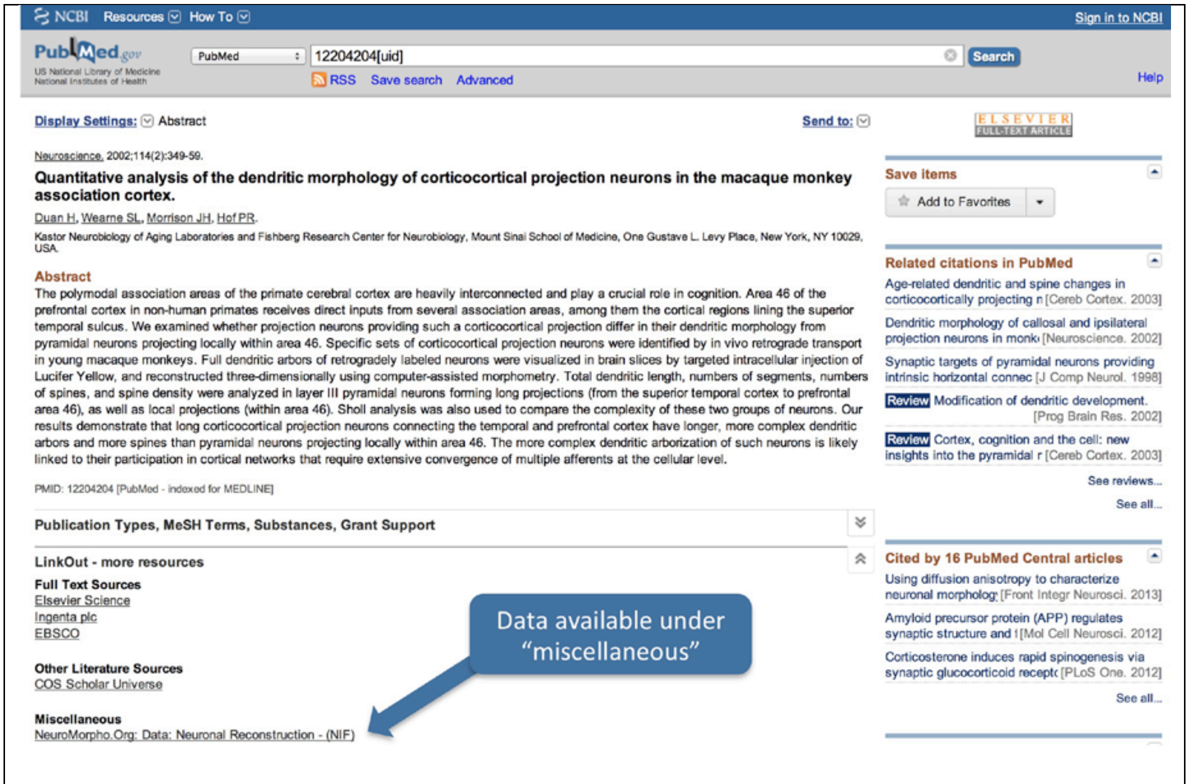
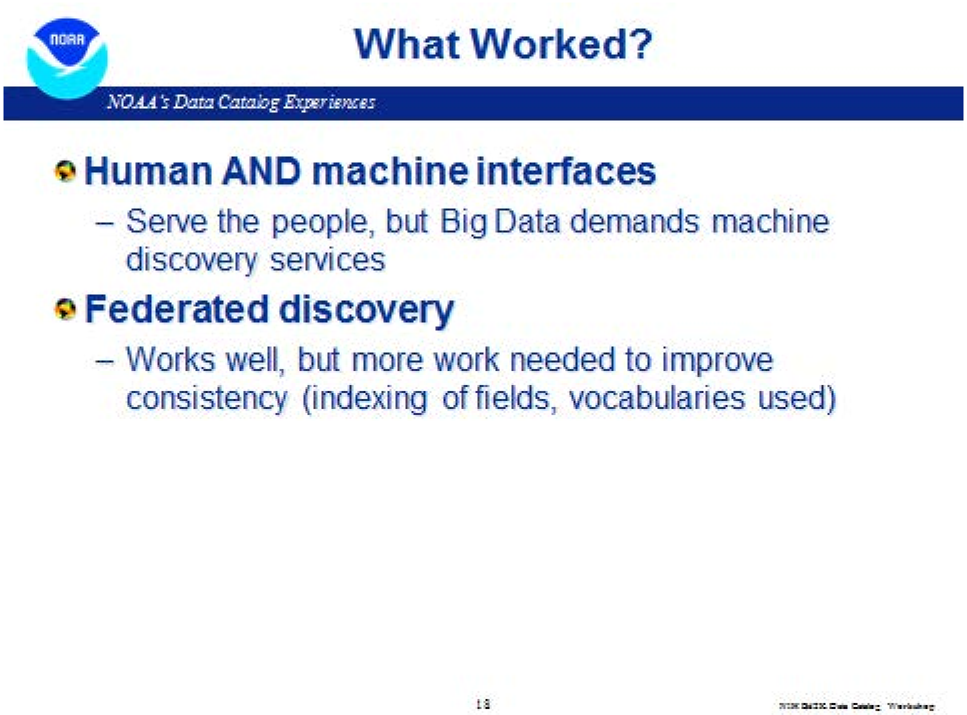



Figure 2: Opening the LinkOut section at the bottom of the abstract page reveals a list of related resources. Under the Miscellaneous section – is the listing for the data set supplied by a domain specific data catalog (i.e. the Neuroscience Information Framework). This data set was cataloged from NeuroMorpho.org that provides neuronal reconstructions.


Figure 3: A mock-up of a PubMed abstract page displaying a download data "Blue Button". This "Blue Button" could take users to a dataset availability page with links back to available data resources.

NOAA Experiences



## What Didn't Work?

*NOAA's Data Catalog Experiences*

- Black-box, whole-appliance solutions
- Reliance on in-house software
- Working alone
- Discovery portals without a solid data management back-end
- Still need to share/consolidate/interoperate on vocabularies

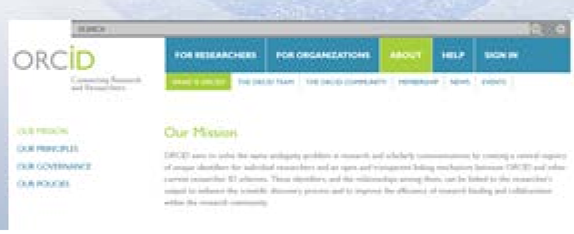


Example of a universal identifier

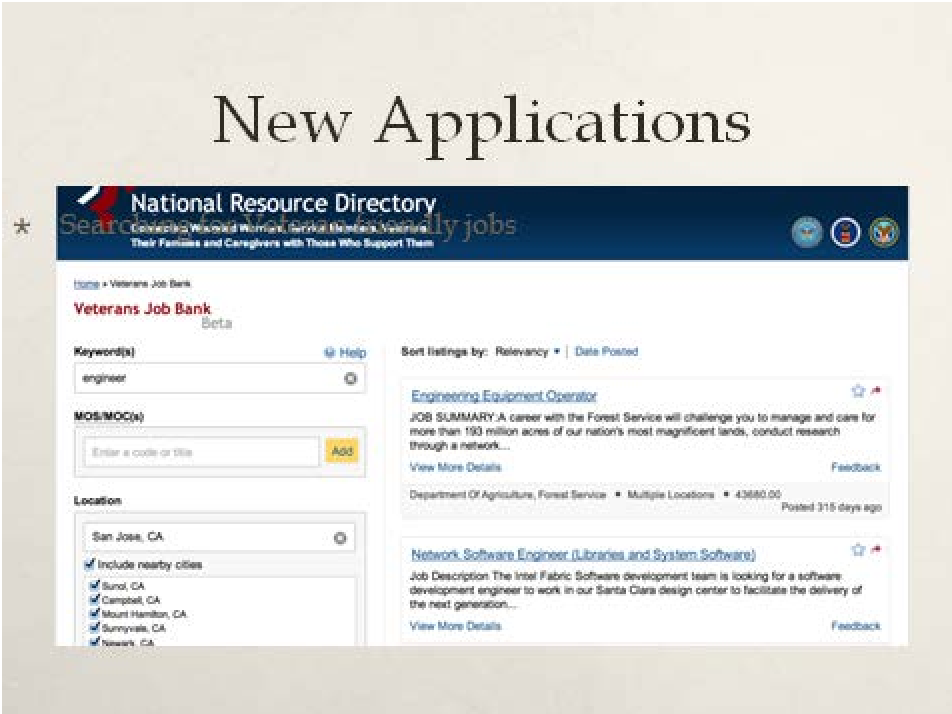
ICPSR Inter-university Consortium for Political and Social Research

## ORCID: Open Researcher & Contributor ID

- Central registry of unique identifiers for researchers
- Disambiguation of authors
- Links to other author ID systems:
  - Thomson Reuters ResearcherID
  - Scopus Author ID



Example of a Data Catalog using Google schema.org



## Additional Potential DDI Pilot Projects:

i. **Demonstrate the utility of DDI to generate applications for grants for third party analysis**. Compare requests for funding that cite accessions at different times within the grant cycle, for example data description at grant award, first deposited dataset, dataset without embargo, datasets linked to publications.

ii. **Demonstrate the utility of DDI to research tool manufacturers**. For example, as judged from DDI metadata, is the number of BAM (CEL or other data format) files generated from a particular proprietary platform (microarray, sequencing machine, mass spectrometer, microscope etc.) within a particular grant program increasing or decreasing with time/ relative to another manufacturer's platform?

iii. **Demonstrate the utility of DDI to employers**. Report on promotions and hiring of researchers, software engineers analysts and curators by institutions and NIH Institutes with and without taking DDI data citation into consideration.

iv. **Demonstrate the utility of DDI in grant respecification and renewal**. Record datasets that led to successful or unsuccessful respecification of a grant using data citation and publications as parallel metrics.

v. **Focus on a set of highly challenging use cases**

   i. **Datasets with incomplete data collection and some non-overlapping fields** are curated and combined using the DDI metadata to select datasets. Download statistics and publications resulting from the new combined datasets and citing DDI accession codes to the new combinations to be used as success metrics.

   ii. **Data to be generated are often imprecisely specified prior to data generation in required data descriptors mandated by NIH and NSF in grant applications.** Obtain metadata and data generation protocols in advance of grant award and version the data description metadata as data are added to repositories in real time. Data citation metrics and applications for data access are short term success measures. Increased publication and citation compared to comparable datasets described and released upon publication are long term metrics.

   iii. Future possible uses:
      - In the future the public will have access to their personal health information on an ongoing basis and with that information will come with questions. The public can use the Index to gather information on and up-to-date health research and information answer their questions about their own conditions. Factoring in a useful context in the form of quality of information controls will be needed.
      - Researchers may use the DDI to take a broad look at the human condition and define "the normal range" based on the data available. As this is done, both researchers and the general public will be using DDI to examine how this "normal" baseline could best be used, and to educate all on the ramifications of "normal", i.e., why no one should be worried/offended/discriminated against based on this defined range and where an individual falls in or out of it. Uses such as here proposed will help to drive a discussion of data literacy and appropriate use of public data.
      - The DDI may be a resource for education. Science will be accessible to children by presenting the data resources in a manner that enables actionable analysis even by elementary school children. Modules of guidance will provide new scientists with a robust way to jump start their own pathways to answers by re-using extant data rather

than the slower approach of setting up a lab, rounding up resources, conducting experiments and finally running analyses to determine outcomes.

vi. A biomedical researcher is working on primary macrophages isolated from atherosclerotic plaques isolated from humans. She carries out transcriptomic analysis on these macrophages both untreated and treated with a number of physiologically realistic ligands. She also carries out phenotypic cell assays on phagocytosis and micropinocytosis. She identifies several important factors that are regulated in response to TZD a drug given for treatment of diabetes and athero. However she finds that the transcription factors which are known to regulate the expression of genes that are changing are themselves not altered. She suspects that either the mechanisms is post-transcriptional or regulated by micro RNA. In the NIH data Catalog there are several independent studies with large measurements of transcription, microRNA and proteomics in murine macrophages from atherosclerosis and a quick search of the data Catalog identifies 4 studies, 2 with murine macrophages, 1 with a cell line and 1 partial study in humans. An integrated study identifies a key miRNA regulation and its other targets which regulate the progression of atherosclerosis.

vii. A parent meets with a pediatrician who suspects that the 18 month old child is displaying symptoms of a type of DMD, based on behavioral observation and imaging of the skeletal muscle. The parent informs the pediatrician that there is a history of DMD in the paternal family and one of the cousins of the child who has a similar pathology is alive. In addition to SNP genotyping of the cohort, the pediatrician also suggests to the parent that there are several measurements on skeletal muscle on children available and the one of the parents who is in a life science profession can access these using the NIH Data Catalog. This will help the parent adjust the life style of the child as well as explore appropriate therapies but must be accompanied by the appropriate tools to help use/make sense of the data.

viii. A patient who was traveling in the pacific islands reports fever and the physician suspects a bacterial infection and prescribes a broad spectrum antibiotic. The patient is non-responsive to the antibiotic, shows unusual symptoms and the physician isolates from the culture a microbe which he sends to CDC. CDC researchers sequence the microbe and compare against known organisms. They find large similarity with another microbe, and go to the NIH Data Catalog to identify data on patient symptoms and response to infection by this related organism. They find several cases and treatment by a very specialized antibiotic.