NIH National Institute of Environmental Health Sciences

Making Biomedical Data Usable: NIH Community-Based Data and Metadata Standards Efforts

Allen Dearry, Cindy Lawler, Rebecca Boyles, Astrid Haugen, Mike Huerta
National Institutes of Health

Abstract

The mission of the NIH Big Data to Knowledge (BD2K) initiative is to enable biomedical scientists to capitalize more fully on the Big Data being generated by research communities. With advances in technologies, these investigators are increasingly generating and using large, complex, and diverse datasets. However, the ability of researchers to locate, analyze, and use Big Data (and more generally all biomedical and behavioral data) is often limited for reasons related to access to relevant software and tools, expertise, and other factors. BD2K aims to develop the new approaches, standards, methods, tools, software, and competencies that will enhance the use of biomedical Big Data by supporting research, implementation, and training in data science and other relevant fields.

One initiative within BD2K is to establish community-driven frameworks for developing and using standards for data and metadata. Such standards enable broad data sharing and reuse of data generated across the full spectrum of NIH-relevant research, from single investigators conducting R01-driven research to large collaborative networks and consortia. Standards for the metadata that describe the samples and experiments associated with the data, in addition to standards for each of the data types themselves, would greatly facilitate (and are probably even required for) large-scale data sharing and data integration. NIH should help establish flexible frameworks for developing data and metadata standards for newly emerging data types that are expected to be used widely, thereby encouraging various biomedical research communities to develop such standards in coordinated ways. Priorities for standardization should be community-driven. Standards should be applicable to both research and clinical data, where appropriate. It will be necessary to address a range of issues, including developing common data formats and data elements for particular types of studies and linking established care standards to meaningful use standards for electronic health records (EHRs), to the extent possible.

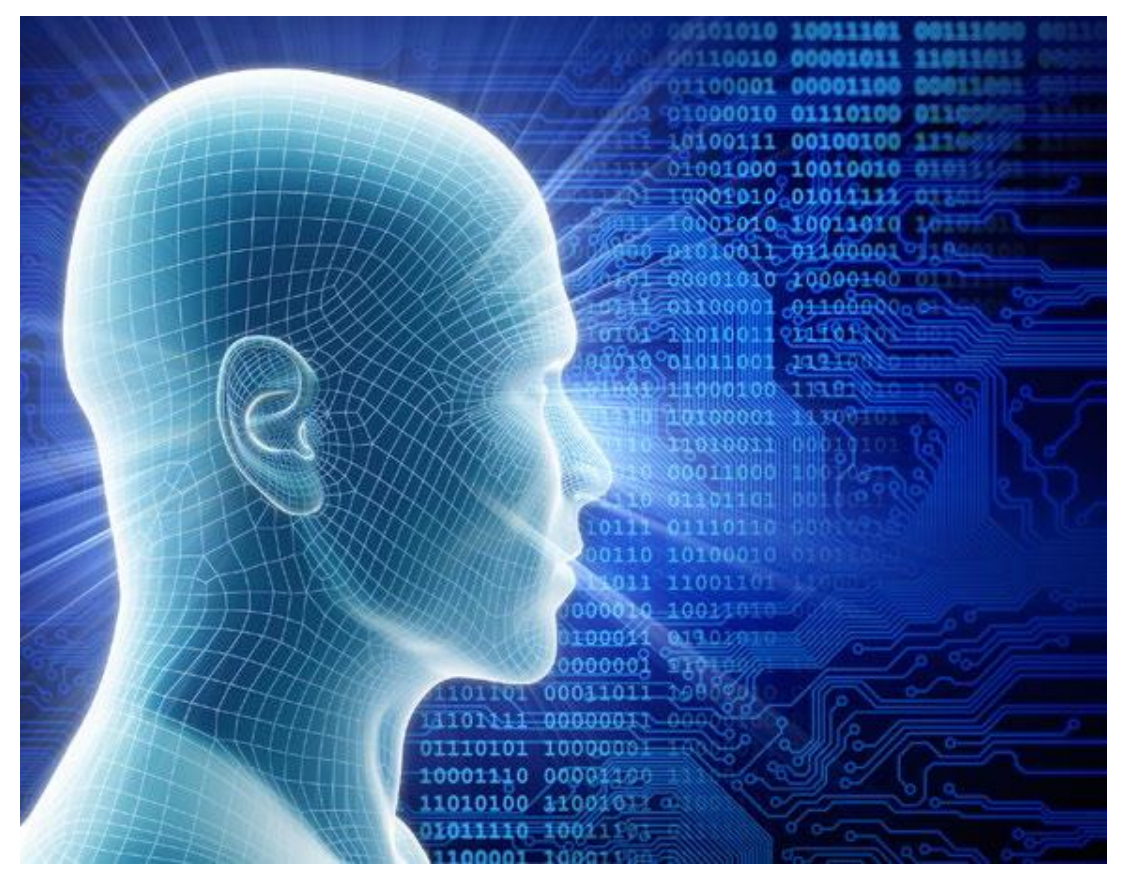
This poster describes the process that NIH is initiating to guide the support and development of community-based standards.

Background

Big Data to Knowledge (BD2K): Overview

Overarching goal:

By the end of this decade, enable a quantum leap in the ability of the biomedical research enterprise to maximize the value of the growing volume and complexity of biomedical data



The BD2K initiative addresses four major aims that, in combination, are meant to enhance the utility of biomedical Big Data:

- To facilitate broad use of biomedical digital assets by making them discoverable, accessible, and citable.
- To conduct research and develop the methods, software, and tools needed to analyze biomedical Big Data.
- To enhance training in the development and use of methods and tools necessary for biomedical Big Data science.
- To support a data ecosystem that accelerates discovery as part of a digital enterprise.

Making Data Useable

Make Data Broadly Useable

Standards allow data to work with:
- Other data
- Software tools
- Data resources

NIH Standards Information Resource
- Encourage the adoption of existing, widely used standards
- Discourage unnecessary duplication of effort / reinventing wheel

Support community-based standards efforts
- Standards are used when community wants & supports them
- BD2K will develop routine ways to provide **time limited** support for **particularly opportune** community-based standards efforts

Purpose

Community-Based Data and Metadata Standards

The purpose of this initiative is to accomplish three main goals:
1) establish an internal NIH framework of policies, governance, administrative procedures, and funding to routinely support community-based standards efforts;
2) use that framework to provide catalytic extramural research support for particularly opportune efforts under BD2K, that are broadly relevant to NIH research; and
3) integrate the framework for standards development into other BD2K activities to identify and capitalize on potential synergisms. The framework for standards development will include catalytic support, in the form of time-limited financial assistance, for convening, organizing, and logistics toward facilitating a community of practice that addresses well-formulated standards-related needs that may include creation or extension of a standard.

Planning Activities

2013 NIEHS/EPA Language Workshop

Purpose:
- Learn about standard language efforts in the field of environmental health science.
- Discuss the way forward for environmental health sciences terminology.
- Develop a local community of standard language expertise within the environmental health sciences.

Findings:
- Active data stewardship/curation adds value and is needed at some level; but we have no funding model to support data stewards and no way to measure the value of their contributions compared to, e.g., new research grants.
- Sociological barriers to data sharing (need for "culture change") within and across communities are as serious as technological barriers.
- Many community-driven and community-developed standards already exist and more are being developed; these different solutions are just starting to meet at the interfaces between research disciplines.

2013 BD2K Stds Framework Workshop

Mapping the Landscape of Community Standards

Formulating, Conducting and Maintaining Community-Based Standards Efforts
- How is the need for a particular standards effort identified?
- What is the process used to assess and prioritize selected activities?
- How do participants contribute to the standards effort?
- What are the characteristics of the ongoing discussions/meetings?
- Are milestones or similar indicators of progress used, and if so, how?
- How is the product of the standards effort updated and assessed?

RFIs for Community Input

Information resources for data-related standards
- Collect, organize, and make available trusted, systematically organized, and curated information about data-related standards

Community-based standards development
- Activities that could advance community-based standards landscape (e.g., creating a collaborative workspace or an advising structure toward standards development, extension, or adoption).
- Gaps in community-based data standards of relevance to NIH research, including real use-cases (e.g., emerging fields, research domains with multiple existing data standards that could benefit from additional work, integration and/or reconciliation).
- Lessons learned from existing field-tested processes and infrastructure.
- Common challenges/pain points in development (e.g., methods for community engagement or building interoperability with other related standards).

2014 NIEHS/EPA Vocabulary Workshop

Purpose:
- Establish a collaborative and cross-disciplinary group to inform development of environmental health science language standards and applications that will aid data sharing, integration and analysis,

Considerations:
- Inventory existing resources
- Propose use cases
- Assess current semantic landscape
- Critical components of a common language framework
- Lessons from successful standards development
- Incentives, sustainability

Planned 2015 CBS Workshop

Themes:
- A Glimpse of Community Standards across the biomedical spectrum
- What is a community for the purposes of standards development? How do you identify change agents?
- Discuss lessons learned from similar community standards efforts. Pain points, and obstacles of efforts that either succeeded or failed.
- Identification of data standards for potential support. What kinds of characteristics should be considered for a need?
- End user engagement: Implementation, adoption, communication, feedback over the lifecycle
- What kinds of targeted support and assistance could accelerate the development and adoption of high quality data and metadata standards for NIH relevant research?

Future Directions

For more on BD2K:

http://bd2k.nih.gov/about_bd2k.html#sthash.qfVYTOK5.dpbs

For Community-based standards development:

- RFI, fall 2014
- Workshop, spring 2015
- Follow up? dearry@niehs.nih.gov

"Now! *That* should clear up a few things around here!"