# Executive Summary

## NIH BD2K Workshop on Community-based Data and Metadata Standards Development: Best practices to support healthy development and maximize impact (February 25-26, 2015)

A multidisciplinary group of scientists and standards experts from both the basic and clinical sciences met for a two-day workshop on February 25-26, 2015 in Bethesda, Maryland. This meeting was supported by the larger Big Data to Knowledge (BD2K) initiative launched by NIH in 2012. BD2K is an ambitious program that plans to increase the researcher's ability to use biomedical data, and is working to develop a data ecosystem that supports data, tools, software, publications and, at its foundation, data and metadata standards.

The main issues that were discussed at the meeting are as follows: 1) gaps in community standards development; 2) common technical, social, and financial pain points in the data and metadata standard development life cycle; 3) how NIH can assist the biomedical research community in addressing the identified pain points to accelerate and improve the quality of community-based data standards development; and 4) the identification of best practices for managing community-based data standards and evaluation of near- and longer-term successes. In short, this workshop served to inform NIH actions including recommendations for policies, processes, and assistance mechanisms.

This workshop built on previous BD2K activities that engaged NIH and the scientific community to discuss ways to construct a standards framework, including a workshop in September 2013 and a Request for Information (RFI) in the fall of 2014. The two-day meeting concluded with recommendations and decisions that will take into consideration the current standards landscape and community, and are intended to enable, not replace, the work already being done.

Over the course of the two-day meeting, which included panel discussions and breakout groups, several major interconnected themes surfaced, including: 1) the standards **life cycle approach**, 2) **management** of the standards development process, 3) **education/resources**, 4) **incentives** for developing and using community standards, 5) **funding,** and finally 6) the role of data science in **NIH study sections**. Within each of these crosscutting themes, meeting participants identified gaps and recognized numerous pain points that will serve as potential targets for future actions.

### Standards Life Cycle

An overarching theme of the workshop was the importance of the **standards life cycle**, which includes the **development** (initiation, establishment of stakeholders/working group, establishment of requirements/use cases, design, test, approval), **dissemination/distribution**, **adoption/use**, **evaluation**, and **maintenance (review, revision, and retirement) of a standard**. Though the focus of efforts is initially on the development stage, every stage of the life cycle requires social, technical, and financial support and must be considered carefully.

From the beginning, it is essential to involve **key stakeholders** early in the **development** stage so that the standard starts out on the right path. Data creators, data users, and data developers form the base of the standards development community, but other relevant stakeholders, such as publishers, vendors, and members of the public — including patient advocacy groups and citizen scientists — are vital to adding breadth to the community. We are not asking patients or other citizen scientists to help create the standards themselves; however, we need to assume that everyone, including the public, has a vested interest in the science and that this broad coalition of stakeholders adds perspective and use cases to the standards development process.

The first step in the **development stage** is establishing a need for a standard. Once that need is known, a researcher must survey the standards landscape to determine if a new standard should be developed or if an existing standard can instead be extended or modified. Workshop participants strongly advocated encouraging the use of existing standards; however, the infrastructure to extend existing standards is not currently in place. Researchers need help identifying and choosing the most appropriate standard to suit their data. This could be facilitated by **a standards registry** that indexes all standards and provides use cases, evaluations, metrics, and ratings to facilitate decision making. Several participants mentioned BioSharing.org as an example of a registry already providing much of this information. NIH could work with BioSharing.org to increase its value to users even more by including use cases, metrics and ratings for standards. In addition, the standards registry could provide attribution and citation information so users of the standard could properly cite the standard, allowing standard developers to receive appropriate attribution for their work.

Participants often spoke of the need to manage both the technical and social aspects of standards development. **Technical aspects** include the **development of open-source tools that support standards development and use.** Researchers need smart, easy-to-use tools that address their specific requirements. For example, tools that facilitate the use of existing standards or terminologies would help researchers find and choose a standard. Vocabulary mapping tools are needed so data can be collected in one standard and then converted to another. Data ingestion tools that validate data or can handle data exceptions could facilitate standards usage by making it easier for researchers to find and correct errors when submitting data. In addition to tools, participants suggested a collaborative, transparent work space for standards development that would provide a full audit trail of people's contributions. Many participants recommended GitHub as an ideal tool for this purpose. GitHub is a web-based revision control hosting service for software development and code sharing; it allows users to change, adapt, and improve software. A "GitHub for standards" would also allow the community to measure standard usage and to track modifications.

It was widely agreed that the **social issues** tied to standards development are more difficult to address than the technical issues. How do people within a standards community deal with others who have differing goals or approaches? How are conflicts mediated and managed? How can people receive the help needed to work together better? Improved mechanisms of communication are needed because one of the most difficult aspects of developing standards is negotiation. Building social networks for standards development and focusing on social structures that allow people to contribute (e.g., GitHub) will help address this issue. Finally, because standards adoption and use are a social and cultural activity, participants recommended bringing in social scientists to study why and how people are using standards.

The **evaluation** stage of the standards life cycle emerged as a strong theme during the workshop. Participants often talked about the need for **formal metrics and evaluation methods** to measure standards usage and usability. Is the standard working? Does it need to be changed, updated, or removed? Should a standards body be involved? How can the results of standards evaluations be used to improve the standard? Participants recognized the obvious need for standards evaluation and metrics research to be funded by NIH.

Both community engagement and funding issues emerged as important themes during the **maintenance** stage of the life cycle. Standards need to be kept up to date to maintain quality, and this takes both the will of the community and money. Attendees noted that the standards life cycle takes too long to complete and that a shorter life cycle is required. The volunteer nature of standards development was cited as one reason for delayed implementation of standards. Providing standards development projects with funding for dedicated personnel including, developers, project managers, and technical writers, was suggested as way to accelerate the process.

## Management and Coordination

Life cycle discussions frequently led to the importance of **improving the organization, management, and coordination** of community standards development. Several subthemes emerged, the broadest being the **need for standards development leadership**. Participants proposed that scientists, funders, standards developers, publishers, software/equipment vendors, and standards bodies gather together in a forum to address the leadership issue. More specifically, this workshop revealed the need for greater coordination between the clinical and basic sciences; these two communities share many of the same demands and challenges and could benefit greatly from a more coordinated effort. Participants called for a forum where researchers in the basic and clinical sciences could meet and discuss issues related to data standards development, with the end goal of coordinating their standards activities. Others went farther and suggested a potential role for NIH to provide some form of a **data standards coordinating center** to help develop, organize, and coordinate standards. Several attendees stressed the need to leverage what other scientific agencies in the U.S. and abroad are doing in standards development; standards development is a worldwide, global initiative.

A problem identified by many attendees is the overabundance of existing and often overlapping standards. The overall consensus is that NIH needs to strongly encourage **the reuse of existing standards** to prevent reinvention of the wheel and to leverage work that has already been done. This leads to questions of how to incentivize use of pre-existing standards. It also leads back to the need for tools that make it easier for researchers to use existing standards or terminologies. In addition to mapping tools, another idea was to develop a graphical browsing tool that would let the user easily find existing standards. Not only does the standards community need coordination and management, it will also need **funding to create the infrastructure** necessary to enhance, expand, and improve existing standards.

## Education/Resources

Several discussions focused on the importance of **education, training, and resources** to both the standards development process and the adoption and use of standards. Researchers need to understand the basics about standards — what they are, how and why they are used, and the qualities of a good standard. Standards education needs to be provided to both new trainees and established scientists. More specifically, grantees need assistance with all aspects of data science — from writing the data management plan to choosing and modifying a standard. Attendees expressed strong interest in the idea of providing **data concierges**, **consultants,** or **mentors** to assist investigators with data standard needs throughout the life of a grant.

Attendees also strongly agreed on the need for a **data and metadata standards resource portal** to help researchers select and adopt appropriate standards. The portal could host information guides, best practices, and the standards registry while also providing access to a suite of easy-to-use tools to identify, evaluate, and implement standards. The general consensus was that BioSharing.org has made significant headway towards this goal and that NIH should work with BioSharing to add functionality to the existing portal.

## Incentives

The need to provide **concrete incentives** (both "carrots and sticks") for standards development, adaptation, and implementation was a reoccurring theme throughout this workshop. Standards development efforts are largely volunteer-driven and, therefore, limited. One aspect of the discussion centered on how to incentivize individuals to participate in standards development or to use pre-existing standards in their research. Participants often mentioned the need for developers and users of data standards to receive **credit or recognition** (e.g., a carrot) for their work from funding agencies and promotion/tenure committees in the way of publications, data citations, GitHub contributions, or mentions in their NIH Biosketch. Many participants suggested initiating standards usage by first taking the "stick" approach by having funding agencies require the use of standards (e.g., common data elements) in funded research or score the data management plan, and then moving to the carrot approach once the standards have wide use. Participants also agreed that it can be difficult to get people to collaborate on standards development efforts and that specific incentives to support community efforts were needed.

## Funding

Incentives can be financial in nature. Attendees agreed on the urgent need for **funding to support standards development**. Standards are not free, and volunteer efforts are limited. Therefore, providing financial support to people who develop them is necessary to both accelerate and improve the data standards process. Funding could be used to support dedicated project staff, such as a developer, project manager, or technical writer, and cover expenses for travel to working group or standards meetings.

The issue of funding touched upon almost every aspect of the data standards development process, and its importance to participants was obvious. In addition to providing an incentive to participate in the development process, funding is also essential for a number of other areas including the development of tools and metrics/evaluation methods as well as infrastructure innovation and development. Every step in the standards life cycle is influenced by funding. Attendees identified funding the lifespan of a standard as a challenge — often a standard will begin its life fully funded but cannot be maintained at the same level once the grant is over.

## NIH Study Sections

The final theme that emerged from this workshop was the considerable need for NIH study sections to be knowledgeable about data standards and to value the role of data science and data standards in biomedical research. NIH study sections should reflect NIH's current focus on data science by including people with information and data science expertise to review grants that use data standards or grants that support standard development, modification, tools or infrastructure. Reviewers with data standards expertise could not only check for a data management plan (DMP), but also assess the plan's quality. Then, by monitoring the DMP, investigators could be rewarded (or penalized) for how they utilized data standards in their research. Several people went beyond adding data standards expertise to existing study sections by calling for creation of a stand-alone NIH study section devoted to data science and data infrastructure that could better review and fund grants related to data standards. The pros and cons of a stand-alone data standards/data science study section will have to be debated; the need and value of a standard should be considered in the context of its domain of use so a stand-alone study section might not be the most effective approach.

## Conclusion

At the end of the two-day meeting, participants agreed on several key take-home messages:

- The life cycle approach to data standards development needs to be emphasized.
- A broad coalition of stakeholders should be present early on in the life cycle so the standard starts on the right path.
- The development of new standards should be limited; reuse, modification, and integration of existing standards should be emphasized.
- Non-technical, social aspects of data standards development are just as important as technical aspects.
- Data standards development, evaluation, maintenance, and general infrastructure require funding.
- NIH study sections need data standards expertise.
- Data standards development needs to be a cross-agency, international effort that spans the basic and clinical sciences.
- Data standards development communities need some kind of forum where people can meet, share knowledge and ideas, and coordinate their activities.

With these key take-home messages in mind, members of the Community-Based Data and Metadata Standards Development (CBS) and the National Standards Information Resource (NSIR) working groups (part of the NIH BD2K executive committee) will now synthesize and prioritize the findings identified during this workshop in order to recommend new NIH policies, processes, and assistance mechanisms to support community-based data and metadata standards development. NIH actions will initially include small steps with low risk but then extend to addressing longer term, more difficult issues.

# Addendum

## Funders' Meeting

Following the close of the BD2K Workshop, members of the Community-Based Data and Metadata Standards Development (CBS) and the National Standards Information Resource (NSIR) working groups (part of the NIH BD2K executive committee) participated in a funder's meeting. The agenda was built around the workshop's discussions and focused on federal coordination and implementation issues. For the first part of the meeting, Bill Miller of the National Science Foundation (NSF) shared his experience with supporting infrastructure networks such as the Research Data Alliance. Members of the working group were interested in NSF's Research Coordination Networks (RCNs), which offer a mechanism for fostering collaboration and community organizing. RCNs serve as a catalyst to bring two groups together to do something that has not been done before.

The second part of the funders' meeting was a discussion of possible NIH intervention points identified from the workshop. The majority of points identified by the committee were reflected in the cross-cutting themes identified during the workshop. They included:

- The importance of considering all aspects of the **standards life cycle**: development (initiation, establishment of stakeholders/working group, establishment of requirements/use cases, design, test, approval), dissemination/distribution, adoption/use, evaluation, and maintenance (review, revision, and retirement) of a standard.
- The need for a **list of standards** (e.g., standards registry) the research community can use to facilitate and promote the reuse of existing standards. BioSharing.org already fulfills much of this need. Another example is the NIH Common Data Elements (CDE) database supported by the National Library of Medicine.
- The need for a **data standards special emphasis panel** and for the **addition of data science specialists to study sections** so data management plans and data standards can be evaluated.
- The need to support development of **easy-to-use tools** to help researchers annotate their data.
- The **need for input from sociologists** to facilitate work with communities to get them comfortable with data sharing.
- Use **of data management concierges/data consultants/data librarians** to assist investigators.
- Creation of a **data standards coordination center.**
- Need to create and fund **data standards metrics and evaluation methods**.
- **Incentives** for stakeholders including both work recognition and funding.

Committee members also proposed alternative funding mechanisms, including a mechanism similar to a U34 cooperative agreement grant to support multi-disciplinary planning activities, and an alternative R13 support for conferences and scientific meetings grant that would allow a person interested in standards to attend several different standards meetings. Members also discussed the possibility of NIH funding an organization and their standards activities instead of a single investigator.