

# Big Data to Knowledge (BD2K) Community-Based Data and Metadata Standards: Overall Strategy and New Concepts

Cindy Lawler (NIEHS) & Sherri de Coronado (NCI)  
Representing the Standards PMWG  
April 25, 2016

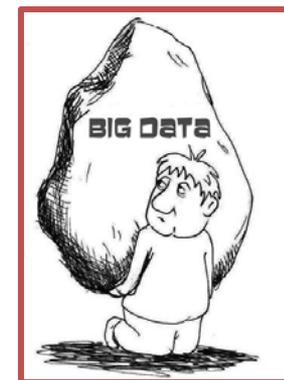
Multi-Council Working Group ► April 25, 2016





# Data and Metadata Standards

- Documented agreements on representation, format, definition, structuring, tagging, transmission, manipulation, use, and management of data
- Enable broad data sharing and reuse of data
  - Findable, Accessible, Interoperable, and Re-usable (FAIR)
- Implementation is essential if NIH is to fully capitalize on 'big data'





**Monitor the data-related standards landscape**

**Consider and develop new BD2K standards initiatives**

**Development of a Multicomponent Program**

**Standards Program Management Working Group (SPMWG)**

**Inform the implementation and conduct of BD2K standards initiatives**

**Engage with relevant standards efforts as appropriate**



# Who is the community?

- Biomedical researchers
- Publishers
- Ontologists
- Data Analysts
- Informaticians
- Software Engineers
- Database Managers
- Vendors
- Professional Societies
- Funders
- Patients





# Needs Assessment

Date	Title
September 2013	<b>Workshop:</b> Framework for Community-based Standards Efforts
August 2014	<b>Request for Information (RFI) :</b> Input on Information Resources for Data-Related Standards Widely Used in Biomedical Science
November 2014	<b>Request for Information (RFI):</b> Making Data Usable--A Framework for Community-Based Data and Metadata Standards Efforts for NIH-relevant Research
February 2015	<b>Workshop:</b> Community-Based Data and Metadata Standards: Best Practices to Support Healthy Development and Maximize Impact



## **BD2K Standards PMWG Activities**

*Standards  
Coordinating Center*

*Support for  
Standards Efforts*

*Infrastructure for  
Community-Based  
Groups*



## Standards Coordinating Center

### Brings together information and guidance about biomedical data standards

- Makes use of a web platform/portal
- Links to information on widely used standards and standards resources
- Considers different user communities (naïve vs. expert)
- Enables community engagement and interaction
- Provides tools to train, educate, and advocate for standards in research
- Publicizes resources



## Support for Standards Efforts

**Recognize and support development of biomedical data standards as a community resource.**

**Incorporate key principles:**

- Focus on standards with broad relevance & impact
- Address important gaps in standards life cycle
- Encourage community engagement
- Provide time limited catalytic support
- Leverage existing efforts
- Plan for sustainability
- Ensure open source products
- Findable through Standards Coordinating Center



## Examples of Standards Activities to Support

- Initiation, establishment of stakeholders/working group(s); refinement of technical requirements/use cases
- Technical development and/or tools to facilitate use
- Validation and 'fit for purpose' testing of the standard
- Development of metrics and overall evaluation
- Development and implementation of strategies to maximize and incentivize awareness, accessibility and uptake
- Engagement with appropriate partners for sustainability



## Phased Approach

### *Pilot Phase*

- Use NIH resource support funding mechanism (R24)
- Two rounds of competition proposed
- Compressed timeline for review and award
- 3 million/year beginning FY17-18

### *Phase II*

- Depends on data from Pilot Phase
- Option 1: Continue use of R24
- Option 2: Establish single large contract or cooperative agreement



# Discussion

*Support for  
Standards Efforts*



## Infrastructure Support for Community-Based Organizations

- Recognizes community-based organizations tackling important biomedical data science issues
- Includes data standards but much broader
- Aligned with BD2K mission and interests
- Much work accomplished by convening various stakeholder groups
- Infrastructure support is challenging



## Approach to Support Community-Based Data Science Organizations

- Use NIH conference/meeting support mechanism (U13)
  - Allows multi-year support
  - Provides transparency and rigor
  - Meetings broadly defined
  - Support for many different activities
  - 2 million/year set-aside
- Characteristics of target organizations
    - biomedical data science focus;
    - work relevant to multiple NIH I/Cs;
    - non profit;
    - community based;
    - no or low cost to participate;
    - products are open source and freely available.



## Support of Data Standards Efforts

- Community engagement
- Leverages existing resources
- Relevance to multiple NIH Institutes/Centers
- Narrow focus on data and metadata standards
- Supports technical and related work to develop standards

## Infrastructure Support for Organizations

- Community engagement
- Partial infrastructure support
- Relevance to multiple NIH Institutes/Centers
- Broad focus (standards or other biomedical data science area)
- Supports only meetings



# Discussion

***Infrastructure for  
Community-Based  
Groups***



# Standards PMWG Members

- Dearry, Allen (NIEHS)
- De Coronado, Sherri (NCI)\*
- Eftekhari, Aras (NCI)
- Federer, Lisa (OD/ORS)
- Fore, Ian (NCI)
- Foster, Christopher (NIMHD)
- Haugen, Astrid (NIEHS)
- Huerta, Mike (NLM)
- Lawler, Cindy (NIEHS)\*
- Lee, Jocelyn (OD)
- Luetkemeier, Erin (OD)
- McKaig, Rosemary (NIAID)
- Moser, Richard (NCI)
- Paltoo, Dina (OD)
- Radman, Thomas (NIDA)
- Ravichandran, Veerasamy (NIGMS)
- Reczek, Peter (OD)
- Roe, Joana (NIAID)
- Rubinstein, Yaffa (NCATS)
- Sen, Taner (OD) – DOA
- Serrano, Katrina (NCI) [F]
- Shabestari, Behrouz (NIBIB)
- Sheehan, Jerry (NLM)
- Sofia, Heidi (NHGRI)
- Williams, Carolyn (NIAID)
- Xia, Ashley (NIAID)
- Zimand, Lori (NIAID)

\* CO-leads



# Evaluating Program Success

## Some Questions:

### Standards Coordinating Center

- Did metrics for the portal and related activities indicate significant use by a broad community?

### Support for Standards Development Initiative

- Did the applications funded address a broad range of topics, appropriate stages of the standards development lifecycle, and a diverse group of stakeholders?
- Did the funded grants fill gaps in standards landscape?
- How much leveraging of existing standards occurred?
- How engaged was the community with the standards development and does the community use the standards?

### Infrastructure Support for Community-based Organizations

Did meetings enable cross-fertilization of BD2K activities and align with the broad objectives of BD2K?



# Guiding Principles

- Build on/leverage existing efforts
- Consider full life cycle
- Facilitate bottom up, community-based approaches
- Focus on standards with broad relevance to NIH
- Identifiers, provenance, versioning
- Address unique needs of different stakeholders
- Give credit/citation
- Ensure fair and open access
- Identify and coordinate activities across NIH



## Meeting Support for Community-Based Data Science Organizations: Timing and Budget

- RFA released in Summer 2016
- Prescreening of applications prior to submission
- Cooperative agreement would enable ongoing coordination with NIH interests.
- Applicants could request multiple years of support and costs up to 500,000/year
- 2 million/year set-aside
- Number of awards dependent on size and duration



# Two Initiative Activities

## Community-based Data Standards Efforts

- narrow data standards focus and addresses gaps
- supports many different kinds of technical and related work to develop data and metadata standards
- Leverages off existing efforts
- Interacts with SCC
- *Various options for funding*

- *community engagement*
- *addresses broad NIH needs (not IC specific)*

## Infrastructure for Data Science, Community-Based Groups

- broader focus (data standard or any other data science area)
- infrastructure support for the costs of bringing people together to work on data science issues
- *Possibly cooperative agreement-based funding*