Big Data Sharing Meeting - September 19, 2017

Logistics

- Location: Shirley Ryan AbilityLab (formerly Rehabilitation Institute of Chicago/RIC)
  355 East Erie Street, Chicago, IL 60610
  10th Floor, Conference Rooms A & B
- Timing: 8:00 a.m. Breakfast; Meeting: 8:30 a.m. – 3:30 p.m.

Meeting mission statement / objectives

1. To reinforce funders' pre-existing inclination toward the importance of data sharing and to present some success stories.
2. To disambiguate the confusion around data sharing today, especially differentiating "data" from "big data", and "data sharing" from "Data Commons."
3. To educate funders about options available for sharing large research datasets and to provide a preliminary plan / roadmap for setting up data sharing within your foundation and options for funders to promote or enable data sharing.
4. To provide a forum at which funders can talk among themselves about jointly sponsoring platforms for data sharing (for example data Commons that are shared and funded by multiple HRA members).
5. Conclude with commitments from membership to assess data sharing needs, in order to establish data sharing platforms and/or work with existing platforms to host member data.

Audience

Representatives of private funding organizations supporting biomedical research that results in imminently shareable data. Specifically, organizations that are already inclined toward data sharing, and have moved on to getting more details about the issues and understanding how to do it.

Assumptions about you (and your organization) as a participant on September 19

- You have already decided that you want the investigators you support to share data.
  - Some of the investigators you support generate "Big Data".
- You may not yet:
  - Have developed formal data sharing policies for the foundation.
  - Require investigators to have a data sharing plan.
  - Score applications based in part on an investigator's data sharing plan.
- You may have:

    - A lot of questions about what is data sharing.
    - Confusion about the boundaries between a technology platform vs. the policies that govern data on that platform.
    - Vague information, or even misinformation, about existing data sharing platforms.
    - Lack of understanding about the implications of sharing data vs. sharing "big" data.
- You are here to learn more, hear some ideas, but probably not make any significant decisions immediately.

Agenda

Breakfast 8:00 – 8:30 a.m.

Welcome & Introduction 8:30 – 9:00 a.m.

- *Maryrose Franko (Executive Director, Health Research Alliance) – 5 minutes*
- *Joanne Smith (President and CEO, Shirley Ryan AbilityLab) – 5 minutes*
- *Adam Levine (President, Circle of Service Foundation) – 5 minutes*
- *Robert Grossman (Frederick H. Rawson Professor of Medicine and Computer Science, University of Chicago) – 15 minutes*

What the introduction should cover:

- *What is this meeting about?*
- *What is this meeting not about?*
- *Let audience know our end goal is simultaneously education and commitment to projects that establish data sharing commons for this community.*

Session 1: What is data sharing and who has done it 9:00 – 10:15 a.m.

Goal: *Provide an overview / the landscape of data sharing initiatives that have launched within the last few years. List some tangible benefits that have resulted from data sharing.*

Maryrose Franko to introduce speakers

- *Warren Kibbe (Duke University School of Medicine)*
- *Brian Nosek (Center for Open Science)*

Goal of Session 1: Historical overview of data sharing initiatives from perspective of some people who have been involved at the policy and implementation levels.

Include a high-level overview of the concepts of open science and data. Introduce conceptual difference between "data" and "big data", and the implications of the difference (e.g. cost, long-term management, etc.). Introduce the landscape of recent data-sharing initiatives.

What the speakers should cover:

- *List well known data sharing initiative(s).*
- *Describe some tangible effects of data sharing, e.g., some clinical trials started, or CF DNA/liquid biopsy markers ID'd as a result of TCGA. Many other examples exist.*
- *Describe which barriers to data sharing are overcome by a platform/Commons, and which barriers are not.*

Break 10:15 a.m. – 10:30 a.m.

Session 2: Data sharing from multiple perspectives 10:30 a.m. – 11:30 a.m.

Goal: Using examples of data sets that were destined to be shared, describe barriers, the solution, and lessons-learned.

Martin Ferguson to introduce speakers:

- *Magali Haas (Cohen Veterans Bioscience)*
- *Kenna Shaw (University of Texas - MD Anderson Cancer Center)*

Goal of Session 2: Educate the audience about the hurdles faced to implement big data sharing platforms, from the perspective of someone responsible for generating big data and making sure they get broadly used. Session will relate experiential descriptions from two individuals who have had that responsibility.

Restate differences between "data" and "big data." Introduce what a "Data Commons" is and what that means.

What the speakers should cover:

- *What data set(s) were you involved with generating?*
- *Was data sharing part of the plan a priori?*
- *How did you do it? What were the barriers?*
- *What lessons were learned?*
- *Haas: will state that the Cohen Traumatic Brain Injury Commons has room for other brain data studies, when similar ontologies are used.*
- *Shaw: will describe carrot/stick policies to get data shared.*

Lunch 11:30 a.m. – 12:00 p.m.

Session 3: Data sharing platforms today 12:00 p.m. – 1:45 p.m.

Goal: Describe and differentiate, from a more technical standpoint, data sharing platforms. Make it known to attendees that significant development on platforms has already occurred; several are operating now and more will be soon. Disambiguate a lot of the confusing terms around data sharing platforms.

Meghan Byrne (PLOS ONE) – Moderator

- *Vincent Ferretti (Ontario Institute for Cancer Research (OICR), Cancer Genome Collaboratory)*
- *Michael Fitzsimons (University of Chicago, Genomic Data Commons – GDC Gen3)*
- *Justin Guinney (Sage Bionetworks, Synapse)*
- *Erik Lehnert (Seven Bridges Genomics, Seven Bridges Cancer Genomics Cloud)*
- *Benedict Paten (University of California, Santa Cruz, Human Cell Atlas Data Coordination Platform and UCSC Computational Genomics Platform)*
- *Anthony Philippakis (Broad Institute, Google Ventures, FireCloud)*

Goal of Session 3: To let the audience know that several data sharing platforms, including "Data Commons" and "big data sharing" type platforms, already exist or are in development. The purpose is to let the audience know they would not be starting from scratch – the technical systems they would need already exist, subject matter expertise already exists, and significant capital investments have already been made and can be leveraged.

Speakers will include representatives of several data sharing platforms, and reinforce the concept of a "Data Commons." Presenters will describe and differentiate their data sharing platforms, and provide more details around the technical, governance and policy, and funding aspects of their projects.

What the speakers should cover:

- *MB: What are data sharing platforms? What is not a data sharing platform?*
- *MB: What makes a data sharing platform a "big data" sharing platform?*
- *MB: Disclaimer(s)*
- *5 Minute intro: Align some examples from Session 1 against these differentiating features.*
    - *Talk about your solution/platform?*
        - *Technical*
        - *Governance and Policy*
        - *Funding*
- *Panel: 10 – 12 minutes each (not in any intended order)*
    - *From the panelist's perspective:*
        - *What data types?*
        - *Platform status?*
    - *From a submitter's perspective:*
        - *How do you get data into a platform?*
        - *What is the biggest challenge to getting data into a platform?*
    - *From a user's perspective – what is easy to do? And what is hard?*
    - *Describe platform functionality, ranging from simple data repository/download source, to having ready analysis tools sitting on the data, to providing a compute space where you bring your own algorithm.*
        - *As asked from the audience's perspective: what can your platform do for me?*
    - *What does "interoperability" mean in the context of "Data Commons"? How can these large-scale systems share data with each other (i.e., interoperability)?*
        - *What might this look like in the next 3-5 years?*
    - *What is a "Data Commons"? What is the difference between "data sharing" and a "Data Commons"?*
        - *A: The latter enables a synergy of discovery between shared, independently generated data sets.*
        - *Data Commons – interoperability in bulk*

Break 1:45 p.m. – 2:00 p.m.

Session 4: Establishing Data Commons 2:00 p.m. – 3:30 p.m.

Goal: Lay out a To-Do list that foundations will have to navigate to establish a place for their investigators to share data. Gather from them the questions they still have, to answer now or in follow-up meetings. Suggested next steps for meeting participants.

Salvo La Rosa (Children's Tumor Foundation) – Moderator

- *Maryrose Franko (Health Research Alliance)*
- *Robert Grossman (University of Chicago)*

Goal of Session 4: To challenge audience members to take the next steps: assessing their data needs, establishing or joining an existing Data Commons, and committing to getting their investigator's data into that Commons.

Speakers to describe the specific steps that are required to build or participate in a Data Commons, from start-up costs, to choosing a technology platform, to developing governance and embargo policies, to understanding operating costs.

What the speakers should cover:

- *Given Session 3, do we need more Commons? Why?*
  - *What should define the data/projects placed together into a Commons?*
  - *Are their gaps in today's available Commons platforms?*
- *What does a To-Do list look like (for example)?*
  - *Create a requirements check list for your organization and the data types you have.*
  - *Develop a governance policy covering what data goes in, who pays for the operating costs, what the embargo policies are, etc.*
  - *Make the data available (this is a bureaucratic issue, not technical).*
  - *Choose a technology platform: e.g., GDC Gen 3 data commons, Sage's Synapse, etc.*
  - *Choose an operator to run the platform (do it internally, choose a group to operate, etc.)*
  - *Start collecting/inventorying data. Understand and cost data transforms that might need to be undertaken.*
  - *Start to interoperate with other Data Commons.*
- *Remind audience of "multi-tenancy" possibility – i.e., there is quite possibly an already running Commons onto which you can place your data and still maintain your branding, data controls, etc.*
  - *Reference the GDC tech stack and the platforms described in Session 3*
  - *Reference and audience discussion with Haas the offer (Session 2) of having other brain disorder data sets hosted on the Cohen Commons.*
- *What are next steps?*
  - *Who would be interested in attending practical workshops on establishing Data Commons?*
  - *(How to ask for more tangible commitments?)*