BIOMEDICAL DATA MANAGEMENT AT SCALE
the rise of data commons and the quest for FAIRness
AGENDA

1. Data sharing is difficult
2. Biomedical big data and data commons
3. Metadata and FAIRness
4. The urgency of COVID-19
SHARING BIOMEDICAL DATA IS DIFFICULT
Sharing data is difficult

• Data are rarely collected for reuse
• The effort to make data reusable falls the data producer; the benefit accrues to the data consumer
• Incentives are not aligned to support data reuse
Challenges with biomedical data

- Privacy and regulatory concerns over data from human subjects
- Clinical data generally collected for treatment, not reuse
- Biologists tend not to converge on standards
Biomedical Big Data
TCGA BY THE NUMBERS

TCGA produced over 2.5 petabytes of data.

To put this into perspective, 1 petabyte of data is equal to 212,000 DVDs.

TCGA data describes 33 different tumor types and 10 rare cancers...including...

...based on paired tumor and normal tissue sets collected from 11,000 patients...using 7 different data types.
Sustaining the big-data ecosystem

Organizing and accessing biomedical big data will require quite different business models, say Philip E. Bourne, Jon R. Lorsch and Eric D. Green.

What is a Data Commons?
A Functional Definition

“Data commons are cloud-based software platforms that co-locate: 1) data, 2) computing infrastructure, and 3) commonly used software applications, tools and services to create a resource for managing, analyzing, integrating and sharing data with a community.”

"Our guiding principles for data access, use, and reuse will adhere to the Findable, Accessible, Interoperable, and Reusable (FAIR) guidelines. While the FAIR guidelines are well-accepted, to bring them to practice will require defining and adopting community-based metrics and rubrics so these can be applied to data, and other types of digital objects, hosted within or available through the Data Commons. At the same time, once FAIR metrics and rubrics are defined, these will be used to measure the level of “FAIRness” of repositories, datasets, and other digital objects. Such evaluations will inform and engage both Data Commons users and digital objects producers."
#metadata
and FAIRness
Metadata & FAIRness

Findable

Accessible

Interoperable

Reusable
Quantifying FAIRness
Researchers produce **Digital Objects**
Datasets, Tools, Workflows, and more

- Find them
- Access them
- Interoperate with them
- Reuse them

**FAIRshake**
https://fairshake.cloud/

Community-driven, semi-automated **Assessments** informed by community-accepted **metadata standards**

- **Bioschemas**
- **FAIRsharing.org**
  - standards, databases, policies
- **doi**
- **SmartAPI**

Community-driven **Metrics** for FAIRness
Commentary

FAIRshake: Toolkit to Evaluate the FAIRness of Research Digital Resources

Daniel J.B. Clarke ¹, Lily Wang ¹, Alex Jones ², Megan L. Wojciechowicz ¹, Denis Torre ³, Kathleen M. Jagodnik ³, Sherry L. Jenkins ⁴, Peter McQuilton ⁵, Zachary Flamholz ³, Moshe C. Silverstein ², Brian M. Schilder ³, Kimberly Robasky ⁶, Claris Castillo ⁴, Ray Idaaszak ⁴, Stanley C. Ahalt ⁴, Jason Williams ⁵, Stephan Schurer ⁶, Daniel J. Cooper ⁶

... Avi Ma'ayan ¹, ², ³

https://doi.org/10.1016/j.cels.2019.09.011
The goal of Common Fund Data Ecosystem CFDE is to federate data from a number of Common Fund Data Coordinating Centers (DCCs) to improve access to data derived from hundreds of studies and samples collected from thousands of human subjects. This project involves a diversity of datatypes has been generated at the genomic, expression, proteomic, metagenomic, and imaging levels, and the DCCs support a tremendous range of scientific discovery efforts.
C2M2 Level 0 defines a **minimal valid C2M2 instance**. Data submissions at this level of metadata richness will be the easiest to produce, and will support the simplest available functionality implemented by downstream applications.

C2M2 Level 1 models **basic experimental resources and associations between them**. This level of metadata richness is more difficult to produce than Level 0's flat inventory of digital file assets. As a result, Level 1 metadata offers users more powerful downstream tools than are available for Level 0.
The Urgency of COVID-19
The Chicagoland COVID-19 Commons is a consortium-driven effort to collect regional clinical and socio-economic data to drive local model efforts and advance the understanding of health disparities.
Final thoughts

• Data sharing (or enabling data access) is key to advancing biomedical research

• Data sharing is difficult; it requires
  • Active data management
  • Effective data governance
  • Adoption of standards, structured vocabulary

• Most importantly, it requires alignment of incentives
Thank you

https://chicagoland.pandemicresponsecommons.org/

Matthew Trunnell
matthew@occ-data.org
@MatthewTrunnell