Data Management in Research-Intensive Environments
A view from the National Library of Medicine

Patricia Flatley Brennan, RN, PhD
Director, NLM
May 3, 2017
My goals today

• Advance the dialogue about data management with special considerations about sustainability and open science
• Learn from you more about how the NLM can help to develop a data management program
• Identify common themes across specialty Federal libraries and research-intensive Universities, including
  • human resources
  • competencies
  • training models for the library workforce engaged with data
  • infrastructure
“The National Library of Medicine should become the epicenter of data and open science ...” (ACD report, 2015)
Crafting a vision of a library as the dynamic interplay of medicine & information
DataScience@NIH

• Opportunities
• Challenges
• Directions

Discovery!
Scientific Discovery

Where do discoveries originate?
Experiments
Computation
DATA

Findable
Accessible
Interoperable
Reusable
A New Era for NIH
Assess and mine high-value datasets

NLM Commitments to Data Science
Preserve with a Purpose
Promote Standards

LOINC  SNOMED CT  UMLS  RxNorm  VSAC
Build Tools for Discovery and Analysis
Support
Open Science
Develop new models of Dissemination
Promote training
Data-driven, Data-Producing Discovery
Data Science Challenges & Opportunities

Scalable, Sustainable Storage
Advanced Analytics
Pipeline Management
Preparing the Data Science Workforce
Where is all that data going to go?

- In the Moment
- In the local repositories
- To the cloud

...Made available for Learning by the system & for science
Distributed data repositories: Protected yet findable
Technology
Creating a distributed infrastructure that supports discovery and fine-grained data access

Remote Data Host

Data Commons

Research Repository

Who Are YOU?

Authentication Service

Registry Service

Access Control Service

What data are in here?

Do YOU have rights to this data?
• In-stream sampling and analysis
• Distributed models
• Discovery
• Beyond Statistics
• Machine/Deep learning
• Optimization
• Visualization
• Visual Analytics
• Display
Pipeline management

- Capturing data collected in the course of work processes
- Data storage
- Best practice agreements for data protection
- Computable processes governing access
Preparing the Data Science Workforce

- Data Scientists
  - Algorithm & tool building
  - Implementation specialists
- Domain Scientists
  - Incentivizing data science approaches
- New management skills
- Data Intensive clinicians
  - Interpreting data-driven reports
- Data Sophisticated Librarians
National Network of Libraries of Medicine: A Pathway that supports workforce development

8 regions; 6500 members
Summary

• Our challenges are similar, the work is complex, so we should partner to the extent possible

• Institution-based investment should
  • Acquire data in an honorable manner
  • Support analytics
  • Prepare for preservation
  • Make judicious investments in storage and compute resources

• Federal investment should:
  • Ensure sustainability and protection of high-value data sets
  • Set policies and guide standards development

• Partnerships should build curricula, leverage investments