Office of Data Science

Realizing the full potential of environmental health data to promote human well-being

Stephanie Holmgren, M.S.L.S., MBA
Charles Schmitt, Ph.D.

NTP Board of Scientific Counselors
7 December 2017
Outline

Data Science Landscape

Data Science at NIH

Data Science at NIEHS/NTP

- Guiding Principles
- Strategic Priorities

Current Initiatives

- NIEHS Data Commons
- Towards Interoperability of Data Systems
- Metadata Catalog
- Automation of Systematic Review
Data Science Landscape

Experimental
Thousand years ago
Description of natural phenomena

Theoretical
Last few hundred years
Newton’s laws, Maxwell’s equations...

Computational
Last few decades
Simulation of complex phenomena

The Fourth Paradigm
Today and the Future
Unify theory, experiment and simulation with large multidisciplinary data
Using data exploration and data mining (from instruments, sensors, humans...)
Distributed Communities

Source: https://goo.gl/yzsXJG
Data Science Landscape

High-Throughput Technologies

Data- and Knowledge-Driven Discovery

Computational Technologies & Infrastructure

FAIR Data
Findable, Accessible, Interoperable, Reusable
What is Data Science?

- an interdisciplinary field
- focused on the research, development, and application of the methods, processes, and systems needed to extract knowledge from data.
- emphasis is on the entire process involved in making data FAIR to generate knowledge.

Source: https://goo.gl/NGVayr
Data Science at NIH

- NIH Public Access and Data Sharing Policies
- Big Data 2 Knowledge (BD2K)
  - NIH Data Commons / NCI Genomics Data Commons
  - BD2K Centers of Excellence
  - BioCADDIE and DataMed
  - Standards Development
  - Software and Analytical Methods
  - ERuDItE training resource
- NLM RFI *Next Generation Data Science Challenges in Health and Biomedicine*
Data Science at NIEHS/NTP
Guiding Principles

- Data and knowledge-driven approaches are fundamental to modern interdisciplinary scientific discovery and team science.

- Building a data science community of practice is as important as building the data science infrastructure.

- Federally-funded data needs to be FAIR (findable, accessible, interoperable, reusable).

- Proper incentives are critical to ensure data are appropriately annotated and shared for the advancement of research.
The Office of Data Science provides leadership and support to NIEHS and the broader environmental health community to enable the discovery, access, and use of data needed to advance environmental health research, policy, and decision-making.
Strategic Priorities

1. Establish a governance framework that ensures availability, usability, integrity, and security of NIEHS data.

2. Promote foundational research in data science methods based on evolving needs of the NIEHS community.

3. Identify and advance the application of existing novel data-and knowledge-driven methods and technologies.

4. Provide data and knowledge systems that support FAIR principles for environmental health data.

5. Increase engagement with community to improve use and utility of data for environmental health science research.

6. Foster a community of practice in environmental health data science.
Strategic Priorities

Data Governance

*Establish a governance framework that ensures availability, usability, integrity, and security of NIEHS data*

- Implement and lead a formal data governance body
- Creation of governance framework to develop policies, resolve inconsistencies/gaps, and set accountability for implementation and adherence
- Engage with NIH data governance processes to represent environmental health needs
Methods - Research
Promote foundational research in data science methods based on evolving needs of the NIEHS community

- Identify exemplar questions to inform EHS data science needs
- Catalog current and projected data science needs
- Engage with methods development community
Methods - Applied
Identify and advance the application of existing novel data- and knowledge-driven methods and technologies

- Identify existing and novel technologies to enable FAIR+ data
- Evaluate and prototype data science technologies
- Promote and advise adoption of best methods for NIEHS applications
Data Cyberinfrastructure

*Provide data and knowledge systems that support FAIR principles for environmental health data*

- Lead development of a NIEHS Data Commons
- Develop core services that support NIEHS data systems and data-centric tools
- Coordinate establishment of a data cyber-infrastructure to support collaborative data-driven research
- Engage with external data cyberinfrastructure efforts towards a sustainable, international system for environmental health data
Engagement

Increase engagement with community to improve use and utility of data for environmental health science research

- Communicate data and knowledge management related governance policies and processes to NIEHS community
- Inform and promote adoption of best practices, methods, and solutions for data-driven research
- Ensure that ODS services are meeting data science needs of NIEHS staff
- Engage with external groups to share, learn, and collaborate
Community of Practice

*Foster a community of practice related to environmental health data science*

- Increase the quantitative skills of environmental health researchers
- Raise awareness of the value of data management practices
- Nurture a data-oriented workforce within the EHS community
- Create a network of data science expertise
## ODS Staffing

<table>
<thead>
<tr>
<th>Current Staff</th>
<th>Proposed Staff</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interim Director</td>
<td>5 Federal Staff</td>
</tr>
<tr>
<td>6 contractors</td>
<td>Director</td>
</tr>
<tr>
<td>• Senior Advisor</td>
<td>• Data Architect</td>
</tr>
<tr>
<td>• Data Systems Engineer</td>
<td>• Data Curator</td>
</tr>
<tr>
<td>• 2 Software Developers</td>
<td>• Data Management Specialist</td>
</tr>
<tr>
<td>• Data Curator</td>
<td>• Data Scientist</td>
</tr>
<tr>
<td>• Scientific Program Manager</td>
<td>6 contractors</td>
</tr>
<tr>
<td></td>
<td>Fellow</td>
</tr>
<tr>
<td></td>
<td>2 Summer Interns</td>
</tr>
</tbody>
</table>
Current Initiatives
<table>
<thead>
<tr>
<th>FAIR PRINCIPLES</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Findable</td>
<td>A data object should be uniquely and persistently identifiable.</td>
</tr>
<tr>
<td>Accessible</td>
<td>Data is accessible by authorized users (human and machine) through a well-defined protocol.</td>
</tr>
<tr>
<td>Interoperable</td>
<td>(Meta) data assigned to the data object is syntactically parse-able and semantically machine accessible.</td>
</tr>
<tr>
<td>Reusable</td>
<td>Data objects must comply with the above three principles and sufficiently documented to allow integration/linkage with other data sources.</td>
</tr>
</tbody>
</table>
Common data publication practices in place:
- submission to dbGAP, GEO, clinicaltrails.gov, …

CEBS: provides FAIR management for NTP data

But, many places where internally we can do better

ODS: several new initiatives to advance FAIR practices within NIEHS
NIEHS Data Commons - a system for:

- Researchers and core labs to access, find, and share research data and metadata
  - For management of data that isn't designated for CEBS

- IT staff to improve data and storage management, without impacting users

- Foundation for integration or federation with external data systems
Early 2018 release
NIEHS Data Commons – the back end

- Metadata
  - Software Interfaces
  - Command Line Tools
  - Web Interfaces
  - Graphical User Interfaces

- Core Data Services: search, access, archive, ...
- Automated processes: lab & informatics workflows, ...

- Virtual File System (iRODS)
  - Storage (DDN)
  - Storage (Netapp)
  - Lab server

- NIH Commons
  - Monitoring, Reporting, Testing
  - Training and Support
  - Future
NIEHS Data Commons – Genomic Data

- Registration of genomic data from core lab with commons
- Extraction of metadata
  - sample details, run details, requestor, etc
Towards a Global Data Fabric

- National Supercomputing Centers
  - NCSA
  - SDSC
  - TACC

- Data Sharing & Archiving Cyberinfrastructures
  - DataONE
  - The NIH Data Commons
  - NCI Data Archive
  - The National DATA SERVICE
  - GAGH

- European Data Fabric
  - EUDAT

- Public Clouds
  - AWS
  - Azure
  - Google Cloud

- CEBS
- NIEHS Data Commons

- Common access to/from other cyberinfrastructure resources
- Policy controls
- Metadata support
- Automation of data processes
Towards Interoperability of Data Systems

Grand Challenge Problem

Research Questions

Data integration = lots of hard & problematic work

- Chemical Databases
- Clinical Databases
- Imaging
- BioAssays
- Genomics
- Chemical Effects Databases
- Personal Health Records
- Expression
- Metabolomics
- Contextual Databases
- Pathways
- AOPs
- Ontologies
- Literature
Towards Interoperability of Data Systems

Web-based Application Programming Interface (API)

ASSETS

APIs

DEVELOPERS
Use APIs to create apps

SOFTWARE & APPLICATIONS
powered by APIs

END USERS

Towards Interoperability of Data Systems

Web-based Application Programming Interface (API)

ASSETS → APIs → DEVELOPERS
Use APIs to create apps
SOFTWARE & APPLICATIONS
powered by APIs → END USERS

In progress...

CEBS
ICE
NIEHS Data Commons

Treatments
Treatment Findings
Data Sets

What additional access would be valuable?

The catalog provides a centralized resource for metadata services for environmental health metadata.
Can we automate extraction of information from research articles using natural language processing?

- Journal Article
  - Studies
    - Experiments
      - Treatment/Animal Groups
        - Type
        - Animal Information
        - Exposures
        - Doses
        - Measures
        - Endpoints
        - Assays
  - Results
  - Risk of Bias

Can we extract these items and relations?
NIST Text Analytics Conference 2018 Challenge

Nineteen-day-old female mice weighing 7-9 g, randomly selected for each treatment group, received CdCl$_2$ (Sigma, St. Louis, MO, USA) dissolved in sterile phosphate-buffered saline (PBS) at 5 (n = 6), 50 (n = 6), and 500 µg/kg body weight (BW) (n = 5) per day for 3 consecutive days via subcutaneous (SC) injections.

Control animals (n = 5) were injected with sterile PBS.

We used EE_2 (Sigma) dissolved in corn oil (50 µg/kg BW, SC) as a positive control (n ≤ 5).

- Preparing training & test sets
- 2018 Challenge – extraction of experimental methods
Techniques will feed into machine assisted data extraction tools
Thank you for your time!

Feedback, questions, concerns?