NTP
National Toxicology Program

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

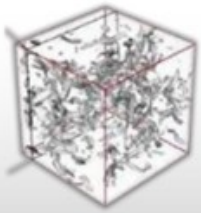
- Guiding Principles
- Strategic Priorities

Current Initiatives

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

Data Science Landscape

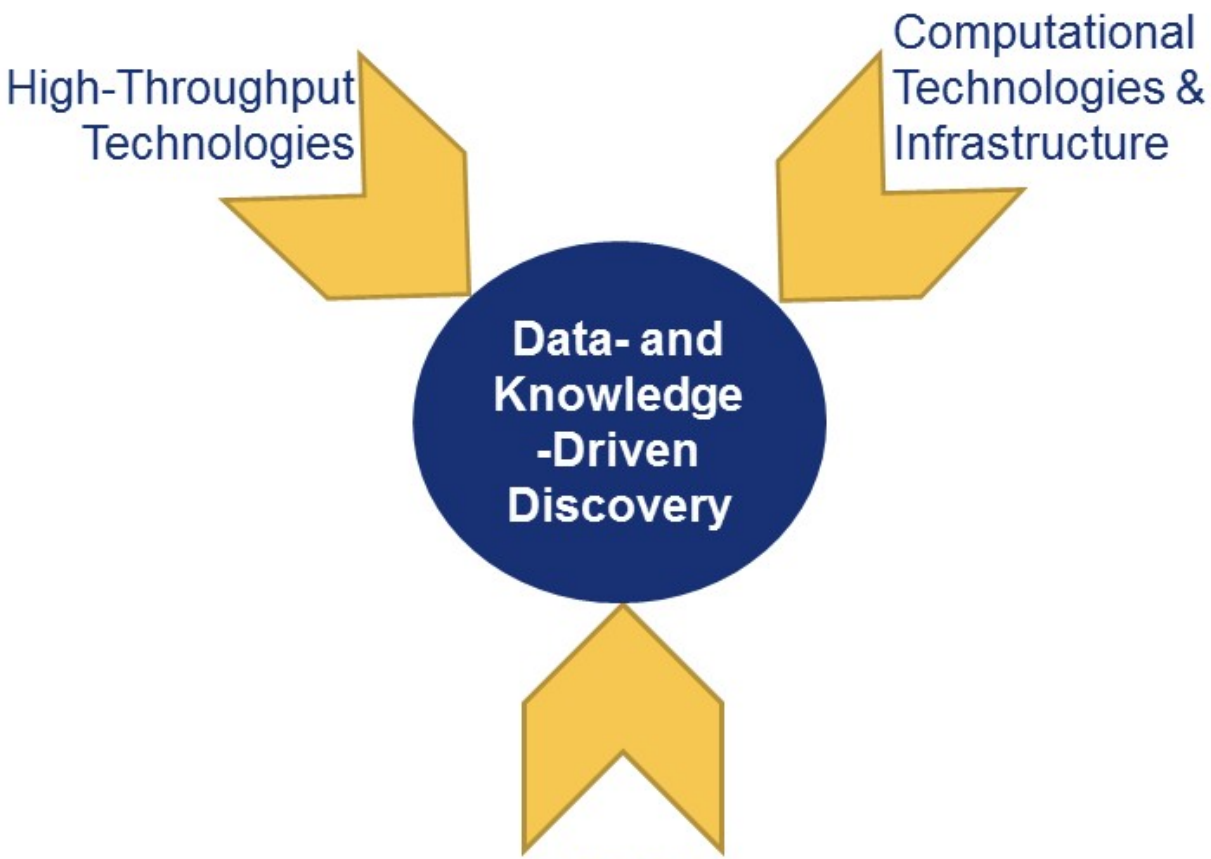
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$



Experimental	Theoretical	Computational	The Fourth Paradigm
Thousand years ago	Last few hundred years	Last few decades	Today and the Future
Description of natural phenomena	*Newton's laws, Maxwell's equations...*	*Simulation of complex phenomena*	*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

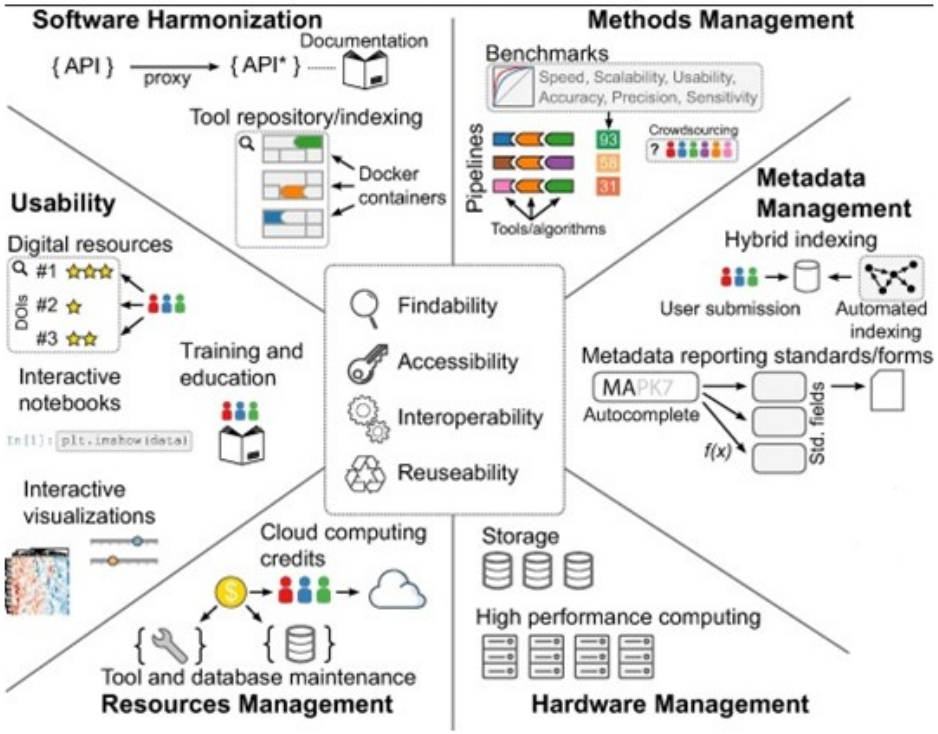
Data Science Landscape



Data Science Landscape

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.



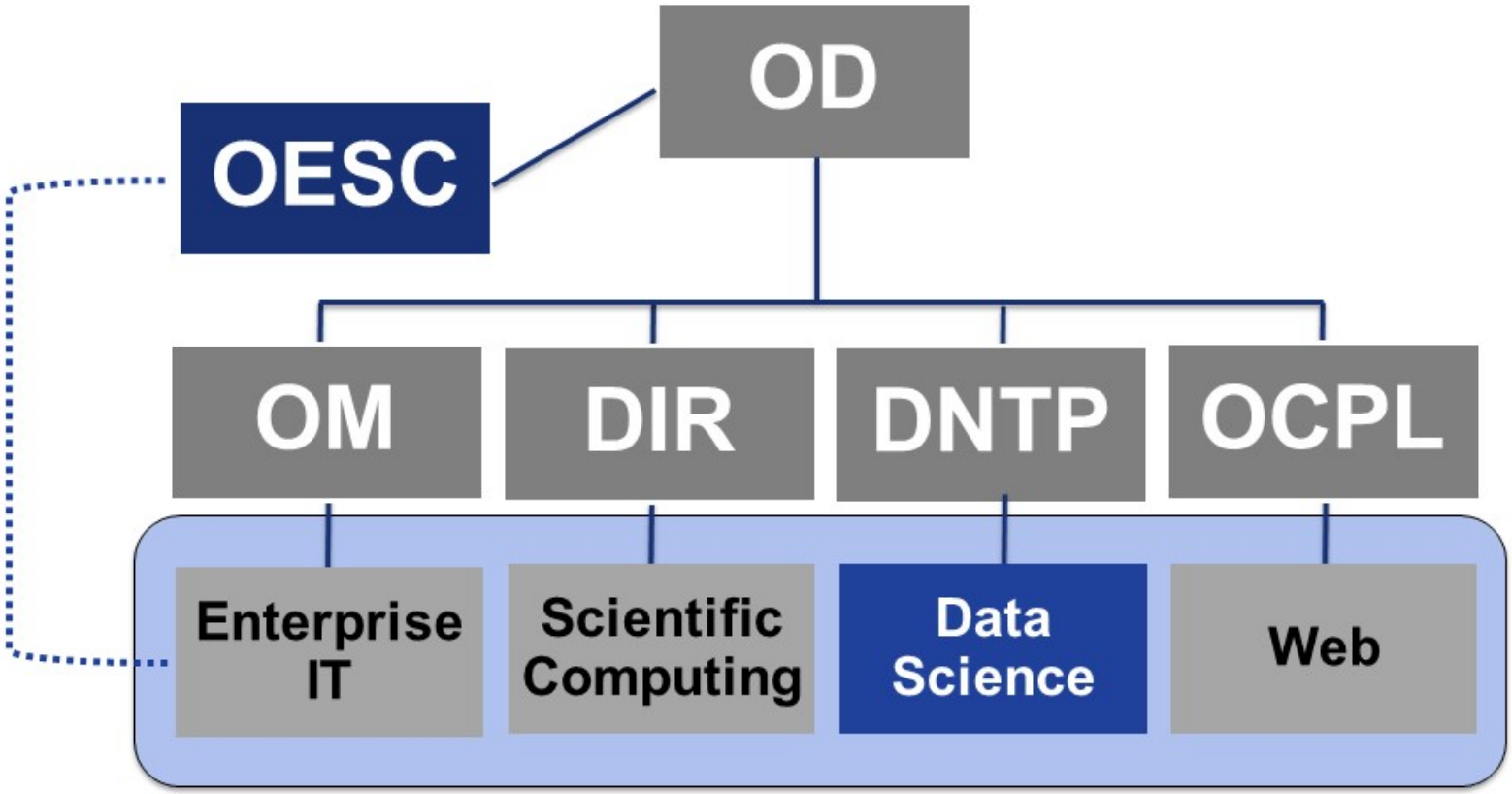
Source:
https://goo.gl/NGVayr

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

Data Science at NIEHS/NTP



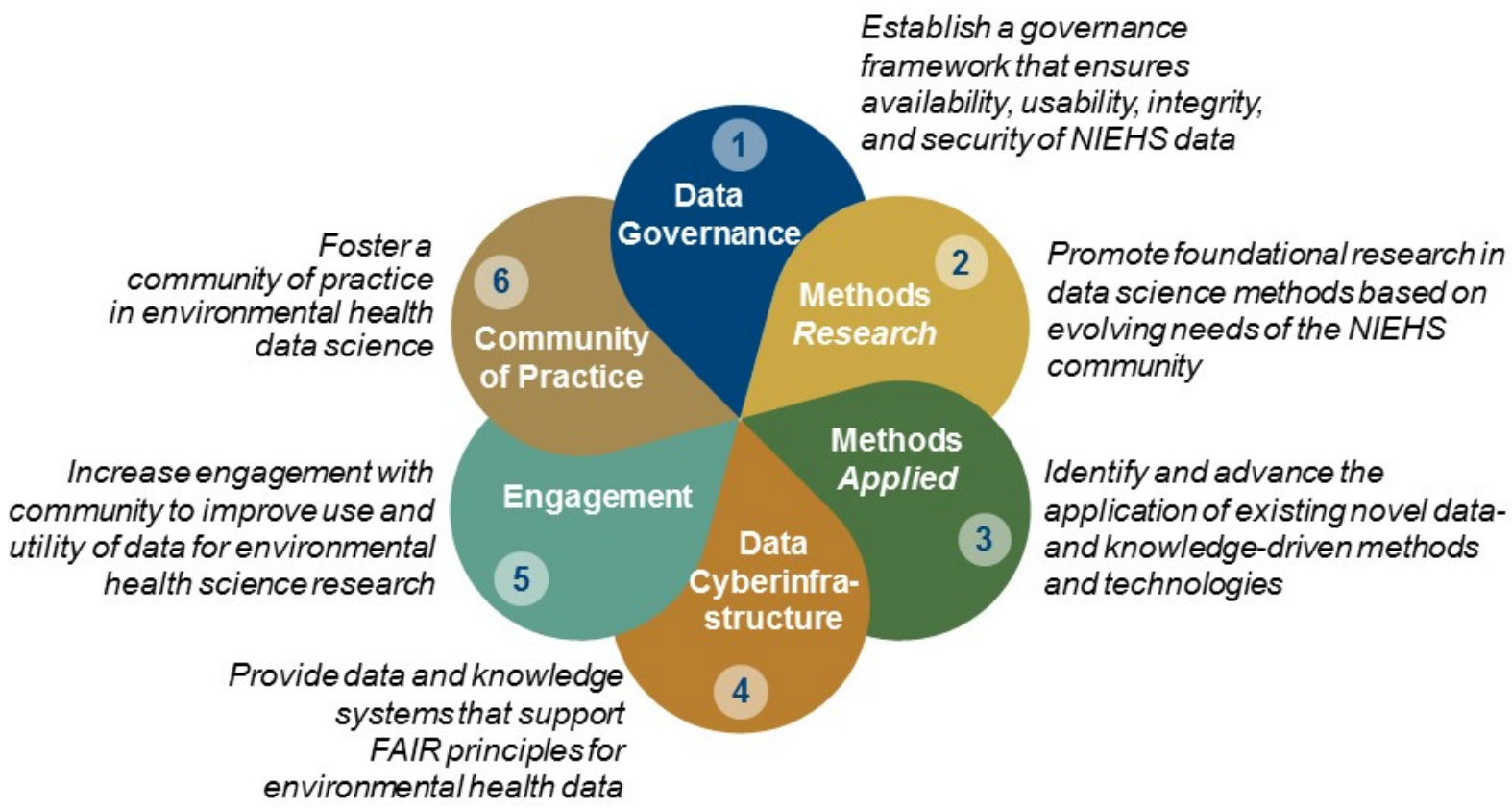
Guiding Principles

- 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.

ODS Mission

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



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

Strategic Priorities

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

Strategic Priorities

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

Strategic Priorities

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

Strategic Priorities

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

Strategic Priorities

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

Current Staff	Interim Director 6 contractors • Senior Advisor • Data Systems Engineer • 2 Software Developers • Data Curator • Scientific Program Manager
Proposed Staff	5 Federal Staff • Director • Data Architect • Data Curator • Data Management Specialist • Data Scientist 6 contractors Fellow 2 Summer Interns

Current Initiatives

State of FAIRness at NIEHS

FAIR PRINCIPLES	
Findable	A data object should be uniquely and persistently identifiable.
Accessible	Data is accessible by authorized users (human and machine) through a well-defined protocol.
Interoperable	(Meta) data assigned to the data object is syntactically parse-able and semantically machine accessible.
Reusable	Data objects must comply with the above three principles and sufficiently documented to allow integration/linkage with other data sources.

State of FAIRness at NIEHS

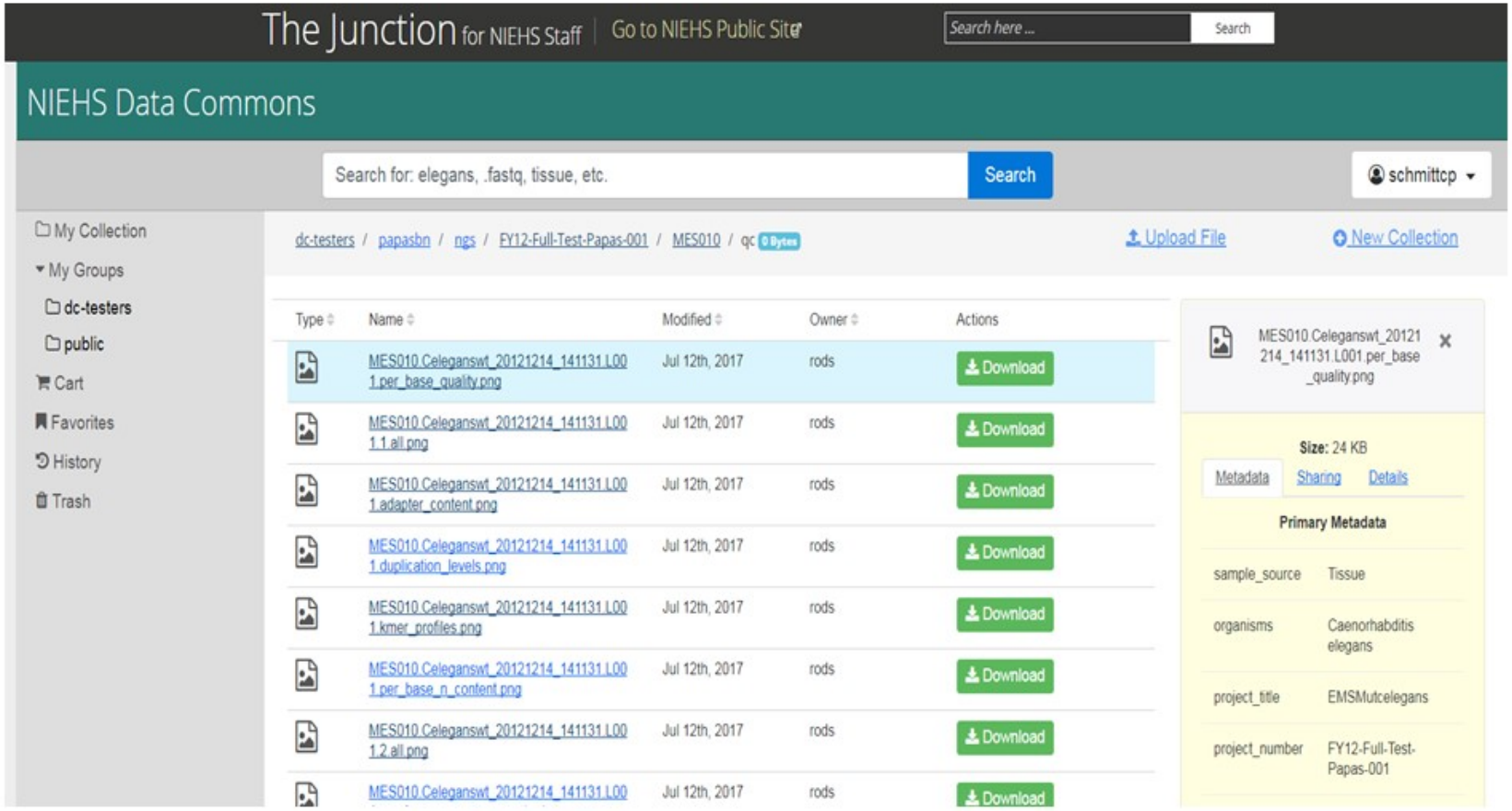
- Common data publication practices in place:
  - submission to dbGAP, GEO, clinicaltrails.gov, …
- CEBS: provides FAIR management for NTP data
- But, many places where internally we can do better
- *ODS: several new initiatives to advance FAIR practices within NIEHS*

NIEHS Data Commons

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

NIEHS Data Commons – the front door view



Early 2018 release

NIEHS Data Commons – the back end

Software Interfaces

Command Line Tools

Web Interfaces

Graphical User Interfaces

Metadata

Core Data Services: search, access, archive, …

Automated processes: lab & informatics workflows, …

Virtual File System (iRODS)

Storage (DDN)

Storage (Netapp)

Lab server

…

NIH Commons

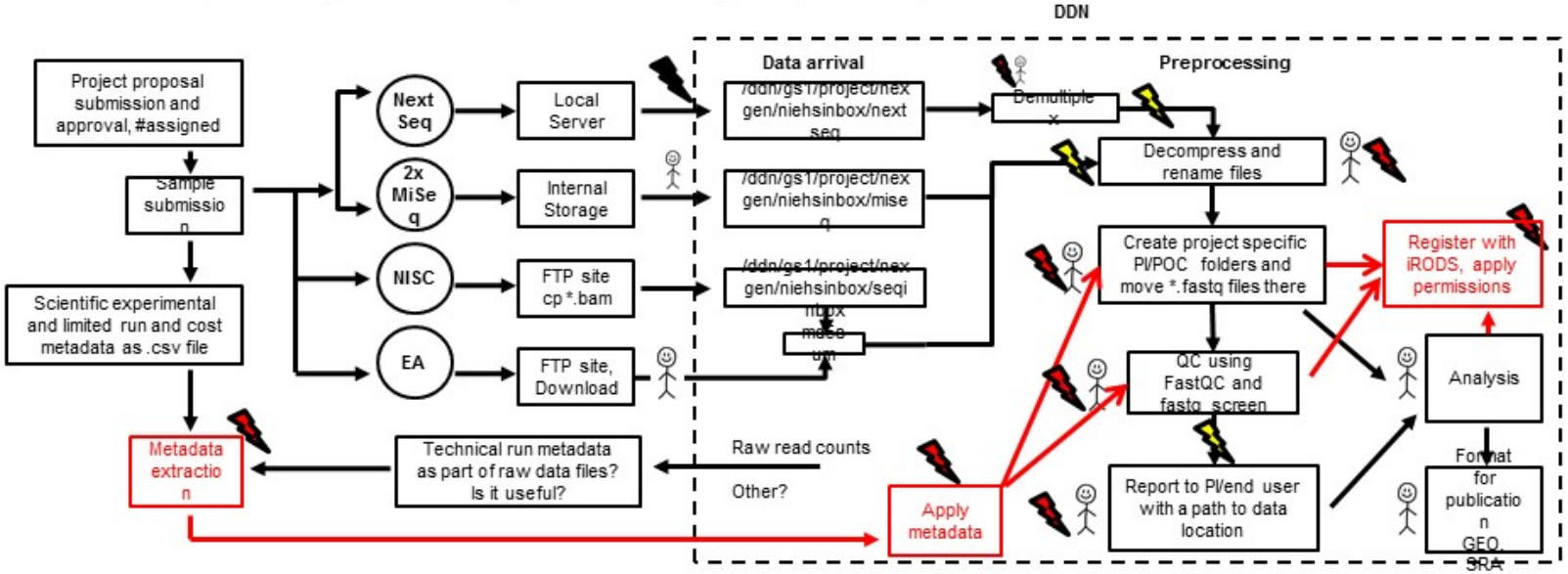
future

Monitoring, Reporting, Testing

Training and Support

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 San Diego Supercomputer Center

TACC

Public Clouds

aws

Azure

Google Cloud

Data Sharing & Archiving Cyberinfrastructures

The NIH Data Commons

DataONE

The Dataverse Project

globus

NCI Data Archive

CYVERSE

The National DATA SERVICE

GAGH

European Data Fabric

EUDAT

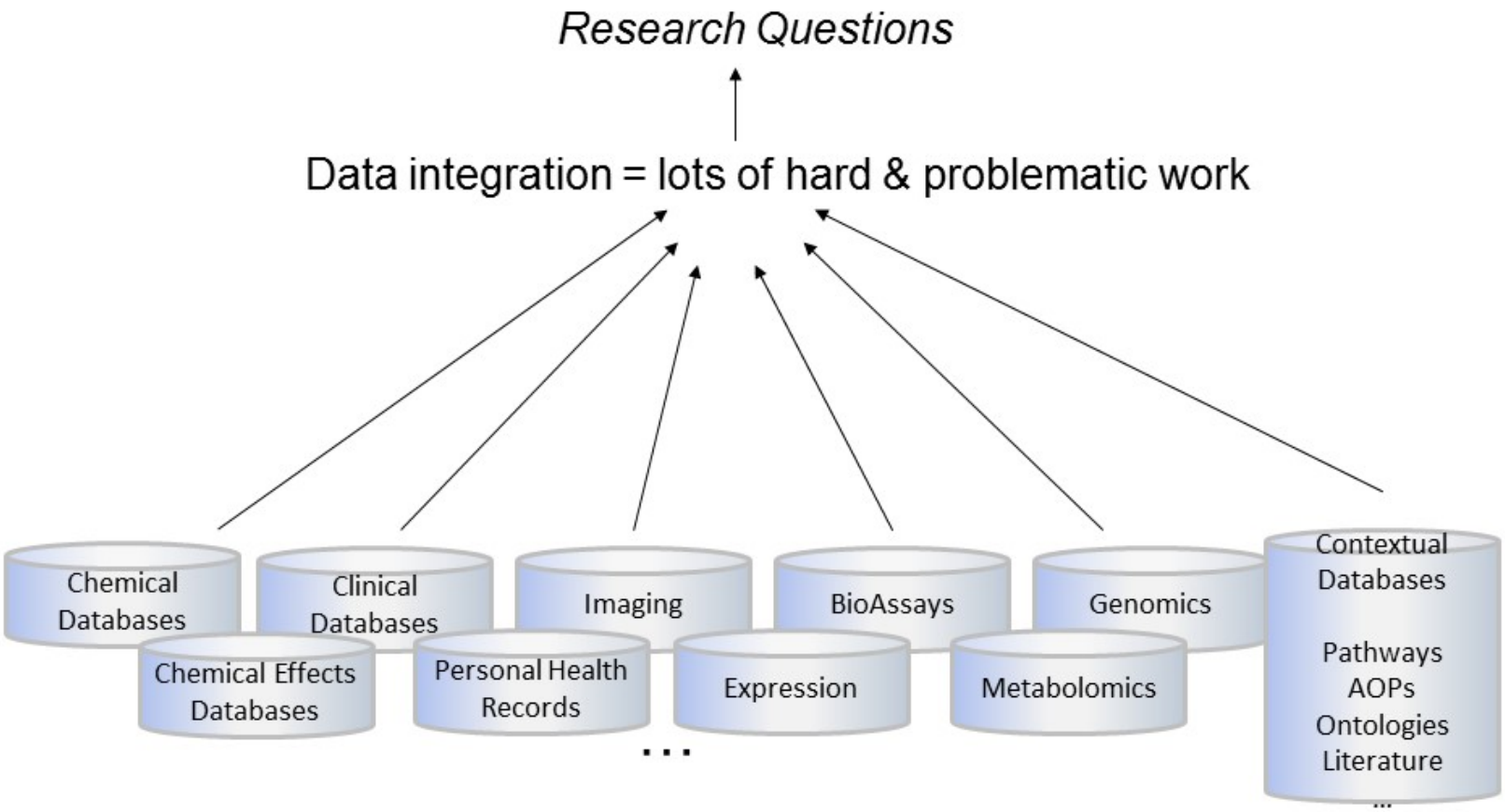
- Common access to/from other cyberinfrastructure resources
- Policy controls
- Metadata support
- Automation of data processes

CEBS

NIEHS Data Commons

Towards Interoperability of Data Systems

Grand Challenge Problem



Towards Interoperability of Data Systems

Web-based Application Programming Interface (API)

Image source: https://www.upwork.com/hiring/development/intro-to-apis-what-is-an-api/

Towards Interoperability of Data Systems

Web-based Application Programming Interface (API)

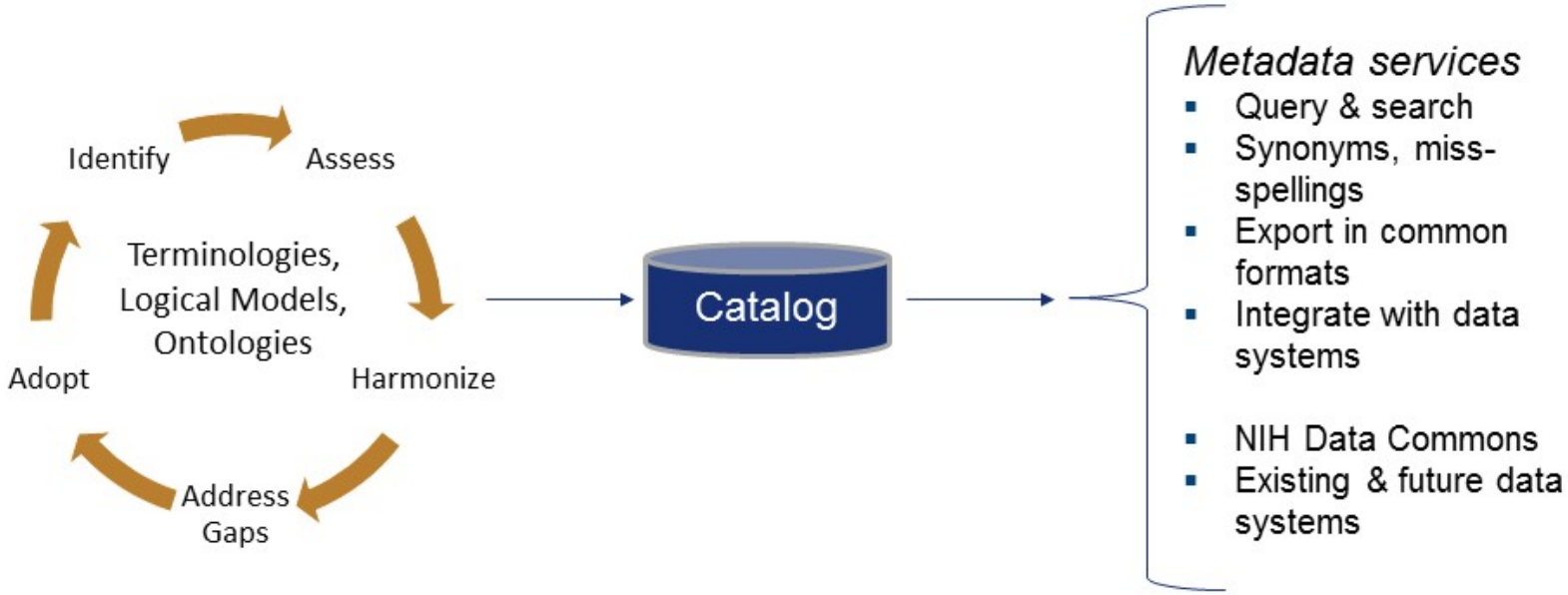
In progress…

CEBS	Treatments
ICE	Treatment Findings
NIEHS Data Commons	Data Sets

What additional access would be valuable?

Image source: https://www.upwork.com/hiring/development/intro-to-apis-what-is-an-api/

NIEHS Metadata Catalog



- The catalog provides a centralized resource for metadata services for environmental health metadata

Systematic Review

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

7 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 ug/kg body weight (BW) (n = 5) per day for 3 consecutive days via subcutaneous (SC) injections.

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

9 We used EE_2 (Sigma) dissolved in corn oil (50 ug/kg BW, SC) as a positive control (n <= 5).

Type | Su

- AgeOfDose
- AgeUnit
- DoseAmount
- DoseDuration
- DoseUnit
- Endpoint
- Exposure
- Group
- GroupSize
- Measure
- Species
- Strain
- Vehicle

- Preparing training & test sets
- 2018 Challenge – extraction of experimental methods

Systematic Review

- **Techniques will feed into machine assisted data extraction tools**

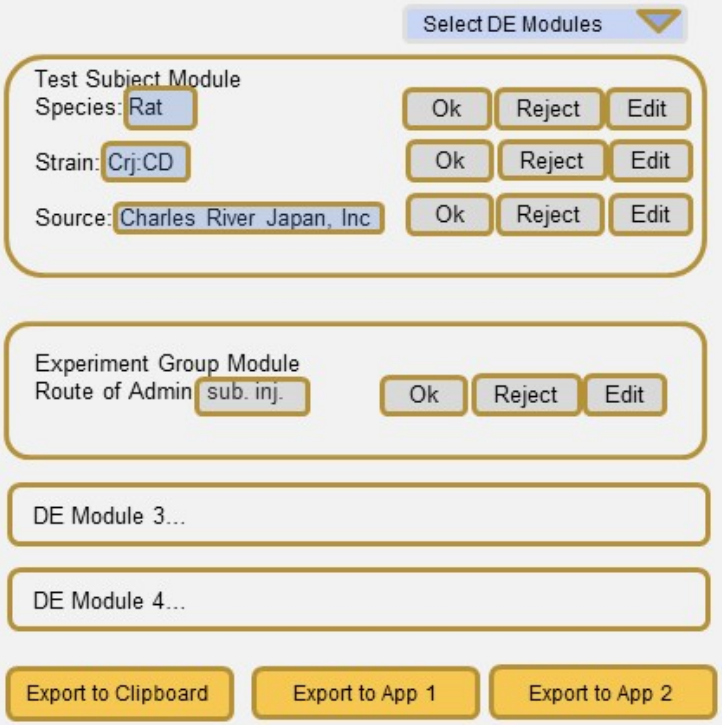
the positive control wells treated with natural ligands (1 nM of 17β-estradiol) ordinary showed maximum response and it showed well reproducibility. Description of PC50 and PC10 is illustrated in Fig. 1.

2.6. Animals

Crj:CD (SD) rats at post-natal day (pnd) 10 and dams were purchased from Charles River Japan, Inc. (Shiga, Japan). Dams and pups were kept in polycarbonate pens until weaning. All rats were weaned at pnd 17 and then housed individually in stainless steel, wire-mesh cages during the study. The immature rats were weighed, weight-ranked and assigned randomly to each of the treatment and control groups. Each group consisted of six rats. Body weights and clinical signs were recorded on a daily basis throughout the study. Rats were provided with tap water and a commercial diet (CRF-1, Oriental Yeast Co., Tokyo, Japan) ad libitum before weaning and with water automatically and a commercial diet (MF, Oriental Yeast Co.) ad libitum after weaning. The animal room was maintained at a temperature of 23 ± 2 °C, a relative humidity of 55 ± 5% and was artificially illuminated with fluorescent light on a 12-h light/dark cycle (06:00–18:00 h). All animals were cared for according to the principles outlined in the guide for animal experimentation prepared by the Japanese Association for Laboratory Animal Science.

2.7. Animal study design

The 21 chemicals, i.e. all of those mentioned above except for dibutyl phthalate and ethynyl estradiol, were injected subcutaneously on the dorsal surface at doses of 2, 20 and 200 mg/kg from pnd 20 to pnd 22, i.e. for 3 days. The high dose was selected on the basis of the previous uterotrophic assay using bisphenol A, in which the uterine response was clearly detected at a dose of 160 mg/kg per day injected subcutaneously (Yamasaki et al., 2000). On the other hand, doses of dibutyl phthalate or ethynyl estradiol were 0, 40, 200 and 1000 mg/kg per day or 0, 0.2, 2 and 20 μg/kg per day, respectively. These doses were based on the results of preliminary studies. The concentration and stability of each chemical was confirmed. The volume of olive oil contained in each chemical solution was 4 ml/kg for subcutaneous injection. A vehicle control group given only olive oil was also established. The animals were killed approximately 24 h after the last ad-



Thank you for your time!

Feedback, questions, concerns?