National Institutes of Health
Office of Data Science Strategy

NIH Data Science Strategic Plan

Belinda Seto, Ph.D.
Deputy Director

Data Science in the next 5 years

- Improve Capabilities to Sustain the NIH Policy for Data Management and Sharing
- Develop Programs to Enhance Human Derived Data for Research
- Provide New Opportunities in Software, Computational Methods, and Artificial Intelligence
- Support for a Federated Biomedical Research Data Infrastructure
- Strengthen a Broad Community in Data Science

Goal 1

Capabilities to Sustain the NIH Data Management and Sharing Policy

Challenges

- Need for the generation of FAIR Data in a manner that will foster greater sharing and the integration of scientific results
- Need for cost effective strategies for sustainable, secure, and accessible biomedical data repositories and knowledgebases

Objectives to Address Challenges

1) Support the biomedical community to manage and share data
2) Enhance FAIR data and greater data harmonization
3) Strengthen NIH's data repository and knowledgebase ecosystem

NIH's Support for FAIR & TRUST Repositories

- Enhancement and Management of Established Biomedical Data Repositories and Knowledgebases (**PAR-23-237**)
- Early-stage Biomedical Data Repositories and Knowledgebases (**PAR-23-236**)

1. Fill a scientific need or gap
2. Encourage adoption of good data management practices
3. Engage the research community to contribute and use data
4. Govern data life-cycle and preservation

Impact of 2 NOFOs in 2020 to 2023

21	awards
2	IDeA States
7	NIH ICOs
Alcohol Research, Virus Taxonomy, Vaccine Information, Chemotherapy Drugs, Drosophila, Human Pathways, GWAS, Neurotrauma	Science focus

Immune Data

Challenges: lack of standardization and consistent metadata across studies

Immune Data is a **unified metadata format** that enables search across 5 immune repositories at NIAID

- ImmPort
- ImmuneSpace
- ITN TrialShare
- The Immunological Genome Project (ImmGen)
- The Immune Epitope Database and Analysis Resource (IEDB)

https://immunedata.org/repository_list.php

Goal 2

Enhance Human Derived Data for Research

Challenges

- Need for acquisition and protection of data obtained from electronic health records, and other real-world data, that preserves privacy and enhances participant consent
- challenges in data quality, privacy and confidentiality, policy, regulatory, and ethical issues associated with healthcare and administrative data
- need to better understand the ethical, legal, and social implications of data linkage

Objectives to Address Challenges

1) Improve access to and use of clinical and real-world data
2) Adopt health IT standards for research
3) Enhance the adoption of social and environmental determinants of health for health equity

Extracting Electronic Health Record Data For Research Use

- Exchanging data between health systems
  - Fast healthcare interoperability resources, FHIR®
  - Data standards and terminology
- Common data elements

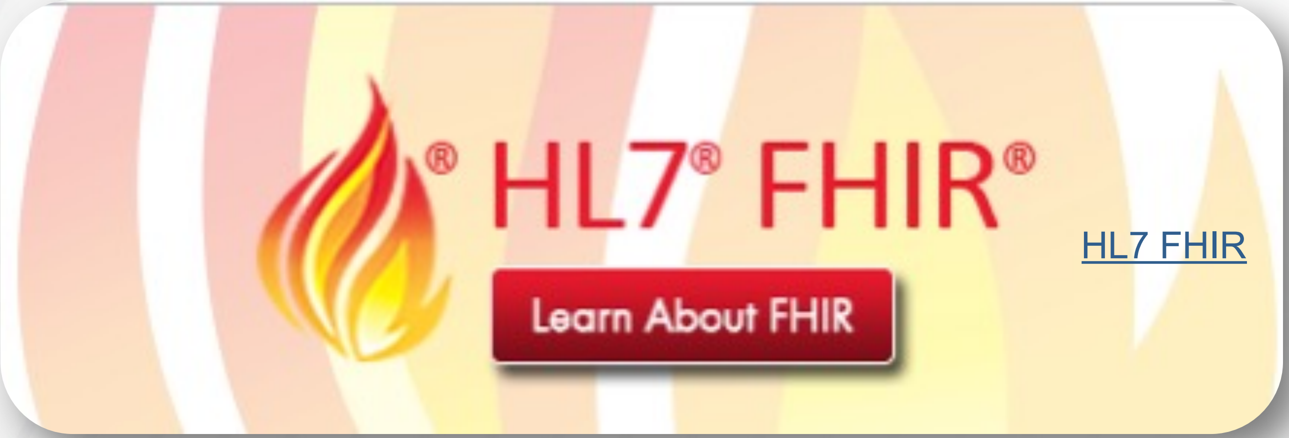
Background

- GUIDE notice, 2019:
    - “encourage NIH researchers to explore the use of the Fast Healthcare Interoperability Resources (FHIR®) standard to capture, integrate, and exchange clinical data for research purposes and to enhance capabilities to share research data.”
- The National Coordinator for Health Information Technology (ONC) has finalized a new rule to support seamless and secure access, exchange, and use of electronic health information.
    - Specifically, by 2022, health care industry is required to adopt standardized APIs by using the FHIR standard to share patient data.

What is FHIR®?

- Fast Healthcare Interoperability Resources (FHIR®)
  - A standard owned and maintained by Health Level 7 International (HL7®)
  - A way of transmitting healthcare data in a standardized way among independent systems

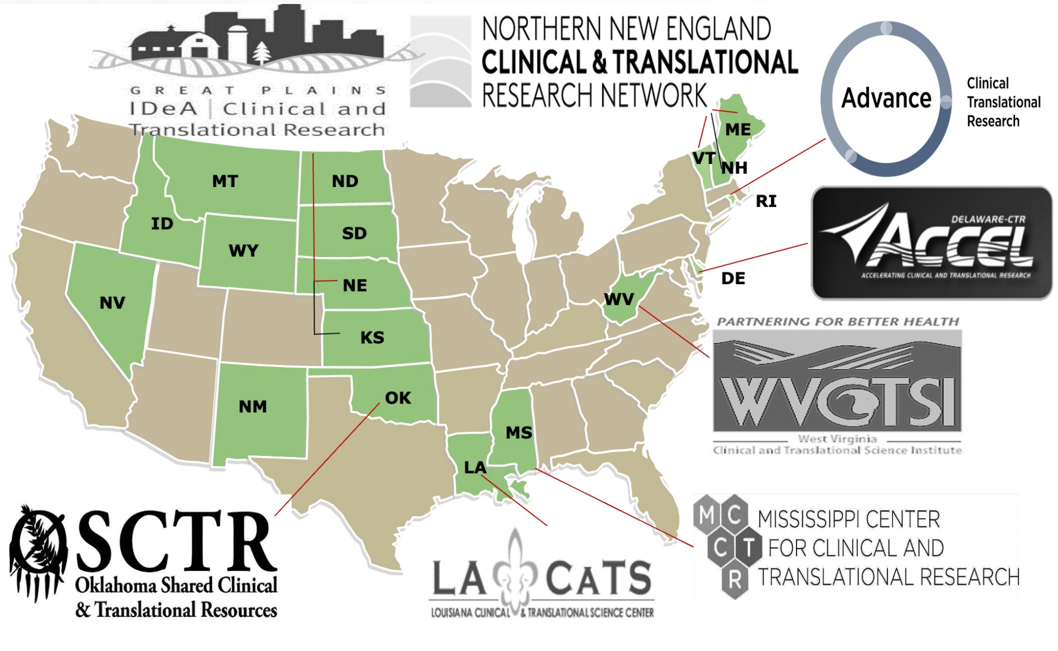
Fast
Healthcare
Interoperability
Resources



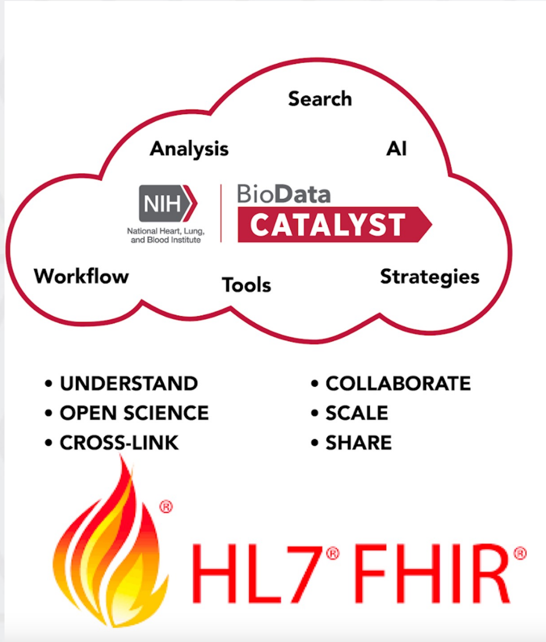
HL7 FHIR

Fast Healthcare Interoperability Resources

ODSS and NIGMS are currently conducting FHIR training at 8 Institutional Development Award Networks for Clinical and Translational Research (IDeA-CTR) institutions



IDeA-CTR FHIR Training



In FY22, ODSS supported NIH IC efforts in FHIR-enabled exchange of clinical data across systems.

- NHLBI's FHIR capabilities allow for FHIR-enabled indexing of clinical studies. This index is invaluable for searching for data and allows researchers to analyze data from electronic health records (EHR)

Data is the Payload of FHIR Transport

- For meaningful use of data, they need to be:
  - Standardized
  - Tagged with terminology definition and codes
  - Captured as common data elements

United States Core Data for Interoperability (USCDI)

USCDI

- Standardized set of health data classes and constituent data elements for nationwide, interoperable health information exchange

Appendix A: USCDI v1 Summary of Data Classes and Data Elements

Assessment and Plan of Treatment

Care Team Members

Clinical Notes
- Consultation Note*
- Discharge Summary Note*
- History & Physical*
- Imaging Narrative*
- Laboratory Report Narrative*
- Pathology* Report Narrative
- Procedure Note*

Goals
- Patient Goals

Health Concerns

Immunizations

Laboratory
- Tests
- Values/Results

Medications
- Medications
- Medication Allergies

Patient Demographics
- First Name
- Last Name
- Previous Name
- Middle Name (including middle initial)
- Suffix
- Birth Sex
- Date of Birth
- Race
- Ethnicity
- Preferred Language
- Address*
- Phone Number*

Problems

Procedures

Provenance
- Author*
- Author Time Stamp
- Author Organization

Smoking Status

Unique Device Identifier(s) for a Patient's Implantable Device(s)*

Vital Signs
- Diastolic blood pressure
- Systolic blood pressure
- Body height
- Body weight
- Heart Rate
- Body temperature
- Pulse oximetry
- Inhaled oxygen concentratio n
- BMI percentile* per age and sex for youth 2-20
- Weights for age* per length and sex
- Occipital-frontal* circumference for

* elements updated or added in the 2019-revised version of US Core Data Interoperability specification.

[HL7 FHIR® US Core Implementation Guide]

Common Data Elements and the NIH CDE Repository

- CDEs are standardized, defined questions paired with a set of specific allowable responses.
  - CDEs are the **foundation for interoperability among data systems.**
- Enable sharing and comparing data systematically across different sites and studies.
- Currently the NIH CDE Repository hosts 23,065 CDEs from 18 collections.
  - Two collections (COVID-RADx and NLHBI CONNECTS, organ support) are labeled as NIH-endorsed CDEs.

https://cde.nlm.nih.gov/guides

USCDI Terminology Standards

- LOINC – Logical Observation Identifiers Names and Codes
- SNOMED CT – Systematized Nomenclature of Medicine Clinical Terms
- CDC ISS – CVX – Vaccines Administered
- National Drug Code (NDC)
- RxNorm – Medications
- HL7 Version 3 – Value sets
- OMB race & ethnicity
- HCPCS, CPT-4, ICD-10

Use of CDEs Promotes Data…

1. ***…Interoperability*** – with other data for re-use across informatics platforms, promoting **re-analysis, meta-analysis, and collaboration**
2. ***…Quality*** –to allow data aggregation across sites and projects, **increasing statistical power** and making data **AI-ready**
3. ***…Reproducibility*** – for rigorous **comparison across studies** or sites with apples-to-apples validity
4. ***…Efficiency of collection*** – use of "off-the-shelf" data elements **reduces time and costs** to investigators
5. ***…Efficiency of use*** – effortful data harmonization not needed

CDE use will add value to the NIH Policy on Data Management and Sharing

NCI Clinical Trials CDEs

- Selected CDEs are bundled into categories that are collected across all studies, non-cancer type specific
- Category example: height
  - Preferred question text: What was the result of the height measurement?
  - CDE Public ID 6606195
- CDEs in 27 Categories covering both clinical and scientific domains for, e.g. Adverse Events, Cytogenetics, Molecular analysis, Medications, Pathology, Protocols, Response/Therapies, Treatment, Tumor Markers
- Impact: Standards of Care, Improved Testing and Treatment, New standards for for FDA approval, and Biomarker Validation

Stay tuned for a RFI on CDEs to be published soon!

Goal 3

New Opportunities in Software, Computational Methods, and AI

Challenges

- Emergence of innovations in trustable artificial intelligence (AI) approaches that reduces bias and risks
- Multi-dimensional data integration remains a significant challenge for biomedical and behavioral research

Objectives to Address Challenges

1) New opportunities to improve biomedical AI analysis
2) Develop cutting edge software technologies
3) Support FAIR software sustainability

NIH's Support for Software to Enhance Open Science

- Enhance software engineering of valuable scientific tools
- Encourage new collaborations between biomedical and clinical scientists and software engineers
- Make research tools reliable and sustainable across multiple computing environments
- Improving reuse and effectiveness of NIH-developed software for open science

Impact of 4 NOFOs in 2020 to 2023

$22.2M ODSS funds

126 awards

10 IDeA States

19 NIH ICOs

Examples of Software to Enhance Open Science

Software Engineering for Cloud-Native Toolkits

Gabor Marth, RUFUS genomic structural variation (NHGRI)

Extracting Data for Sharing on Cloud

Melanie Fried-Oken, Brain-Computer Interface software to collect & share severe speech defect data using cloud (NIDCD)

Extracting Data for Sharing on Cloud

Melanie Fried-Oken, Brain-Computer Interface software to collect & share severe speech defect data using cloud (NIDCD)

Community Outreach

Gerardo Andres Cisneros, multi-scale modeling & dynamic "graphic novel" on Twitter for LatinXChem (NIGMS)

NEW Funding Announcement in FY24!

Research Software Engineering Program (R50)

Pilot a new model to support research software engineers in biomedical and behavioral research

Software Sustainability Program (R03)

Foster software foundations to increase robustness, reproducibility, and reusability of NIH supported open software

https://dpcpsi.nih.gov/council/january-19-20-2023-agenda

Request for Information From ODSS:

Sharing NIH Supported Research Software

Purpose:

- Soliciting input on best practices for sharing research software.
- Informing NIH guidance on development, implementation, and sharing of NIH-supported research software.
- Framing guidelines for the sharing of research software.

Respond by: February 1, 2024.

Office of Data Science Strategy

Respond to the RFI here:
http://bit.ly/3M22jAt

Goal 4

Support for a Federated Biomedical Research Data Infrastructure

Challenges

- Creation of opportunities for exploration of new technologies and computing paradigms for biomedical research

Objectives to Address Challenge

1. Develop, test, validate, and implement ways to integrate NIH data and infrastructure
2. Ensure a robust and connected data resource ecosystem that includes collaborative data management platforms, curation, analysis, and sharing of data and metadata
3. Develop new capabilities for data search and discovery

STRIDES Initiative

Value to Participants

Participants in the NIH Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative benefit from:

- **Competitive** pricing & financial benefits
- **Flexible** business model
- **Expert** support from cloud providers
- **Training** expertise and scaling capacity
- **Professional** service consultations
- **Expanded** communication reach
- **Reach-through** to additional partners

Google AWS Microsoft Azure

Impact *as of November 30, 2023*

224+	PETABYTES OF DATA
549M+	COMPUTE HOURS
1,984+	RESEARCH PROGRAMS
$82M+	COST SAVINGS
5384+	PEOPLE TRAINED

Supporting Researchers to use STRIDES

Encourage and enable researchers to explore and test opportunities by incorporating cloud capabilities

- *Michael Valerius* (Brigham and Women's Hospital)- ***Atlas-D2K Exploring Cloud Optimization***
- *William Howe* (Virginia Polytechnic Institute)- **An evaluation of the costs and benefits of cloud computing for modern systems neuroscience**

Impact from 2022 (NOT-OD-23-070)

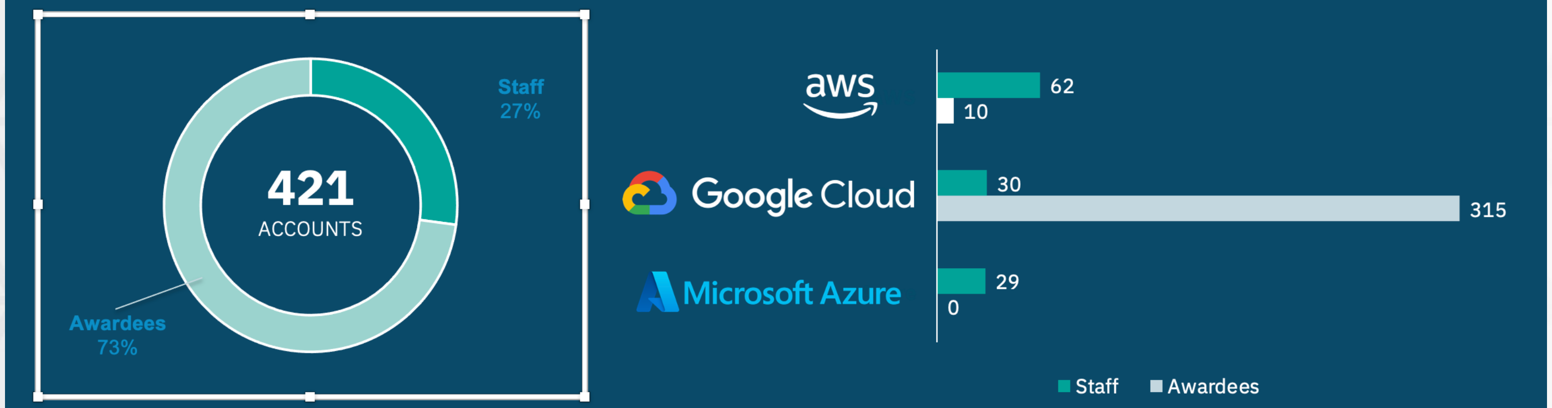
$1.8M	ODSS in funding
40%	Intramural Researchers
41%	For new Staffing costs
7	NIH ICOs

NIH Cloud Lab

Experiment in the Cloud

Through this resource, NIH-funded researchers will become more efficient and comfortable in leveraging the cloud for their research purposes.

Accounts Issued as of December 2023



cloud.nih.gov/resources/cloudlab

Use Cases

Evaluate Utility & Cost

Reduces the financial, labor, and time commitments required to evaluate the cloud's utility/cost for a project

Develop New Tools

Allows experienced teams to prototype new architectures and evaluate software and hardware combinations

Share Ideas

Enables researchers across the world to share ideas on how to conduct biomedical research in the cloud

Learn New Skills

Simplifies access to tools and cloud environments that participants can use for training purposes

Goal 5

Strengthen a Broader Community in Data Science

Challenges

- Develop and nurture data science talent from a diverse array of scientific interests and institutions

Objectives to Address Challenge

1. Increase training opportunities in data science
2. Develop and advance initiatives to expand the data science workforce
3. Broaden and champion capacity building and community engagement efforts
4. Enhancing data science collaboration within the NIH Intramural Research Program

Training, Workforce, and Community Engagement in FY23

- Institutional Training Awards:
  - 14 R25
  - 7 T32
- Individual Training Awards:
  - 9 Diversity K
  - 1 Diversity F
  - 7 Diversity Supplements
- Intramural Training:
  - GDSSP

Promote Training Pipeline

Expand Data Science Workforce

Data Science to Serve All Populations

Enhance Data Science Capacity

Provide Tools and Resources

- DATA Scholar Program:
  - 10 new scholars
  - 7 scholars completed program
  - ODSS gained FTE slot
- DataPath Program:
  - Concept approved and program launched
  - 3 new fellows
- Leadership Scholars Program

- 12 Enhancing Capacity Supplements
- 4 NARCH awards
- 1 Tribal Epidemiology Center
- 4 Cloud-Based Learning Modules supplements (IDeA)
- 1 Education Hub (NHGRI)
- 2 DS-I Africa awards
- 2 FIC LAUNCH program

- Completed Phase 1 of DataSciZone
- Assumed management of the NIH Coursera Program
- Conducted codethon in collaboration with ASBCB

The The Hawaii Advanced Training in Artificial Intelligence for Precision Nutrition Science Research (AIPrN)

- Co-funded 5-year T32 training program to develop individual-level dietary and nutritional intake by accounting for individual variability in genes, phenotype, environment, and lifestyle

- Novel concept to include the assessment of biological, clinical, social, and environmental parameters including multi-omics, genomics, proteomics, metabolomics, and sustainability

- Partnership between ODSS and NIDDK

Request for Information From ODSS:

NIH Strategic Plan for Data Science

- Read and submit your comment on the draft NIH Strategic Plan for Data Science, 2023-2028

- **The NIH seeks comments on any of the following topics**:
  - The appropriateness of the goals of the plan, the strategies and implementation tactics proposed to achieve them; including potential benefits, drawbacks or challenges
  - Opportunities for NIH to partner to achieve these goals
  - Emerging research needs and opportunities that should be added to the plan.
  - Any other topic the respondent feels is relevant for NIH to consider in developing this strategic plan.

- **Last day to submit: March 15, 2024**

https://bit.ly/3vc4MTq