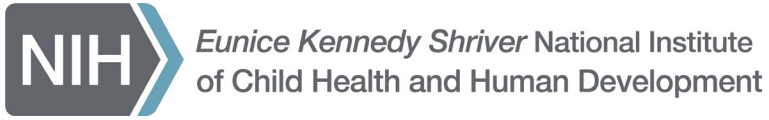
Strategies for Using Diverse Data Sources for Disaster Science

Rebecca F. Rosen, Director and Valerie Cotton, Deputy Director, NICHD Office of Data Science & Sharing

Symposium on Disaster Data Science

April 2, 2024

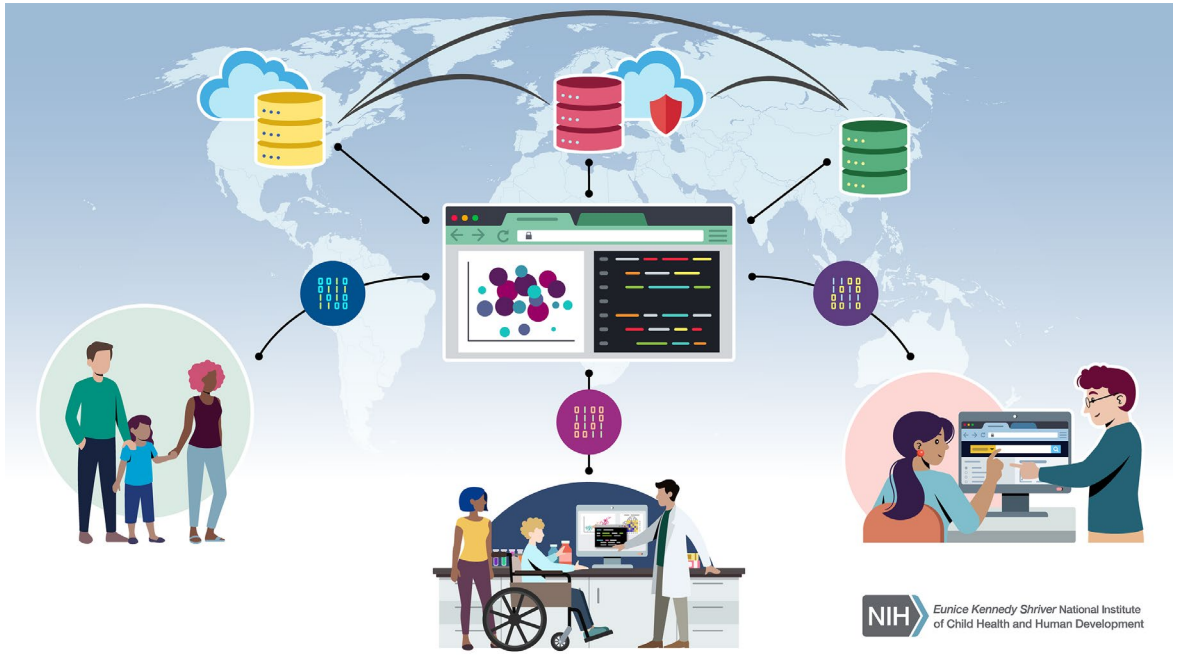


NICHD Office of Data Science and Sharing Vision and Mission

Vision: A culture of responsible and innovative use of data and biospecimens that accelerates research and improves health for NICHD populations

- *Children*
- *Pregnant and lactating people*
- *People with disabilities*

Mission: NICHD ODSS will develop a diverse, secure, and interoperable research data ecosystem and will advise on best practices for data collection, standards, management, sharing, and use across the research and funding lifecycles, in order to advance scientific discovery in support of NICHD's mission to understand human development, improve reproductive health, enhance the lives of children and adolescents, and optimize abilities for all.



NICHD Data Ecosystem

The NICHD Data Ecosystem is comprised of **people, data, processes, and technologies** to align with the NICHD Strategic Plan and support NICHD communities' data science and sharing needs

PRINCIPLES

- Leverage existing NICHD and NIH investments (e.g. datasets, data repositories, and tools)
- Follow human centered design practices, where real users inform and test all work
- Invite contributions from diverse researchers, developers, and other community members
- Utilize open metadata, data, and software standards
- Follow privacy and security by design principles

CURRENT STRATEGY

1. Assess all NICHD-relevant data repositories to inform:
    - **Sustainability:** Data repositories that do or can meet NICHD needs for long-term data sharing and data security
    - **Interoperability:** Approach to building connections between NICHD, NIH, and external data repositories and tools
2. Collect and prioritize user stories, and then implement work to address researcher, staff, participant, and community needs

Consider Multiple Layers of Interoperability

Governance

Rules and controls that define and enforce appropriate data collection, access, sharing, linking, and use.
Examples: policies, regulations, consent, data use agreements

Data

Semantics, formats, structure, and mappings.
Examples: data and metadata standards, common data elements, record linkage

System

System to system connectivity.
Examples: web services/APIs, single sign-on, cloud-based analysis platforms

Example Use Cases

Researchers gain authorized access to sensitive datasets by following data and repository-specific governance requirements

Researchers integrate disparate datasets by mapping common data elements, standards, geographic areas, or individual participant identities

Authorized researchers co-analyze large datasets by calling repository APIs from cloud analysis tools, using centrally issued identity and entitlement tokens

GOVERNANCE: Record Linkage Implementation Checklist

- Describes **governance** and technical considerations for implementing record linkage (e.g., PPRL) based on an assessment of existing record linkage implementations.
- Driven by pediatric COVID-19 user stories identified by NICHD and NIH researcher communities, given the federated nature of the NIH data ecosystem.

User Story	Current Problem
What does the user want to be able to do?	***Why can't the user do this today?***
As a researcher/clinician, I want to combine participant-level data collected from multiple studies and data repositories to merge multiple data types for each participant and avoid working with inflated sample sizes to effectively study COVID in children	We believe the same children were recruited for multiple studies with different data collection protocols and the data are shared through multiple data repositories, but we don't have a way to identify which children are the same without sharing personally identifiable information which is not allowed by our IRB

Implementation Checklist

Governance Considerations :

- ❑Determine the scope of linkage (which datasets to link)
- ❑Obtain approval to link
- ❑Identify policies relevant to specific data type(s) or participant population(s)
- ❑Establish which party should link the data
- ❑Use a variety of controls to mitigate re-identifiability risk

Technical Considerations:

- ❑Collect & standardize PII elements for high linkage quality
- ❑Select a technology that meets basic requirements
- ❑Consider PPRL Tool Sustainability for Long-term Implementations

DATA: FHIR Information Exchange Standard for Interoperability

User Story	Current Problem
What does the user want to be able to do?	***Why can't the user do this today?***
As a researcher/clinician in maternal health, I would like a tool that makes it easier to search for reproductive health studies in dbGaP and other NIH data repositories.	Open access study metadata from NIH data repositories are often not structured in or accessible via FHIR

FHIR – Fast HealthCare Interoperability Resources, an HL7 standard for healthcare data exchange. NIH encourages researchers to use FHIR to capture, integrate, and exchange clinical data for research purposes and to enhance capabilities to share research data
API – Application Programming Interface, a way for two or more computer programs to communicate with each other

Local EHR data

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM

db GaP GENOTYPES and PHENOTYPES

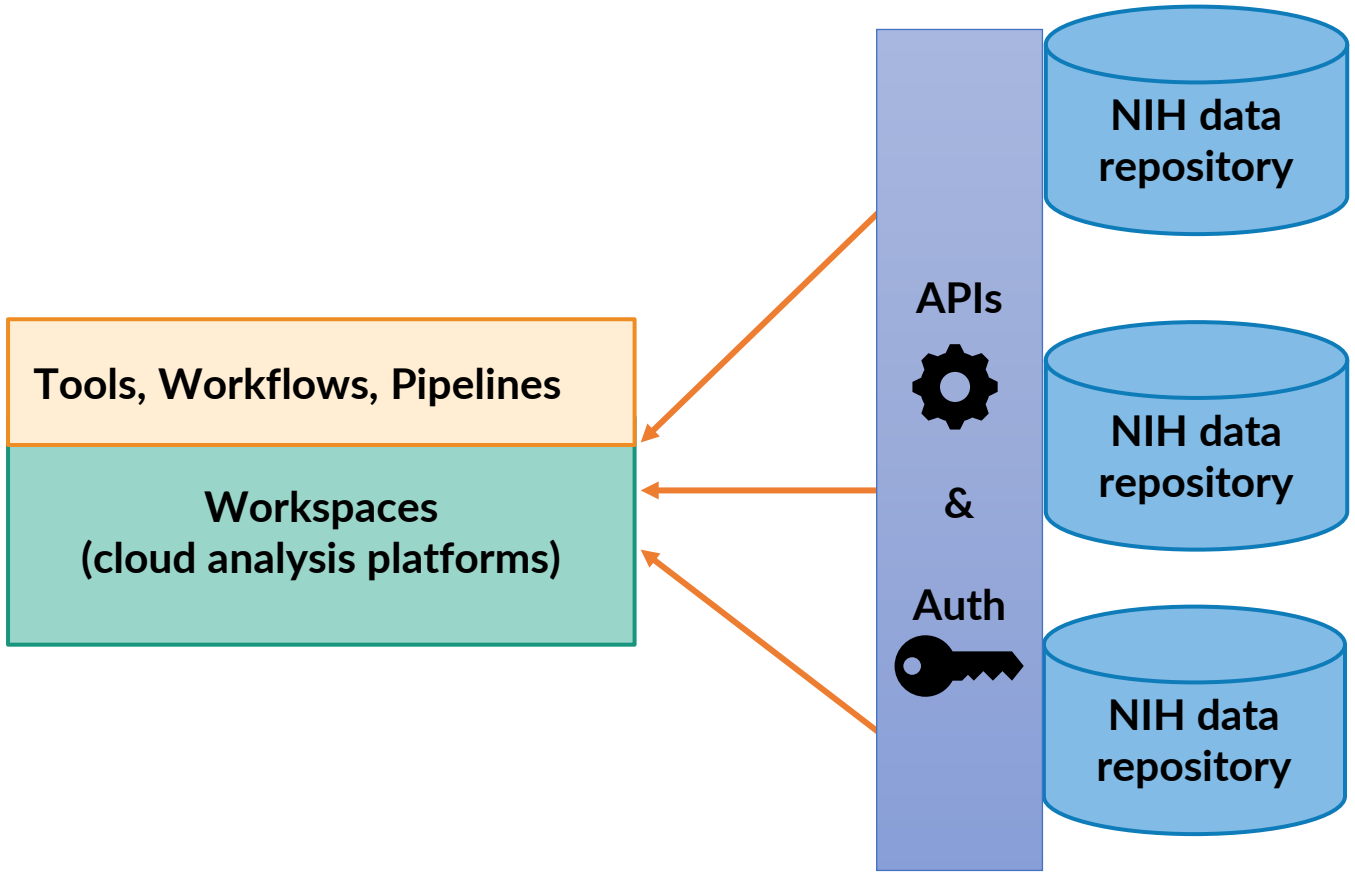
IMMPORT Shared Data
Your site for searching and downloading shared data

Coming soon:

Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD)
DASH Data and Specimen Hub

SYSTEM: Cloud Workspaces to Co-analyze Data from Diverse Sources

User Story	Current Problem
What does the user want to be able to do?	***Why can't the user do this today?***
As a perinatology researcher, I want to easily access and co-analyze NuMoM2b cohort data from DASH and BioData Catalyst	This is not possible today because data from two studies performed with the NuMoM2b cohort are stored in different NIH data repositories because of different Institute sharing expectations. DASH lacks APIs and connectivity to cloud platforms that are part of or interoperate with BioData Catalyst.



NIH Cloud Platform Interoperability (NCPI) program is actively testing and adopting APIs and other interoperability standards:
https://www.ncpi-acc.org/resources/technologies

Considerations for Today's Workshop

- Our esteemed organizers used a human centered design approach for today's workshop structure:
  1. **Planning**: Collected Use Cases from co-organizers and their communities (scientific questions, user stories) to identity user needs
  2. **White Paper**: Published foundational research to understand current state of relevant data systems and data sources
     - *Advancing Disaster Data Science: A Commissioned Paper for the Action Collaborative on Disaster Research*
  3. **Today's Workshop**: Initiating discussions amongst users, system/data source owners, and other stakeholders to identify potential solutions to address use cases
     - Are there opportunities to improve existing data sources or systems (new features, data curation)?
     - Should we build new tools to interact with these systems?
     - Is there an opportunity to educate researchers on how to make use of the existing data, systems, and tools?
     - *How can we work together on next steps?*

THANK YOU!

Apply Human Centered Design

Tracking NICHD User Stories (what users want to do and why) and how they fit with NICHD goals

- Drives our strategy for ecosystem improvements
  - Technical solutions, interoperability work, funding opportunities for new projects, responses to NIH and Federal data calls (e.g., HHS)
  - Informed repository assessment variables
- To eventually post on NICHD's public GitHub repository
  - Fosters collaborative solutions and adoption of tools/solutions that will be developed
- Serves as foundation for formal use cases (documentation for desired functionality)

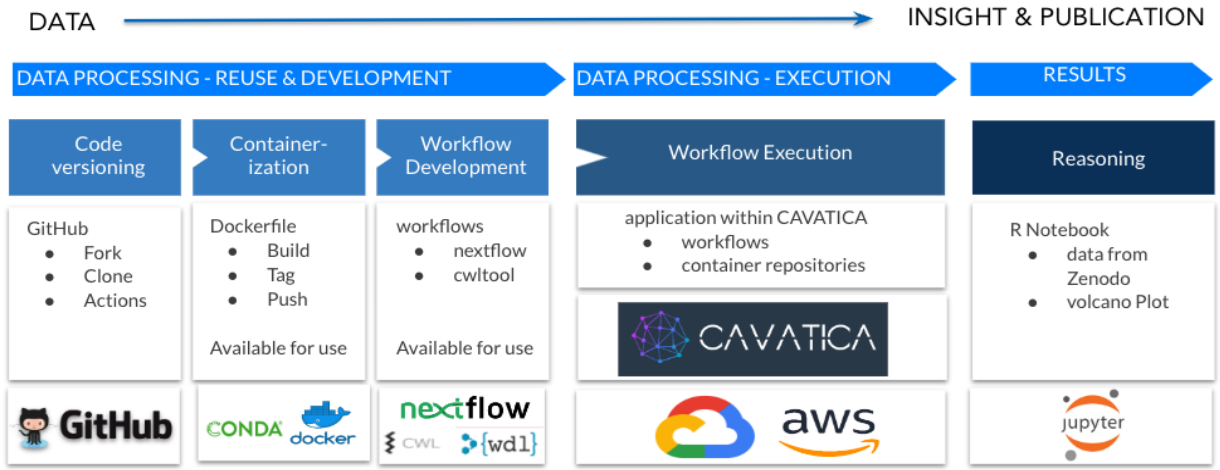
User Story	Current Problem	User
What does the user want to be able to do? *Detail a specific situation in the format "As a (user), I want to (action), so that (action)."*	**Why can't the user do this today?** *Describe in the format "Today, when I try to (do this) (what happens?)" or "This is not possible today because (why)"*	**How would you describe the type of user?**
As a program officer, I want to build a central place for my researchers to find and use global HIV health data	Some data reside in DASH, ImmPort, some in other places, most are HIV data but also some control data but there is no specific portal that makes all the relevant data findable	NICHD Staff
I am a parent that just found out my child has Down syndrome. I am wanting to know how long people with Down syndrome live, what health conditions they have, and what kind of community I am joining. I am not a reader of scientific papers and like visuals.	There are resources available but nothing in a user-friendly data dashboard	Participant Community

Encourage Use of the Ecosystem with Training

Elements of Style in Workflow Creation and Maintenance

by NICHD DATA Scholar, Anne Deslattes Mays

- Goal: Building solutions from data to insight & publication by containerizing at the process level, stitching together with workflow languages, and analyzing with JupyterLab Notebooks
- Audience: Broad (no previous command line skills) to build literacy, capacity, skill, and understanding to perform analysis on any platform.
- Public sharing – everything documented on NICHD GitHub for continuous and offline learning



NICHD GitHub https://github.com/NIH-NICHD
5 Day course – available online for self-study

One attendee said ***"I feel as if I have just had my skills modernized"***

INCLUDE Data Coordinating Center

Gabriella Miller Kids First PEDIATRIC RESEARCH PROGRAM Data Resource Center