Digitizing Data Linkage Governance Metadata Towards Streamlining Decision Making for Patient-Centered Outcomes Research Data Linkages

NICHD Office of Data Science & Sharing

April 2024

NIH *Eunice Kennedy Shriver* National Institute of Child Health and Human Development

Speakers:

Rebecca Rosen
Director

Valerie Cotton
Deputy Director

Elizabeth Clerkin
Data Science and Policy Specialist

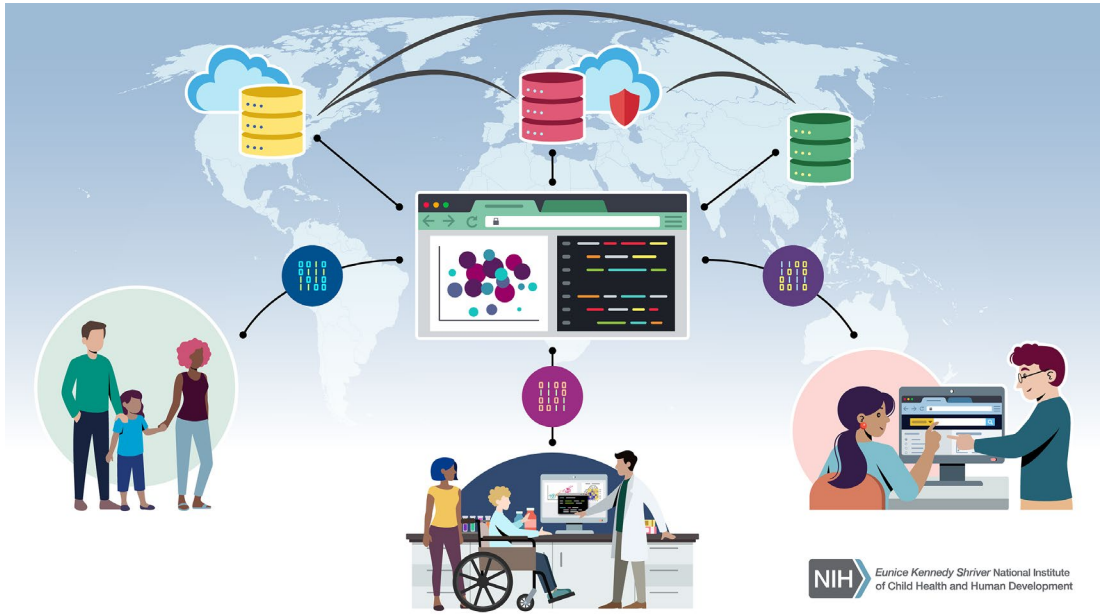
NICHD Office of Data Science & Sharing

Develop a diverse, secure, and interoperable research data ecosystem 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*

Work funded by:
- HHS OS-PCORTF
- NIH Office of Data Science Strategy

With support from Essex Management, Booz Allen Hamilton, and The MITRE Corporation

Contributions from many NICHD, NIH, and other federal and external collaborators!



Today’s Presentation

- Motivation and Approach
- Project Outcomes
- Discussion

Predecessor Project: Privacy Preserving Record Linkage (PPRL) for Pediatric COVID-19 Studies

- The 2022 report describes **governance** and **technical** approaches for PPRL based on an assessment of existing record linkage implementations.
- The assessment was designed to address pediatric COVID-19 use cases identified by NICHD and NIH researcher communities, given the federated nature of the NIH data ecosystem.
- The report proposes a set of governance and technical considerations that could inform the design of any PPRL implementation in a federated data ecosystem.

Human Centered Design Approach

Human Centered Design: Deeply understand users' needs and experiences to create effective solutions – *all work is driven by **use cases** or **user stories***

User Story	Current Problem	User Goal
What does the user want to be able to do?	***Why can't the user do this today?***	***What is the user's ultimate goal?***
As a researcher/clinician, I want to combine participant-level data collected from multiple studies 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, but we don't have a way to identify which children are the same without sharing personally identifiable information (PII)	My goal is to link data for the same children across multiple studies and data repositories without sharing PII

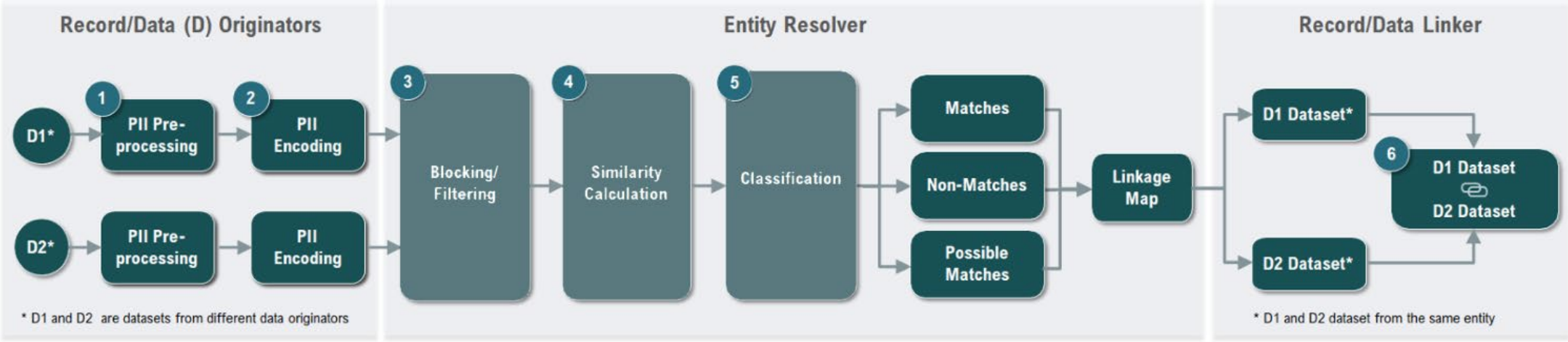
Can PPRL solve this?

*Work driven by needs that emerged for pediatric COVID-19 research, **but the solutions apply broadly***

PPRL Solutions

PPRL: An approach that uses secure software to enable users to link data from multiple sources to the same individual without revealing personally identifiable information (PII).

- PII is entered into the software to create cryptographically encoded (hashed) codes or "tokens“ but the PII does not leave the data originator.
- The same software must be used to enable linkage across datasets
- The outcome is a linkage map matching cross-study participant IDs (or GUIDs)



Assessing Data Governance

Data governance: The collective set of rules and controls that define and enforce appropriate collection, linking, sharing, access, and use of data.

We examined the **people, policies, processes, and controls** across the data lifecycle for 13 existing record linkage implementations, by asking questions like:

- *Which party **matches the tokens** (or other information for non-PPRL, entity resolution) and which party **links the data** (merging, de-duplicating)?*
- *What is the scope of linkage? Is it specific to one study, or does it encompass a broader database to support many studies?*
- *How is access to linked data approved?*
- *What risk mitigation procedures (controls) are in place?*

We then aligned these practices with the **needs of the pediatric COVID use cases** to develop considerations

Linkage examples using PPRL

- NIH BRICS & NIMH Data Archive
- NCATS N3C (3 classes)
- PEDSnet
- CDC CODI

Linkage examples using non-PPRL

- NIH dbGaP
- NIH *All of Us* (before CLAD)
- Georgetown FSRDC (US Census Bureau)
- CDC NCHS + NDI Mortality Data

NICHD Record Linkage Implementation Checklist

Led to the creation of a checklist for guiding decisions that must be made prior to designing and implementing a strategy for linking data from multiple sources and sharing and using the linked datasets for research

Implementation Checklist

Governance Considerations :

- ❑Determine the scope of linkage (which datasets to link)
- ❑Obtain approval to link
- ❑Identify policies and rules that apply to each dataset, 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

NICHD Record Linkage Implementation Checklist

Use case: For participants with Down syndrome who are represented in multiple studies across NIH repositories, how can I merge multiple data types?

INCLUDE Data Coordinating Center

Implementation Checklist

Governance Considerations :

- ❑Determine the scope of linkage (which datasets to link) → Research data shared through multiple NIH data repositories (INCLUDE Data Hub, NIMH Data Archive)
- ❑Obtain approval to link → Data contributors determine if linkage is appropriate (obtain additional approvals, as needed), NIH program approval
- ❑Identify policies and rules that apply to each dataset, specific data type(s), or participant population(s) → NIH Genomic Data Sharing Policy, consent-based data use limitations, other policies
- ❑Establish which party should link the data → Approved researchers will link/merge datasets
- ❑Use a variety of controls to mitigate re-identifiability risk → Share linkages after data access committee approval, sign Data Use Agreement (prohibits re-identification), standard de-identification of all datasets

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

NDA (The **N**ational Institutes of Mental Health (NIMH) **D**ata **A**rchive)

GUID (**G**lobal **U**nique **Id**entifier) **Tool**

NICHD Record Linkage Implementation Checklist

Prior to designing and implementing a record linkage strategy, funders, researchers, data repositories, and other stakeholders should collaborate to make a series of governance and technical decisions.

Implementation Checklist

Governance Considerations :

- ❑ Determine the scope of linkage (which datasets to link)
- ❑ Obtain approval to link
- ❑ Identify policies and rules that apply to each dataset, 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

Motivation for current project:

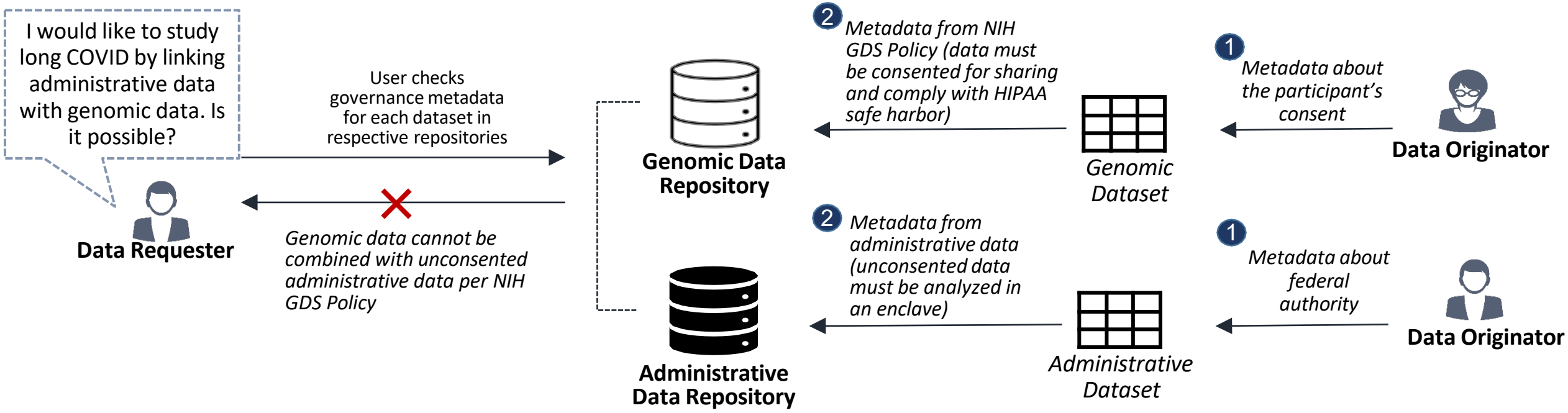
It is important to understand how a linked dataset **inherits rules associated with each original dataset contributed to the linkage**

Rules can come from laws, policies, consent forms, agreements etc....

How do we capture, structure, and exchange dataset level rules (in a standard manner) to inform record linkage decisions?

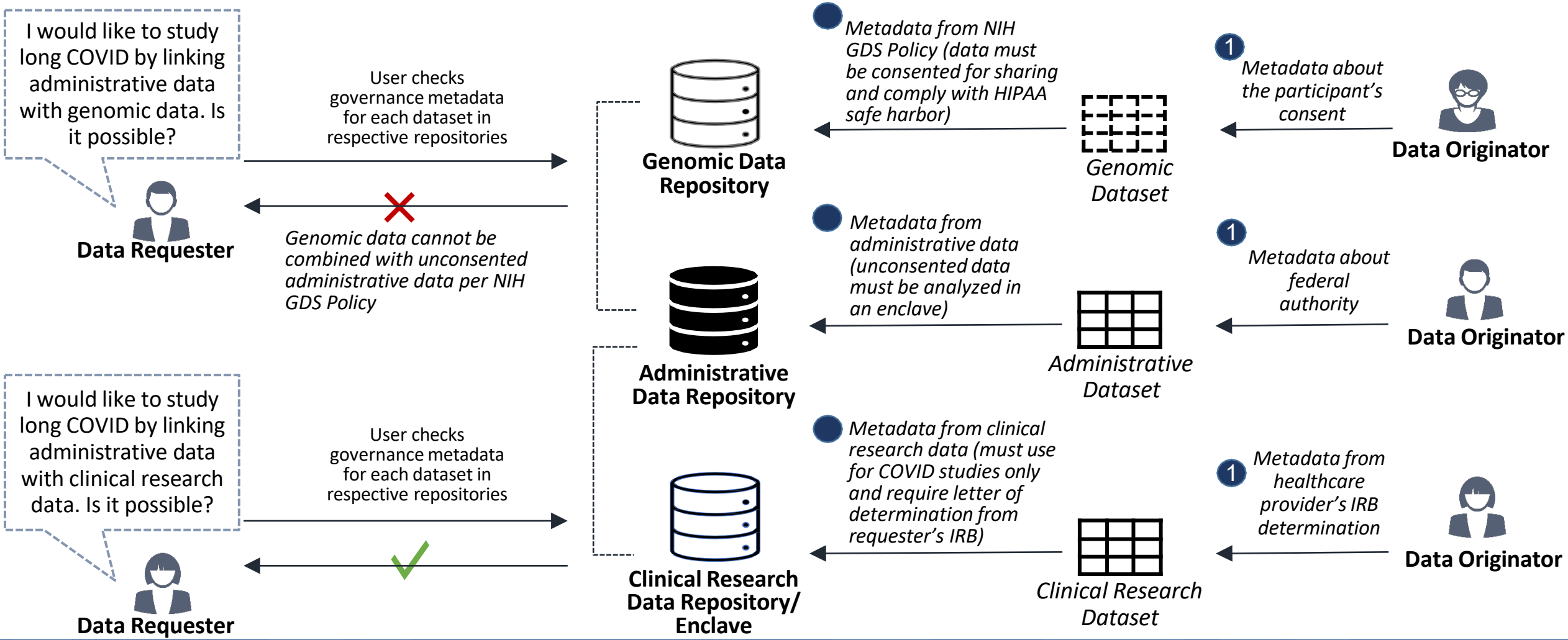
Project Motivation

Example data linkage decision making based on cumulative governance information



Project Motivation

Standardizing the cumulative governance information for datasets and combinations of datasets as structured metadata would streamline the decision-making process for data linkage and use



User Stories for Current Project

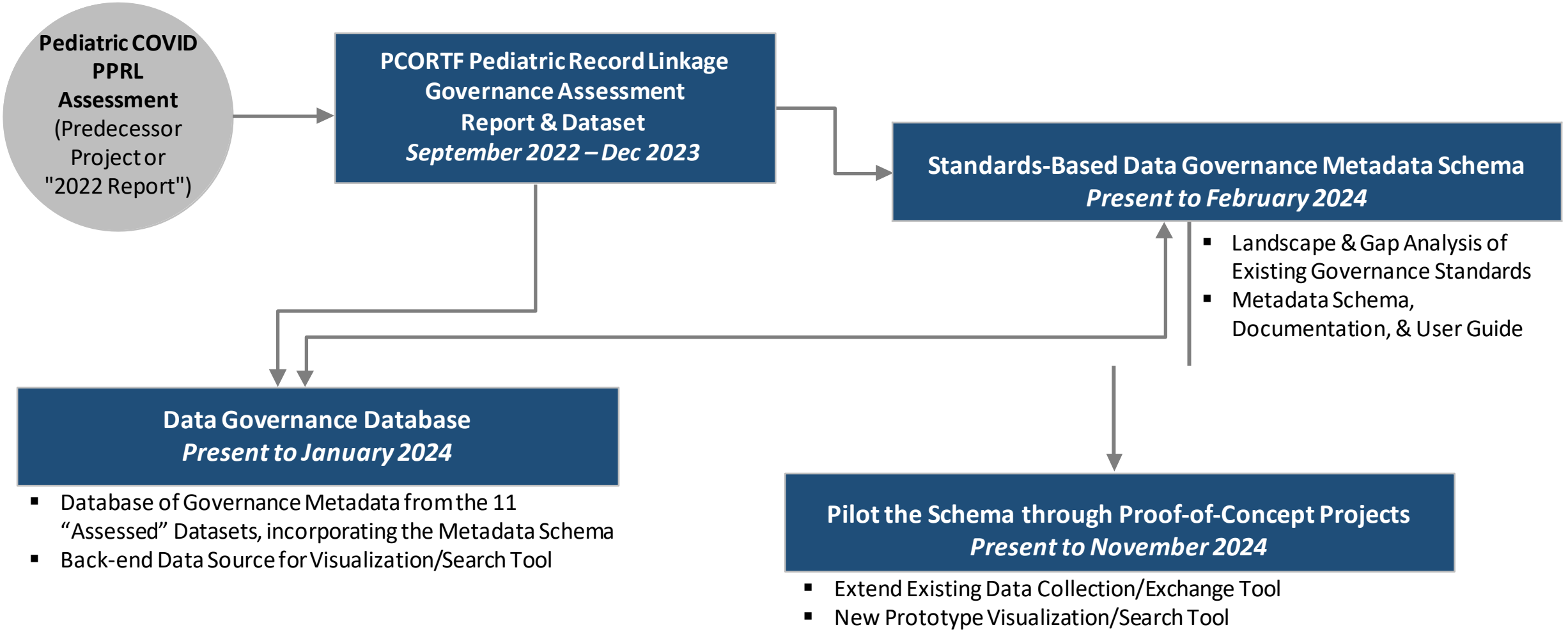
User Story	Current Problem	User Goal
What does the user want to be able to do?	***Why can't the user do this today?***	***What is the user's ultimate goal?***
As a researcher, I want to **access, link, share, and use data from multiple research and administrative data sources** (such as CDC NHANES, SAMHSA NSDUH, NIH/NIDA MTF, and ACF AFCARS) so I can study the effects of COVID-19 pandemic on mental health of children and determine whether related outcomes are more severe for children in foster care.	**Each dataset is subject to different rules often stored as unstructured narrative text** within policy documents, data use agreements, consent forms, laws, and other sources of governance information. **It's difficult to extract this information and understand how these rules intersect.**	My goal is to understand **whether certain datasets can be linked, and if so, what rules and controls apply to the resulting linked dataset** so I can appropriately share and use the linked data to study pediatric COVID.
As a researcher, I want to access, link, share, and use data from multiple research and administrative data sources (such as NIH/NCI NCCR, CMS T-MSIS, and ACF AFCARS) so I can study the impact of COVID-19 infection of COVID-19 infection on pediatric cancer survivors and the impact of COVID-19 infection on future pediatric cancer outcomes.		
As a researcher, I want to access, link, share, and use data from multiple research and administrative data sources (such as NIH/NCATS N3C, PCORnet, NIH RADx, and EPA Air Quality) so I can study whether SARS-CoV-2 vaccination results in reduced asthma-related school absences at 3/6/12+ months post-vaccination.		

Current Project: Streamlining Governance of Pediatric COVID-19 Research Data Linkages

- Purpose: Develop and test a generalizable governance metadata schema that
  1) will streamline the decisions of data stewards, researchers, and other stakeholders on whether two or more datasets can be linked and how linked data can be used, and
  2) can be incorporated into future PCOR data collection workflows.

- Objectives:
  - Develop a user-friendly and machine-readable standards-based metadata schema that describes data governance requirements for each dataset
  - Develop the metadata schema based on dataset governance requirements extrapolated from patient-centered outcomes research use cases that involve linking and using high priority NIH and HHS clinical and administrative pediatric datasets
  - Pilot test the metadata schema in data collection and visualization tools, engaging researchers, software developers, and policy experts

High-Level Project Structure and Interrelationships



Datasets for Pediatric COVID-19 Use Cases

Use Case	Data Source	Agency	Data Types
Use Case 1	National Health and Nutrition Examination Survey (NHANES)	CDC	Survey (dietary, health, etc.), Clinical, Demographics, Laboratory
	National Survey on Drug Use and Health (NSDUH)	SAMHSA	Survey (substance use/treatment), Demographics, Mental health
	Monitoring the Future (MTF)	NIH/NIDA	Survey (drug use, self-esteem, violence, crime, etc.), Demographics
	Adoption and Foster Care Analysis and Reporting System (AFCARS)	ACF	Demographics, Adoption/placement data, Outcomes data
Use Case 2	National Childhood Cancer Registry (NCCR)	NIH/NCI	Clinical, Demographics, Incidence/Survival Data
	Transformed Medicaid Statistical Information System (T-MSIS)	CMS	Clinical (Diagnoses and procedures), Patient demographics, Charges, Enrollments and Service Use
	COVID Data Tracker	CDC	Clinical (vaccine effectiveness, ER visits, etc.), Claims, Demographics
Use Case 3	National COVID Cohort Collaborative (N3C)	NIH/NCATS	EHR, Clinical (procedures, medications, etc.), Demographics
	PCORnet	PCORI	EHR, Clinical (procedures, medications, etc.), Demographics
	COVID Rapid Acceleration of Diagnostics (RADx)	NIH	Clinical, behavioral, survey/interview, diagnostic/test, biological sample, sequencing and imaging data
	Air Quality Data Collected at Outdoor Monitors Across the US (AIR Data)	EPA	Pollutant Data, Remote Sensing Data, Community Multi-scale Air Quality, Satellite Data, Biomass Burning Data

Example Linkage Determination

Use Case 1: Effects of COVID-19 pandemic on mental health of children. Are related outcomes more severe for children in foster care?

INDIVIDUAL DATASET GOVERNANCE LIMITATIONS
(Select Governance Examples)

NHANES Mental Health - Depression Screener – Youth

1. Can only be used for statistical purposes/research
2. Access requires Data Use Agreement, NCHS Review Committee approval, and data requesting institution's IRB determination
3. Must be accessed on-site at the NCHS Research Data Center (RDC) enclave (download not permitted)

Monitoring the Future (MTF)

1. Does not have explicit authorization for linking
2. Access requires Data Use Agreement, NAHDAP staff approval, and data requesting institution's IRB determination
3. Must be accessed via the ICPSR virtual enclave (via a virtual isolated machine; download not permitted)

NHANES and MTF datasets can be linked provided the data user:

A. Addresses the following <u>gap</u>:

1. Obtains linkage authorization for MTF dataset

B. Complies with the following <u>implementation controls</u>:

2. Signs Data Use Agreement, obtains approval from NCHS Review Committee and NAHDAP staff, and provides letter of determination from their institutional IRB

3. Accesses data in the required location (requires resolution of conflicting controls, e.g., by obtaining approval to export MTF data into the NCHS RDC)

C. Complies with the following <u>data use limitation</u>:

1. Uses the linked dataset only for statistical purposes/research

LINKED
NHANES Mental Health - Depression Screener – Youth
+ Monitoring the Future (MTF) Dataset

1. Can only be used for statistical purposes/research
2. Access requires data use agreements, approvals from NCHS Review Committee and NAHDAP staff, and a letter of determination from their institution's IRB
3. Must be accessed in the required location (requires resolution of conflicting controls, e.g., by obtaining approval to export MTF data into the NCHS RDC)

LINKED DATASET GOVERNANCE

Record Linkage Governance Assessment Report: S*elect Findings*

- Dataset documentation often does not explicitly **authorize linkage or specify scope of linkage**
  - Approval can be granted on a case-by case basis, often by a committee
- Linked dataset governance converges on the **most constraining requirements** of the contributing datasets
- Apparent **conflicts in governance** between contributing datasets introduce complexity in defining appropriate use of linked datasets - **stakeholders need to make decisions**
  - Common conflicts include where the data can be shared/accessed (i.e., which enclave)
- Linkage implementations must consider **how the linked dataset is de-identified**; issues include:
  - Risks added through increased richness of data about a given individual
  - Conflicts between HIPAA Safe Harbor requirements and datasets with dates & locations
  - Location information used for geo-linking data (e.g., environmental data); even if dates & locations are subsequently removed, *can location be deduced?*

Record Linkage Governance Assessment Report: *Considerations*

To facilitate dataset linkage, a data governance metadata schema should:

- Publicly share the data governance information in a **predictable and easy-to-find location**
- Include the following:
  - Explicit statements about **whether linkage is permissible**
    - And if so, guidance for what **types of linkages** are allowed or prohibited and the **rules and controls** the linked data would inherit
  - The **provenance** of data governance
  - The **roles and responsibilities** of multiple stakeholders
  - Any **decisions made** for previous linkage
- Describe data governance in a standard way to enable **human interpretation and machine-readability**

Governance Metadata Schema

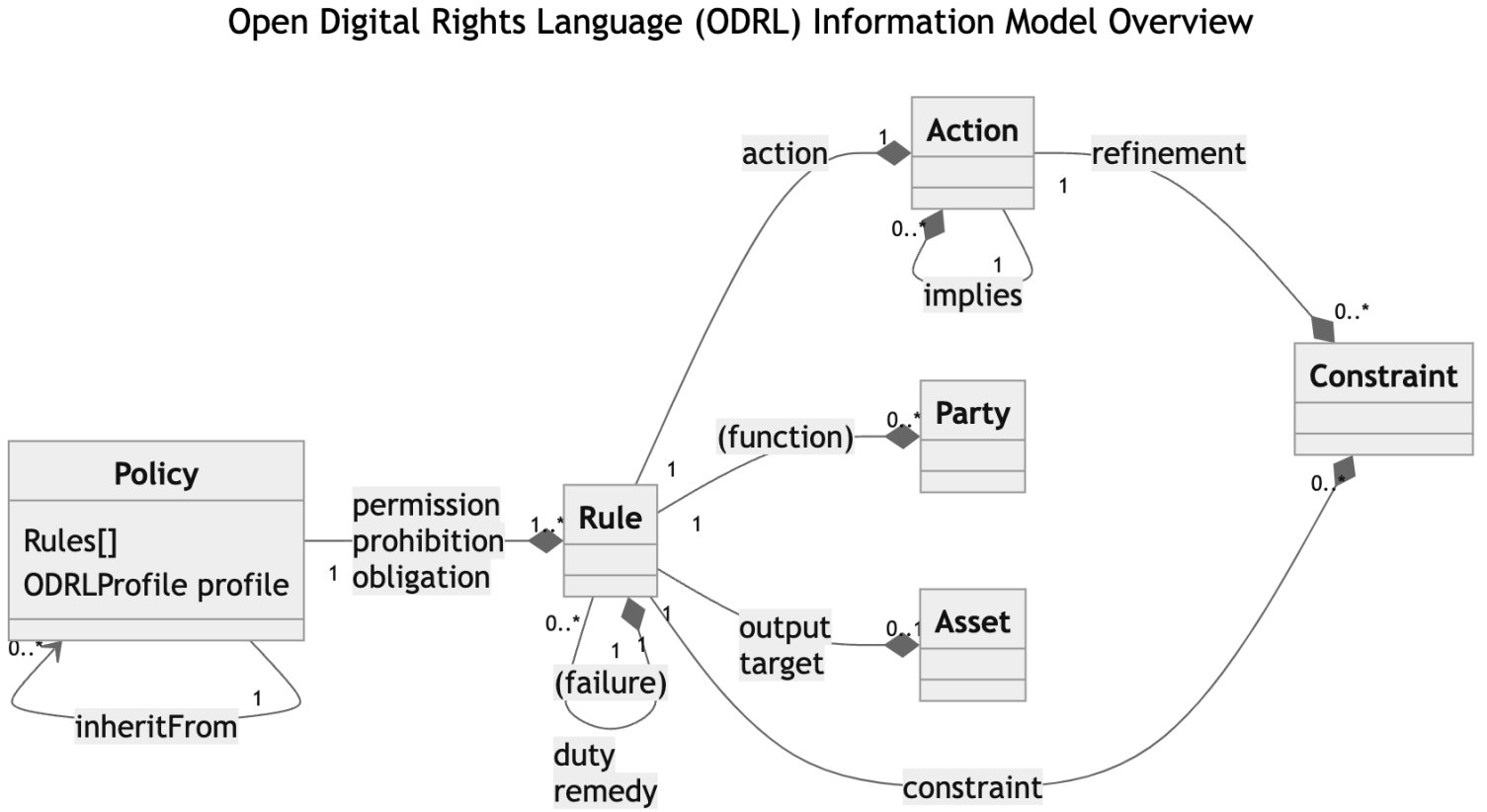
Asset: Dataset for which governance information is being provided

Policy: A non-empty group of rules of subclass set, offer, or agreement

Rule: An abstract concept that represents permissions, prohibitions, and duties

Action: An operation on an Asset

Constraint: A logical expression that expresses the conditions applicable to a Rule

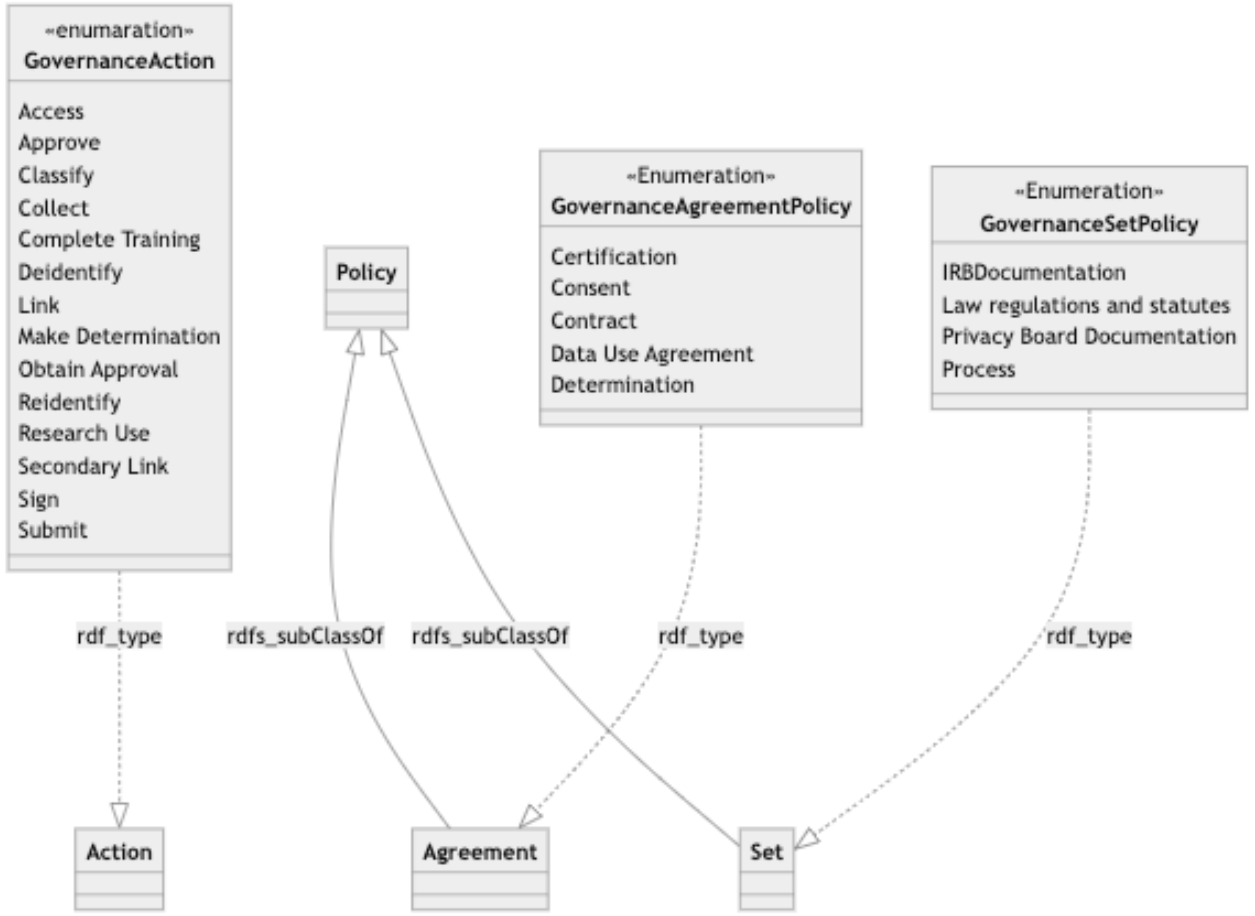


Data Governance ODRL Profile

The Data Governance ODRL Profile defines th additions to the ODRL ontology to represent data governance information.

The data governance profile includes >50 additional terms and annotations required to accurately represent governance metadata. Terms were added to represent policy types, governance actions, and constraints.

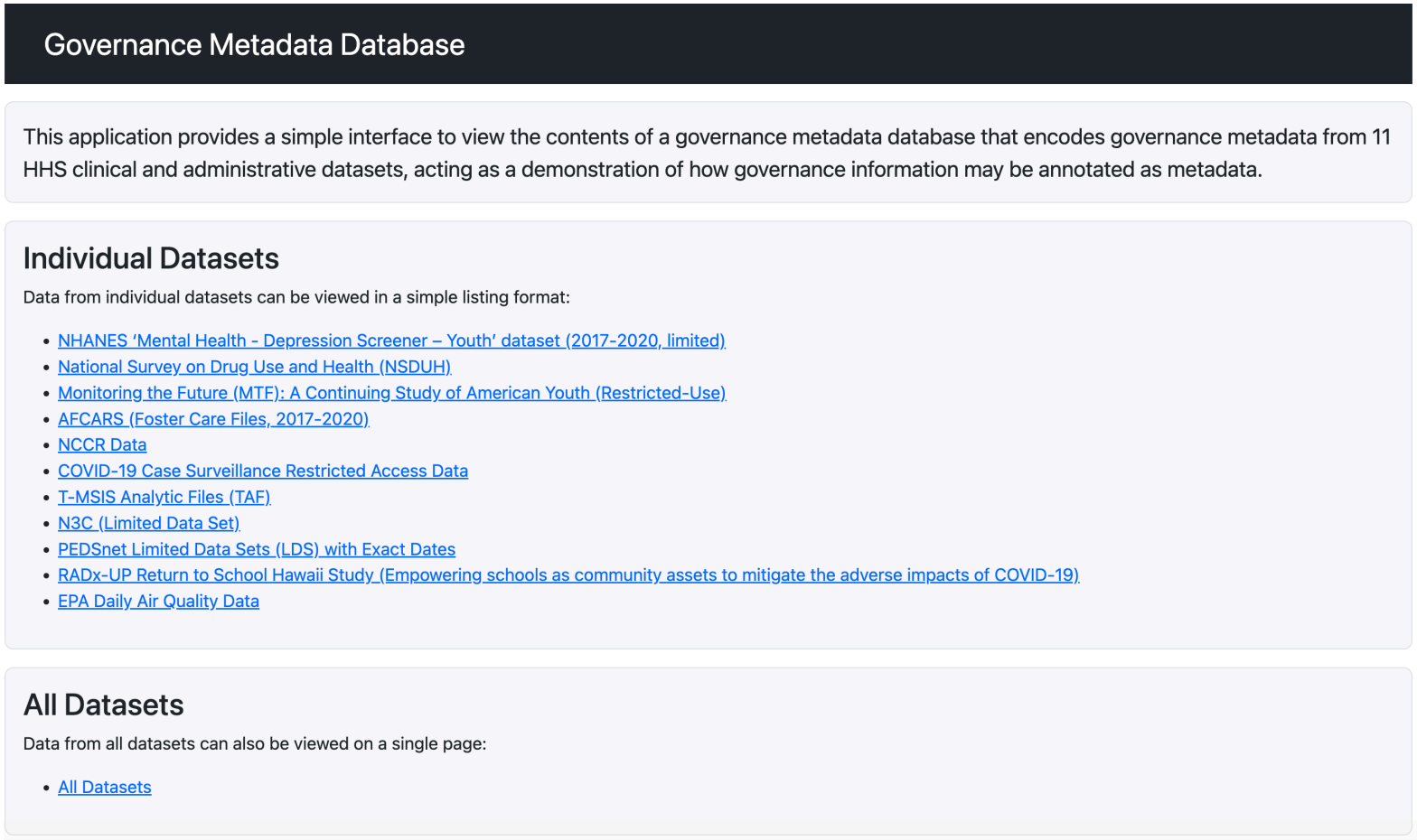
Additional terms were mapped to existing standards (Data Privacy Vocabulary and HL7) when possible.



Testing the Governance Metadata Schema
Database Development

We applied the schema design and vocabulary to a relational database

The database is the first implementation of the schema – testing how well the schema can be used to annotate governance information from 11 HHS (and EPA) datasets



Example from Data Governance Database

Consent: MTF Consent

Rule:

Type: Permission
Action: collect
Assigner: Guardian
Assignee: PrincipalInvestigator
Commentary:
Lifecycle: Data Collection
Language: Consent forms are sent to the parents of the targeted respondents 3 weeks prior to the targeted survey dates; this...
Interpretation: Consent from parents authorizes data collection
Source: NAHDAP Meeting
Source: https://monitoringthefuture.org/wp-content/uploads/2022/12/mtf2022.pdf (Accessed: 2023-04-18)

Rule:

Type: Permission
Action: researchUse
Assigner: Guardian
Assignee: PrincipalInvestigator
Commentary:
Lifecycle: Data Collection
Language: The consent specifies that the data can be used for broad research.
Interpretation: Consent specifies that the data can be used for broad research
Source: U-Mich Legal Meeting

```
---
policy:
- type: Consent
  title: MTF Consent
  uid: MTFConsent
  profile: https://www.nichd.nih.gov/data_governance_odrl
  target: MonitoringTheFuture
  permission:
  - action: collect
    assigner:
    - Guardian
    assignee:
    - PrincipalInvestigator
  - action: researchUse
    assigner:
    - Guardian
    assignee:
    - PrincipalInvestigator
```


Testing the Governance Metadata Schema
Prototype Tools - Coming Soon!

Data Collection Tool – In progress
Facilitate collection of data governance information

- What are the questions to solicit governance metadata?
- Can researchers answer questions about governance metadata?
- Do the responses generate metadata that fits within the schema?
- Do the value sets specified in the schema support governance metadata collection goals?

Search and Visualization Tool – Up next
Facilitate viewing data governance information to streamline researcher decision-making

- Can governance metadata be transformed into actionable governance information in a user-friendly search/visualization tool?
  - What governance information informs a researcher's determination about whether linkage is feasible?
  - How are the rules that flow from data sources into a linked dataset best presented?

Example from Governance Data Collection Tool

Enables collection of data governance information for a dataset, using NLM's LHC FHIR Tools

Section 5: Consent	
Were participants consented for the collection of this dataset?	Yes
Enter a link to the consent form if available	N/A
Will minor participants be reconsented in the future?	Yes
Permissions	
Does the consent permit dataset linkage?	It doesn't say
Does the consent permit dataset sharing?	Yes, with conditions
Select the conditions that the consent applies to dataset sharing	× The dataset may only be shared as a de-identified dataset × The dataset may only be shared if approved by a review body (if so describe) Select one or more or type a value
Enter the name of the review body needed for approval for sharing	Organization X Review Body
Does the consent permit dataset access?	Yes, with conditions
Select the conditions that the consent applies to dataset access	× The dataset may only be accessed in a data enclave Select one or more or type a value
Does the consent permit dataset [secondary] use?	Yes, with conditions
Select the conditions that the consent applies to dataset use	× This dataset may only be used for the approved purpose × This dataset may only be used for research on a specific topic Select one or more or type a value
Prohibitions	
Select or enter the prohibitions	× Individuals in the dataset may not be reidentified × The dataset may not be used for commercial purposes Select one or more or type a value

Discussion

While the record linkage checklist and governance metadata schema are ready for adoption, they are expected to evolve through real-world implementation within the OS-PCORTF Data Infrastructure.

Open Access to Project Resources (coming soon):

- Public GitHub repository with schema, schema user guide, and prototype tools
- Checklist and reports posted on NICHD and OS-PCORTF websites

Collaborations with HHS colleagues:

- As a PCORTF project proposal lead, I want to link a longitudinal cohort study dataset with state health department data and make the linked dataset available for research use as part of the OS-PCORTF data infrastructure
- As a PCORTF project lead, I want to create a governance playbook so that new organizations can be integrated into my secure data ecosystem, and their data linked in a streamlined fashion

We welcome additional user stories from the audience today!

THANK YOU!