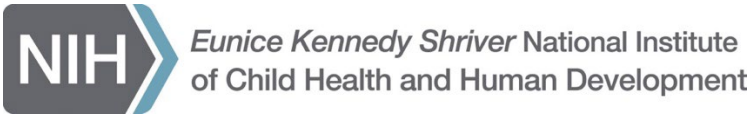
Implementation Report: Development of a Governance Metadata Visualization Prototype

Services in Support of Standardizing Governance Metadata for Pediatric COVID-19 Data Linkage

Prepared for:



Eunice Kennedy Shriver National Institute of
Child Health and Human Development (NICHD)
Office of Data Science and Sharing (ODSS)
31 Center Drive, Bldg. 31, Rm. 2A03, Bethesda, MD, 20892

Prepared by:

CMS Alliance to Modernize Healthcare (The Health FFRDC)

A Federally Funded Research and Development Center

November 4, 2024

Department of Health and Human Services (HHS), National Institutes of Health (NIH), *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD)

NICHD was founded in 1962 to investigate human development throughout the entire life process, with a focus on understanding disabilities and important events that occur during pregnancy. Since then, research conducted and funded by NICHD has helped save lives, improve well-being, and reduce societal costs associated with illness and disability. NICHD's mission is to lead research and training to understand human development, improve reproductive health, enhance the lives of children and adolescents, and optimize abilities for all.

NICHD Office of Data Science and Sharing (ODSS)

NICHD ODSS was established in 2021 to lead and coordinate NICHD's activities within data science, bioinformatics, data sharing policy and compliance, and emerging technologies. ODSS's vision is to enable a culture of responsible and innovative use of data and biospecimens that accelerates research and improves health for NICHD populations. The office's mission is to:

- Develop a diverse, secure, and interoperable research data ecosystem
- Advise on best practices for data collection, standards, management, sharing, and use across the research and funding lifecycles
- 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

ODSS is a trusted informational resource for NICHD staff and researchers on all NIH data and specimen sharing policies. ODSS serves as NICHD's primary liaison with the NIH Office of the Director's Office of Data Science and Strategy, to ensure engagement in large NIH data-science and emerging technology programs and ensure alignment with NIH, HHS, and federal programs and policies.

For additional information about this subject, you can visit the NICHD ODSS home page at https://www.nichd.nih.gov/about/org/od/odss or contact the NICHD Project Officers at:

NIH NICHD Office of Data Science and Sharing, 31 Center Drive, Bldg. 31, Rm. 2A03, Bethesda, MD, 20892
Rebecca Rosen, PhD, Director rebecca.rosen@nih.gov

Citation

CMS Alliance to Modernize Healthcare (The Health FFRDC). Implementation Report: Development of a Governance Metadata Visualization Prototype. Prepared under Contract No. 75N94023F00171. November 2024.

Authors

Susan C. Hull, MSN, RN, NI-BC, NEA-BC, FAMIA
Emily Kraus, PhD, MPH
Peter Krautscheid
Sean Mikles, PhD
The MITRE Corporation, McLean, VA

Rebecca Rosen, PhD, corresponding author
Valerie Cotton, BSc
Elizabeth Clerkin, PhD
U.S. Department of Health and Human Services, National Institutes of Health, *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), Office of Data Science and Sharing (ODSS)

Acknowledgments

This report represents a team effort, in which many individuals made contributions, particularly the leadership team of NICHD ODSS and community experts in the form of a Technical Experts Panel, to whom we extend our sincere appreciation.

This report was prepared by The MITRE Corporation under Contract No. 75N94023F00171 from the Office of Data Science and Sharing (ODSS), *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), National Institutes of Health. The authors are solely responsible for the document's contents, findings, and conclusions, which do not necessarily represent the views of NICHD ODSS. Readers should not interpret any statement in this product as an official position of NIH, NICHD, or HHS.

Notice

This technical data report was produced for the U.S. Government under Contract Number 75FCMC18D0047/75FCMC23D0004, and is subject to Federal Acquisition Regulation Clause 52.227-17, Rights in Data-General. No other use other than that granted to the U.S. Government, or to those acting on behalf of the U.S. Government under that Clause is authorized without the express written permission of The MITRE Corporation.

For further information, please contact The MITRE Corporation, Contracts Management Office, 7515 Colshire Drive, McLean, VA 22102-7539, (703) 983-6000.

Executive Summary

Linking individual-level data across biomedical datasets and the U.S. Department of Health and Human Services (HHS) administrative and survey datasets provides opportunities to maximize the value of existing data by enabling researchers to deduplicate participants across datasets, introduce new variables in analyses, reduce costly redundancies in data generation, perform longitudinal analysis, and ask new scientific questions of the enriched dataset. However, linking datasets effectively while ensuring adherence to each dataset's governance is extremely challenging given the complexities of governance information for which no standards exist. To progress the field, governance metadata must become easier to collect, exchange, and visualize to inform decisions by researchers, repositories, funders, policy/legal experts and other community members involved in linking data for research.

Recognizing this, the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) Office of Data Science and Sharing (ODSS) partnered with the Health Federally Funded Research and Development Center (Health FFRDC), operated by MITRE, to develop a first-of-its-kind metadata schema[a] for data governance[b] information relevant to linking individual-level participant data and then sharing and using linked datasets for research. The data governance metadata schema implements and extends the Open Digital Rights Language information model and enables the development of tools to collect, standardize, exchange, and visualize information about data governance, including rules from consent, policies, laws, and other sources. With funding from the HHS Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) and the NIH Office of Data Science Strategy, the NICHD ODSS and the Health FFRDC team developed the data governance metadata schema and built a governance information visualization prototype that tests the schema's capacity to accommodate real-world governance information about datasets.

This effort aligns with NICHD ODSS's larger goal of developing a governance and technology strategy for implementing individual-level record linkage for patient-centered outcomes research with NICHD populations (children, pregnant and lactating women, and people with disabilities), initially driven by pediatric COVID-19 research use cases. The data governance metadata schema and visualization prototype will also contribute to NIH-wide strategic goals for data science. The overall goal of this effort is to provide high-quality information that can be used to determine whether certain datasets can be linked, and if they can be, what rules and controls apply to the linked dataset.

This prototype project has two aims. The first aim is to test the data governance metadata schema by examining how the schema structure and design can support decisions about a proposed linkage implementation. The second is to demonstrate how governance information collected from real-world datasets can be visualized to inform how datasets of interest can be linked and then shared and used by

[a] A metadata schema is a structured set of metadata elements and attributes, together with their associated semantics, that are designed to support a specific set of user tasks and types of resources in a particular domain. A metadata schema formally defines the structure of a database at the conceptual, logical, and physical levels.

[b] Governance or data governance comprises the collective set of rules and controls that define and enforce how data are handled across the data lifecycle including: appropriate data collection, sharing, linking, access, and use. Data governance addresses privacy protections, ethics, compliance, risk management, and other requirements and derives from a variety of sources such as participant consent, IRB determinations, laws, agreements, and policy documents.

focusing on five questions about how researchers, as proxies for a broader user group including repositories, funders, and policy/legal experts, understand and assess governance information:

- Can governance metadata be transformed into human readable governance information?
- Can governance information (from metadata) be consumed and interpreted accurately by a researcher?
- Can a researcher make an assessment about the feasibility of dataset linkage for research?
- Does a researcher understand the conditions that must be met to link datasets for research?
- Does a researcher understand what actions are required to link datasets for research?

Development work reflected an agile design process, progressing in stages from examining the source governance information from 11 HHS and other federally funded datasets, building a back-end governance metadata relational database populated by real-world governance information, and then developing a visualization tool to render governance information for linkage implementation decision making. The project team collaborated with researchers as co-designers, conducted usability testing, and engaged with community experts in the form of a Technical Experts Panel who provided key guidance and feedback on this work. The results are a Governance Metadata Visualization prototype based on the metadata schema and open-source documentation to support others to develop further. The prototype includes six pages that support users to make an assessment about the feasibility of a linkage implementation based on governance metadata from the 11 federally funded datasets. All who interacted with the prototype were supportive about its value for visualization of structured governance metadata and its potential to advance linkage implementations for research.

The development and testing of the Governance Metadata Visualization prototype also highlighted ways that future iterations of the data governance metadata schema and visualization tools could be improved.

Future work to develop production-ready governance metadata visualization tools will require collaboration among data providers, repositories, funders, and policy/legal experts, to bring multiple perspectives about decision making for linkage implementations.

If widely adopted, this work would contribute to streamlining appropriate access to sensitive data for patient-centered outcomes research and promoting trust and appropriate oversight in linking individual-level participant data when collected and combined from different resources. A refined governance metadata schema and governance information visualization tools could be leveraged throughout the HHS and NIH research ecosystem, supporting innovative and responsible research to improve health outcomes for all Americans.

Contents

Figures

Tables

1 Introduction

1.1 Background

The *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) Office of Data Science and Sharing (ODSS) at the National Institutes of Health (NIH) has developed a robust metadata schema[c] for data governance[d] information relevant to linking individual-level participant data and sharing and using linked datasets for research. The data governance metadata schema implements and extends the Open Digital Rights Language information model and enables the development of tools to collect, standardize, exchange, and visualize information about data governance, including rules from consent, policies, laws, and other sources. The metadata schema allows data governance metadata to travel with data across the lifecycle, promoting appropriate and responsible adherence to governance that addresses requirements such as those related to ethics, privacy protections, compliance, and risk management. With funding from the HHS Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) and the NIH Office of Data Science and Strategy , this effort aligns with NICHD ODSS's larger goal of developing a governance and technology strategy for implementing individual-level record linkage for patient-centered outcomes research with NICHD populations (children, pregnant and lactating women, and people with disabilities), initially driven by pediatric COVID-19 research use cases. Learn more about the metadata schema on the NICHD GitHub Data Linkage Governance Repository.[1]

NICHD ODSS engaged the Health Federally Funded Research and Development Center (Health FFRDC), operated by MITRE, to test the data governance metadata schema through two proof-of-concept implementation projects to: (1) collect governance information, and (2) visualize governance information to support decision making about linking datasets for research. The data governance metadata schema and collection and visualization prototypes will contribute to HHS- and NIH-wide strategic goals for data science, and patient-centered outcomes research. The overall goal is to provide researchers, repositories, funders, and policy/legal experts with high-quality information they can use to determine whether certain datasets can be linked, and if they can be, what rules and controls apply to the linked dataset.

This report focuses on the development of the second proof-of-concept implementation project, the Governance Metadata Visualization Prototype.

The Health FFRDC project team, under the oversight of NICHD ODSS, engaged community experts in the form of a Technical Experts Panel (TEP) to guide the prototype development and subsequent efforts to evaluate the prototype's usability. See Appendix A for TEP membership.

[c] A metadata schema is a structured set of metadata elements and attributes, together with their associated semantics, that are designed to support a specific set of user tasks and types of resources in a particular domain. A metadata schema formally defines the structure of a database at the conceptual, logical, and physical levels.

[d] Governance or data governance comprises the collective set of rules and controls that define and enforce how data are handled across the data lifecycle including: appropriate data collection, sharing, linking, access, and use. Data governance addresses privacy protections, ethics, compliance, risk management, and other requirements and derives from a variety of sources such as participant consent, IRB determinations, laws, agreements, and policy documents.

Alignment with NIH Controlled Data Access Goals

NIH has identified the need to improve efficiency and harmonization among controlled-access data repositories to make NIH data more findable, accessible, interoperable, and reusable (FAIR) and to ensure appropriate oversight when data from different resources are combined. Toward addressing this need, NIH released a Request for Information for public feedback and established an internal working group in 2021 that delivered a series of recommendations for streamlining access to controlled data in NIH data repositories.

These recommendations aim to streamline access and use of controlled-access data across the NIH ecosystem to accelerate research; for instance, by assessing standards for defining consent-based data use limitations, drafting standard data submission and data use certifications for adoption by controlled-access repositories, and identifying the need to protect privacy particularly when linking participant-level data from multiple studies. Implementation of these recommendations would benefit from a harmonized approach to collecting, exchanging, and visualizing information about controlled-access data governance.

1.2 Foundational Governance Work

NICHD ODSS has been leading data governance work since 2022, developing frameworks and tools to support responsible use of individual-level record linkage (privacy preserving record linkage or other linkage methods) for research in support of the NICHD mission.

Privacy Preserving Record Linkage (PPRL) for Pediatric COVID-19 Studies Report[2]

Published in September 2022, this report assessed 13 existing record linkage implementations and developed technical and governance considerations for appropriately linking data. The resulting report summarizes the current state of pediatric COVID-19 studies that could benefit from use of privacy preserving record linkage (PPRL), a method for linking records associated with an individual represented across multiple datasets without exposing any personally identifiable information (PII). The report documents decisions made for existing record linkage implementations, develops and defines considerations for the governance components necessary for enabling PPRL and dataset linkage, and develops considerations for implementing potential PPRL tools. This work also resulted in the publication of the NICHD Record Linkage Implementation Checklist,[3] which guides technical and governance decisions that must be made prior to designing and implementing a strategy for linking data from multiple sources and sharing and using linked datasets for research. The checklist advises that the design of a new record linkage strategy requires funders, researchers, data repositories, and other community members to collaboratively consider all items described in the checklist, such as considering additional controls to mitigate potential risks. The report acknowledged that the checklist item "identify policies that apply to each dataset including rules specific to certain data types or participant populations" requires significant effort, given how difficult it is to identify and interpret dataset-level rules from complex documents and sources. This finding was the motivation for NICHD ODSS to develop a governance metadata schema.

OS-PCORTF Pediatric Record Linkage Governance Assessment[4]

To gather real-world evidence to inform the structure of a new governance metadata schema, NICHD ODSS collected and examined the governance information from 11 HHS and other federally-funded datasets that represent three theoretical pediatric COVID-19 research use cases, following a Governance Information Framework designed to capture information necessary to make a determination about the ability to conduct linkage and the subsequent limitations and controls that would apply to such a linkage. This 2023 report describes the outcome of that governance information collection effort, linkage determinations made for the three pediatric COVID-19 use cases, and key considerations for the development and implementation of a standardized and machine-readable data governance metadata schema.

In the process of collecting governance information, NICHD ODSS uncovered a rich and complex governance information ecosystem, for which no data governance metadata schema previously existed. NICHD ODSS's research also arrived at a select set of findings relevant to this report, including:

- Dataset documentation often does not explicitly authorize linkage or specify the scope of linkage.
- Linked datasets converge on the most constraining requirements.
- Conflicts in governance introduce complexity in defining the approach to linkage.
- Linkage determination must consider how the linked data is deidentified.

Governance Metadata Standards: Landscape and Gap Analysis Report[5,6]

Published in 2024, this report describes the results of a landscape analysis conducted by the Health FFRDC that identified existing data standards that could be used to develop the data governance metadata schema. The analysis consisted of an inventory of existing standards, an assessment of utility of those standards, and a gap analysis based on 11 domains of governance information.

The landscape analysis recommended the Open Digital Rights Language[7] (ODRL) standard and information model as the primary standard to base the metadata schema design on. ODRL is a versatile policy articulation language that offers an adaptable and interoperable data model, vocabulary, and encoding systems for expressing statements about the utilization of content and services. ODRL's foundational elements are policies made up of rules that are employed to denote permitted (allowed) and prohibited (forbidden) actions on a specific asset, as well as the responsibilities that parties are required to fulfill (i.e., obligations). Rules can be subject to constraints (e.g., locations of data access) and duties (e.g., as obtaining approvals) that can be imposed on permissions. This system of policies, rules, parties, and constraints serves as an ideal basis for governance metadata schema development, and a useful representation of most data governance information relevant to linkage.

1.3 Data Governance Metadata Schema

The Health FFRDC and NICHD ODSS project team developed and published a data governance metadata schema[8] in 2024. The schema provides an information model and vocabulary, built with and expanding on ODRL, that is designed to support the collection, annotation, and exchange of governance information in a structured format (see Figure 1).

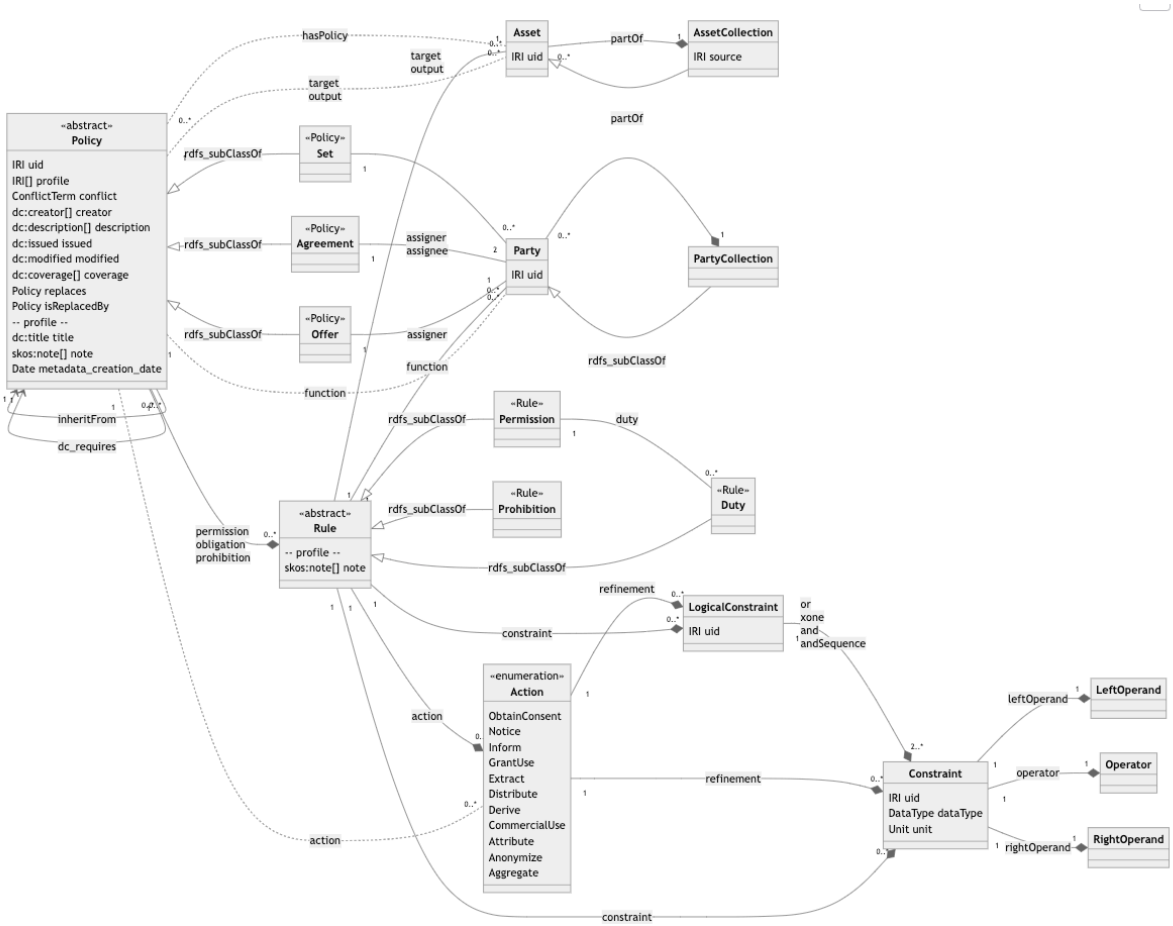


Figure 1: Data Governance Metadata Schema

The data governance metadata schema holds dataset-level governance information where a dataset is modeled as an asset. Each asset has one or more governance policies (e.g., consent form or data use agreement [DUA]) and each policy holds one or more rules. Rules may be permissions, prohibitions, or obligations and each rule contains an action (e.g., permission to *link* or prohibition to *reidentify*). A duty is defined as the requirement to perform an action. Rules may be assigned zero, one, or multiple parties. Rules may also contain constraints that are conditions of the rule's application (e.g., permission to link [rule] if the product is a deidentified dataset [constraint]). Rules can be related to other rules, most often as a permission to [action] with a duty to [action] (e.g., a permission to use data with a duty to obtain approval from an institutional review board (IRB)).

The data governance metadata schema includes extends the ODRL vocabulary, adding more than 70 additional terms and annotations required to accurately represent governance metadata. This new Data Governance Profile (see Figure 2), captures data governance-specific concepts such as policy types of DUA and consent and actions to *reidentify* and *deidentify*. Terms were added to represent policy types, governance actions, and constraints. Profile terms were mapped to existing standards (such as Data Privacy Vocabulary and Health Level 7) when possible.

The schema also adopts the Open World Assumption, meaning it only captures explicitly stated rules and a lack of rules should not imply permission or prohibition for an action.

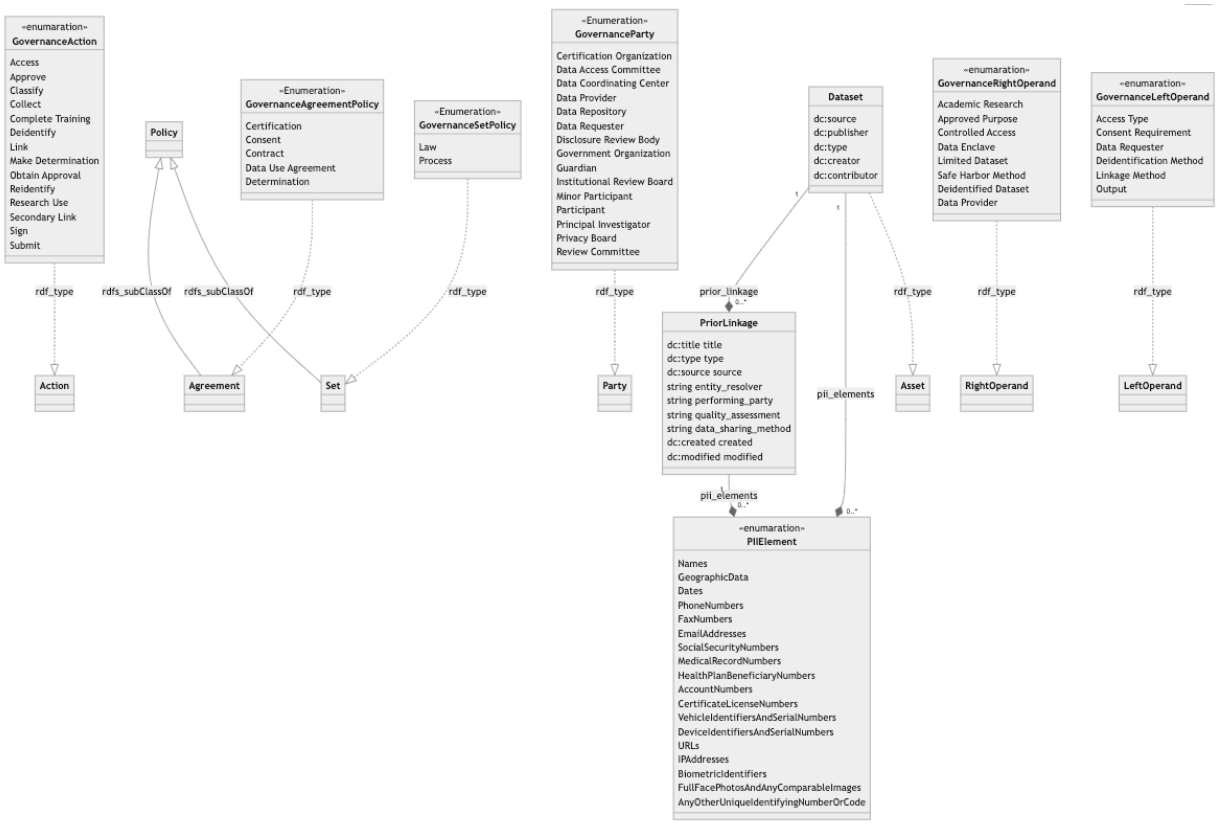


Figure 2: Data Governance Profile

The data governance metadata schema should provide a foundation for tools to collect, exchange, and/or visualize data governance information by defining the contents of governance information to be collected, a structure and design that collected governance information will be aligned to, and a vocabulary that may be used to describe governance. Notably, the schema vocabulary is not fixed; the schema will evolve over time, potentially expanding its vocabulary of terms to represent more governance concepts.

Testing is essential for the data governance metadata schema's improvement, adoption, and sustainability across the NIH and HHS data ecosystems. To test the metadata schema, NICHD ODSS has piloted two prototype tools: (1) a data collection tool for entering and sharing dataset-level governance information as metadata, and (2) a visualization prototype to learn how datasets of interest might be linked and used based on their governance.

1.4 Purpose

The purpose of this report is to describe the end-to-end approach to a proof-of-concept implementation project to develop a data governance metadata visualization prototype and underlying database to test the data governance metadata schema. The report describes the methods and approach to develop the prototype, key findings from the development and usability evaluation, and recommendations for the metadata schema and future governance metadata visualization work.

The objectives of the proof-of-concept project are twofold:

1. To explore how governance information collected from real-world datasets, encoded with the data governance metadata schema, and stored in a relational database can be visualized and to ascertain which governance information is the easiest and most challenging to display.
2. To test how the data governance metadata schema performs in a data visualization prototype designed to support parties in making decisions about a proposed dataset linkage implementation for research. Testing the metadata schema is focused on five questions about how researchers, as proxies for a broader user group including repositories, funders, and policy/legal experts, understand and assess governance information:
    - Can governance metadata be transformed into accurate and human-readable governance information?
    - Can governance information derived from metadata be consumed and interpreted accurately by a researcher?
    - Can a researcher make an assessment about the feasibility of dataset linkage for research?
    - Does a researcher understand the conditions that must be met to link datasets for research?
    - Does a researcher understand what actions are required to link datasets for research?

1.5 Audience

The intended audience of this public report includes: (1) researchers planning to link datasets from multiple studies or programs , including researchers and data scientists across HHS and NIH agencies (2) stewards of data repositories that accept and expose governance metadata for datasets they host, (3) policy experts aiming to streamline the search of, access to, and responsible linkage and use of datasets, and (4) the patient-centered outcomes research community.

2 Approach and Methods

The project team, including Health FFRDC experts in biomedical research, data governance, informatics, metadata, standards, and software engineering, developed and evaluated the proof-of-concept visualization prototype through these steps:

1. Constructing a data governance metadata database and basic query tool
2. Generating capability statements that describe the users' objectives
3. Developing an early prototype using agile principles and methods[9]
4. Refining the prototype with co-designers based on user-centered design principles and methods[10]
5. Completing prototype development based on feedback from co-designers and TEP members
6. Conducting a usability evaluation using existing frameworks such as the Reach Effectiveness Adoption Implementation Maintenance Reach (RE-AIM) framework[11] and the extended unified theory of acceptance and use of technology (UTAUT)[12]

The project team worked in collaboration with NICHD ODSS, regularly sought guidance from the TEP, and defined a user story to guide the prototype development, described in Table 1.

Table 1: User Story to Guide Prototype

User Story: What does the user want to do?	Current Problem: Why can't the user do this today?	User Goal: What is the user's ultimate goal?
As a small business innovation researcher, I want to study the success of algorithms developed for diagnosing COVID in children by linking/combining research data on these diagnostics with electronic health records and claims data, so I can extend selected algorithms for the diagnosis of other pediatric infectious disease.	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 is 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.

2.1 Data Governance Metadata Database

The governance visualization prototype requires a source of schema-encoded governance metadata for visualization. As part of the OS-PCORTF Pediatric Record Linkage Governance Assessment,[13] NICHD ODSS collected the policies, limitations, processes, and controls for the 11 selected datasets in a Microsoft Excel workbook structured with the Governance Information Framework (herein referred to as the source governance information and which is Appendix D of the OS-PCORTF Pediatric Record Linkage Governance Assessment). The 11 datasets are National Health and Nutrition Examination Survey[e] (NHANES), National Survey of Drug Use and Health [f] (NSDUH), Monitoring the Future[g] (MTF), Adoption and Foster Care Analysis and Reporting System[h] (AFCARS), National Childhood Cancer Registry[i] (NCCR), Centers for Disease Control and Prevention COVID Data Tracker[j], Transformed Medicaid Statistical Information System[k] (N3C), National COVID Cohort Collaborative[l] (N3C), PEDSnet[m], COVID Rapid Acceleration of Diagnostics (RADx) Data Hub[n], and Environmental Protection Agency Air Quality System[o].

In the Excel spreadsheet, NICHD ODSS recorded the policy names and types, raw policy language such as the words from a consent that permit dataset sharing for each phase of the data lifecycle, dataset

[e] NHANES Website: https://www.cdc.gov/nchs/nhanes/index.htm
[f] NSDUH Website: https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health
[g] MTF Website: https://monitoringthefuture.org/
[h] AFCARS Website: https://www.acf.hhs.gov/cb/data-research/adoption-fostercare
[i] NCCR Website: https://cancercontrol.cancer.gov/research-emphasis/supplement/childhood-cancer-registry
[j] CDC COVID Data Tracker Website: https://covid.cdc.gov/covid-data-tracker/#datatracker-home
[k] T-MSIS Website: https://www.medicaid.gov/medicaid/data-systems/macbis/transformed-medicaid-statistical-information-system-t-msis/index.html
[l] N3C is now named National Clinical Cohort Collaborative; N3C Website: https://ncats.nih.gov/research/research-activities/n3c/overview
[m] PEDsnet is Clinical Research Network in PCORnet®; PEDSnet website: https://pedsnet.org/
[n] RADx Data Hub Website: https://radxdatahub.nih.gov/
[o] Environmental Protection Agency Air Quality System Website: https://www.epa.gov/outdoor-air-quality-data

collection, linkage, sharing, access, and use. The project team constructed a relational database containing the real-world governance information collected by NICHD ODSS transformed into governance metadata that reflected the metadata schema design.

Annotation

The project team began translating real-world governance information into metadata by examining the source governance information and experimenting with approaches to dissect and decompose unstructured text governance information to produce a digital representation of the Governance Information Framework in a relational database.

Translation of governance information to metadata required an intermediary format that could be processed by simple tools into a schema representation of the governance information. Natural language processing tools were unavailable for use. The project team developed and implemented a manual annotation method for policies, raw policy language, and policy interpretation from the source governance information. The project team added an annotation column to the Excel workbook and assigned governance information to data governance metadata schema classes and terms using annotation tags. The project team then loaded the annotated data into a relational database using an import script that translated the annotations to match the schema.

Query Tool

The project team built a basic query tool to review the annotation output alongside the source information (i.e., raw policy language and policy interpretation) and validated the accuracy of the annotations as well as the overall annotation approach.

Technical Development

PostgreSQL was chosen as the relational database for storing data governance metadata based on the following criteria:

- Availability as a free and open-source tool, released under the PostgreSQL license
- Well-understood performance and scaling characteristics
- Wide adoption and strong community support
- Data modeling flexibility, including support for JavaScript Object Notation (JSON) datatypes
- Development flexibility, including support for dozens of client libraries and robust query capabilities

The Ruby on Rails framework was chosen to provide infrastructure for managing the database, specifically the following capabilities:

- Database migrations for managing the database schema in alignment with the data governance metadata schema
- Extension libraries for producing project artifacts such as data model diagrams
- Task management capabilities for importing annotated governance information into the database, validating consistency of imported data, and tracking provenance of imported data

- Web interface development capabilities to support development of a query tool and eventually Application Programming Interfaces (APIs) to support the visualization prototype

2.2 Visualization Prototype

The project team developed the prototype governance visualization web tool by iteratively refining it based on the user story and the team's exploration, with simple wireframes and mockups to outline the user interface and data flow. As the project team developed the prototype, the implementation of the database was evaluated by using the real-world governance information. The project team focused on creating a user-friendly and intuitive experience and incorporating features to help the user explore governance information at their own pace. The project team tested and validated the database and prototype functionality regularly throughout its development to identify and address any issues.

The prototype was designed for users to visualize dataset governance information and make their own assessment on the potential for a linkage implementation. It served as a minimum viable product for co-designers to engage with and provide feedback.

The project team drafted seven capability statements to describe a user's priorities and expectations (i.e., in terms of features and functions) for the visualization prototype Table 2. Each capability statement was then translated to a function that the prototype would be expected to support. Five prototype functions were named and matched to each capability statement: select dataset, compare governance, assess linkage feasibility, view action steps, and user experience

The project team sequenced capability statements to explore and plan the workflow through the prototype functions. Capability statements guided prototype development with iteration and refinements to prepare for feedback co-designers.

Table 2: Capability Statements Mapped to Visualization Prototype Function

Capability Statements	Prototype Function
1. As a user, I want to learn basic information about datasets that can be included in a linkage implementation.	Select Datasets
2. As a user, I want to select the datasets that I would like to include in a future hypothetical linkage implementation.	Select Datasets
3. As a user, I want to see the number of policies categorized by policy type so that I can visualize the quantity and diversity of governance information sources. This presentation helps me understand how much governance information there is and where governance rules come from for each dataset. It also helps me know where to go next as I want to dive a level deeper into the rules.	Compare Governance
4. As a user, I want to see permissions, prohibitions, and obligations (with constraints) categorized by policy type so that I can see what the rules are for this dataset (e.g., there are probably between 3 and 40 rules for a given dataset) and where those rules are from. As I user I want rules to be linked to raw policy excerpts, when available.	Compare Governance

Capability Statements	Prototype Function
5. As a user, I want to see the rules that apply to dataset linkage, sharing, access, and use. I want the rules to be deduplicated, and I want to see the conditions for each data lifecycle action. I want to know which policies the rules originate from.	Assess Linkage Feasibility
6. As a user, I want to see the rules presented by assigned party to understand who these rules apply to. I specifically want rules applied to "no party" to be included (because they apply to everybody).	View Action Steps
7. As a user, I should be able to navigate and use this prototype with zero knowledge of the data governance metadata schema.	User Experience

The project team utilized the capability statements to guide prototype development. The team brainstormed various visualizations to present governance information such as matrices and diagrams in the context of selecting datasets, comparing governance, assessing linkage feasibility, and viewing action steps. The project team reviewed existing web tools with similar functionality for ideas about how to represent governance information in an innovative way. NICHD ODSS and other members of the project team gave feedback and selected the clearest visual representation for implementation.

The project team selected a color scheme and icons for every action that appeared in the governance metadata. To ensure the prototype was accessible, the project team reviewed and followed The Web Content Accessibility Guidelines published by the Web Accessibility Initiative of the World Wide Web Consortium, (W3C).

The visualization prototype includes six pages: Home, About, Select Datasets, Compare Governance, Assess Linkage Feasibility, and View Action Steps. The project team collaborated with NICHD ODSS to draft text for the home and about pages and instructions for each of the four main visual pages (Select Datasets, Compare Governance, Assess Linkage Feasibility, and View Action Steps). The prototype is designed for the user to move in a sequential and intuitive workflow, starting at the home page, gaining context on the prototype's purpose, and then select datasets to compare governance, assess linkage feasibility, and view action steps.

Technical Development

The visualization prototype was developed as a client-side web application providing visualization capabilities and a user interface driven by data from the Data Governance Metadata Database. Connectivity between the visualization prototype and the database was implemented by extending the Ruby on Rails application used to develop the database query tool with a simple JSON API for retrieving the required dataset governance information. While the visualization prototype was used primarily with the specific set of datasets represented in the database, the use of a JSON API means that the database and visualization prototype components are not tightly coupled and allows for the potential of using the visualization prototype with other data sources.

The visualization prototype user interface was developed as a JavaScript web application using the React user interface library, a JavaScript library designed for building interactive user interfaces, accompanied by the Material User Interface component library, a library of React user interface elements. For ease of

system coordination, the React visualization prototype was built within the Rails application; however, these components can be easily decoupled if desired.

Prototype elements included core prototype functions as defined using the capability statements along with related supporting elements such as an about page and a glossary of terms. Prototype development also included the creation of "humanization" utility functions intended to convert the schema representation of policies, rules, duties, constraints, functions, and parties into easily interpretable language intended for human consumption.

System Architecture and Deployment

The visualization prototype is deployed as a small set of Docker containers on a single Amazon Web Services (AWS) Enterprise Cloud Computing (EC2) instance (Figure 3). Docker is a platform that allows applications to be configured, tested, and deployed by packaging application components into containers. The AWS EC2 instance was accessible through an application gateway service, Web Application Rapid Prototyping (WARP), which provides user management and access control capabilities. Only invited users could access the prototype.

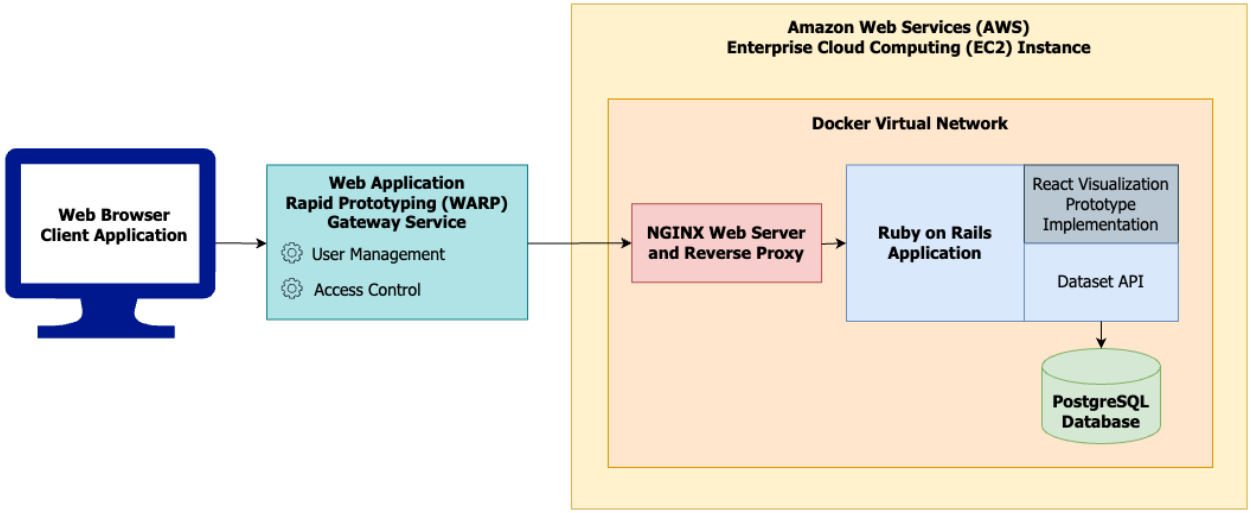


Figure 3: Visualization Prototype System Architecture

The prototype is organized into three separate docker containers:

- An NGINX web server that acts as a simple reverse proxy, orchestrating connections between the WARP service and the Ruby on Rails application
- A Ruby on Rails application that orchestrates access to the React Visualization Prototype App and provides a dataset API that pulls data from the PostgreSQL database
- A PostgreSQL database that contains the data governance metadata

Co-design with Researchers and Informaticists

The project team engaged five biomedical researchers and informaticists in a series of co-design sessions to support the prototype's development (see Appendix B). Co-designers were invited to attend three one-hour co-design sessions to provide feedback on the prototype during and between meetings.

The project team sent an agenda ahead of time, and the sessions were recorded. Co-design sessions displayed the prototype at various stages of development. Co-designers were provided access to the prototype through the MITRE WARP environment. The project team took notes during co-design sessions to capture feedback about expressed concerns/observations and made interval changes to the prototype between the sessions.

Members of the project team took notes during co-design sessions and captured user feedback about their expressed concerns and observations. The team reviewed, analyzed, and prioritized the feedback for consideration for implementation by:

- Labeling the concern or observation with a topic category (e.g., the page to which it applied)
- Noting specific proposals for change to the prototype
- Rating each suggested change for level of effort and alignment with the project's user story
- Assigning each suggested change as a candidate for immediate implementation or for future enhancement
- Gaining additional feedback through interval discussion with NICHD ODSS and the TEP
- Previewing new versions of the prototype during each co-design session, gaining additional feedback

2.3 Usability Evaluation

The project team conducted a usability evaluation, with a small group of user testers to evaluate whether the governance metadata data visualization prototype was easy to understand and use and to gather feedback for future enhancements. The project team designed a usability test guided by human-computer interaction methods to understand how the prototype performs in a realistic setting.[14] The MITRE IRB reviewed the procedures for this evaluation and deemed them exempt from human subjects' review.

During usability testing, testers representing a segment of the anticipated user population were observed visualizing real-world dataset governance information in the prototype, and guided through six pages:

- Welcome and home page
- Select datasets page
- Compare governance information page
- Assess linkage feasibility page
- View action steps page
- About page and glossary

The evaluation framework was guided by concepts from the RE-AIM framework and the UTAUT. Questions asked during the user tester sessions included:

- Are the instructions in the prototype easy to understand?

- Is the terminology used to describe dataset governance and linkage implementation understandable?
- Is the prototype organized in an intuitive way?
- Do the visualizations facilitate an understanding of how to implement dataset linkages?
- Does the prototype require substantial time and effort to use?
- Do users feel confident determining if a linkage implementation is practical and feasible?
- Would researchers and research institutions view this prototype positively?

The project team's user-centered design expert formulated the usability evaluation data collection and analysis procedures and facilitated each 1:1 user tester session. Two additional team members with experience in qualitative research methods attended all sessions to take notes and collaborated on the analysis.

Recruitment

The NICHD ODSS team recruited a convenience sample of six experienced biomedical researchers and policy analysts, all federal employees, as user testers to participate in the usability evaluation Table 3. NICHD ODSS selected user testers who were familiar with dataset linkage and who represent different domains of biomedical research.

Table 3: Participating Biomedical Researchers and Policy Analysts in Usability Evaluation

Tester Number	Background	Years Involved with Research Data	Biomedical Field of Research	Experience Linking Datasets	Involvement with Data Consortia
1	Privacy Board Research Review	7	Health Services	Yes	No
2	Research and Administration	18	Infectious Disease and Protein Trafficking	Yes	Yes
3	Research	20	Data Science	Yes	No
4	Scientific Policy	13	Genomics, General Biomedical Research	No	No
5	Research and Administration	20+	Maternal Health, Genomics, Data and Technology	Yes	Yes
6	Policy Analyst	8	Genetics and Ethical, Legal, and Social Issues	No	Yes

Orientation

The project team conducted an introductory virtual half-hour group meeting to introduce the user testers to the project and the governance metadata visualization prototype. The project team presented the rationale and goals of the governance metadata project, a description of the underlying data governance metadata schema, a description of the visualization prototype, and the expected sequence for the 1:1 user tester session. The user testers did not view the visualization prototype during this session.

User Tester Sessions

The project team conducted one 90-minute usability evaluation session with each user tester through Microsoft Teams. At the beginning of the session, the facilitator introduced the two project team members as observers and notetakers, reviewed the session procedures, defined key terms, and asked the user testers basic questions about their research and data linkage experiences.

The facilitator then directed the testers to log in to the visualization prototype via MITRE's hosted WARP environment and instructed them to share their screen so that the facilitator and notetakers could observe their actions and to "think aloud" and verbalize their thought process as they navigated the prototype. The facilitator guided user testers through a scenario where the testers were instructed to explore the governance information for a hypothetical linkage including three datasets: PEDSnet, RADx, and T-MSIS. The facilitator prompted user testers to navigate to the different screens in the prototype and asked questions about the information displayed. Testers were asked to verbalize when they had completed the section or when they determined that they could not complete the section to gather timing information. Notetakers documented the user testers' responses and response accuracy to a set of questions, such as these examples:

- Which dataset has the largest number of relevant "Policies," and how many "Policies" does it have?
- If you want to link these three datasets together, how many "Consent" type policies do you have to consider?
- Looking at the Action Steps for the Data Requestor, how many documents need to be signed to link the RADx database?

After the user testers navigated through all of the visualizations on the screen, the facilitator asked them to rank the feasibility of linking the three datasets on a 10-point scale, with 1 meaning that linkage is trivial and easy, and 10 meaning that linkage is not feasible or possible.

At the end of the session, the facilitator asked the user testers discussion questions about their experience reviewing governance information, the perceived usefulness of the prototype, the perceived ease of use of the prototype, its ability to help the user tester perform their work, and whether they thought the prototype would be adopted by researchers in their field. The facilitator also asked which aspects of the prototype aided information review and interpretation, what challenges the participant faced as they were reviewing information, and if they had any suggestions for future enhancements.

Discussion questions were informed by elements of the RE-AIM framework and the extended UTAUT. Questions were a mixture of statements where testers rated their agreement or disagreement using a four-point Likert scale and open-ended inquiries to encourage dialogue. Refer to Appendix C for the session script and evaluation questions.

Analysis

After each usability evaluation session concluded, the project team analyzed the session transcripts, notes, and discussion question responses, and analyzed questionnaire data using quantitative and qualitative methods. The team's video recording served only as a back-up as needed for analysis, and these recordings were discarded within two weeks of each session.

The project team created a preliminary codebook to support the analysis based on concepts from RE-AIM, the UTAUT, and important concepts identified in prior usability evaluations for a governance metadata collection tool. Refer to Appendix D for the codebook. Two team members reviewed the transcripts, notes, and discussion question responses from all sessions and applied codes from the codebook or codes for emerging themes inductively generated from the text as needed. Team members also applied structural codes to categorize data by the four main visuals in the prototype. After all the text was coded, project team members met to review code applications and generated themes. The project team estimated the time user testers took to review each page in the prototype by calculating when the tester started and ended reviewing each page. The project team also summarized question responses and calculated the percentage of user testers who provided the expected answers.

3 Outcomes and Findings

The project team generated findings based on the development of the data governance metadata database and visualization prototype and the evaluation of the prototype through usability testing.

3.1 Data Governance Metadata Database

The project team developed, tested, and implemented the annotation methodology to convert governance information into metadata while also constructing a relational database to store that governance metadata.

The project team then connected these two processes by applying the annotation approach to generate metadata, then loading that metadata into the relational database. Annotation and database development occurred in parallel with the schema design, iterating between the schema and database to inform the schema design and vocabulary.

Annotations

The project team categorized governance information in the sources as dataset information, history of linkage, or policy information. All three categories contained non-standardized textual representation with some policy content duplicated throughout the source documentation.

The project team developed an annotation approach that isolated discrete governance metadata elements and tagged those elements to schema classes and terms. Annotation of dataset information and history of linkage was straightforward as each row in the source governance information mapped to one corresponding schema class that held unstructured text. Policy information was varied and required the project team to manually decompose the raw policy information and policy interpretation into policies, rules, duties, constraints, and parties.

The initial annotation method was expanded to allow for more terms that were added to the schema, differentiation between duties and obligations, and annotation across multiple policies within a single cell. This improved the information to metadata translation and enabled alignment with synchronous changes to the schema.

Database

The project team defined a database data model based on the ODRL information model, iteratively revising the database to reflect the schema as it was finalized. Figure 4 presents the database data model displaying schema classes as a database table and schema terms as attributes in a table.

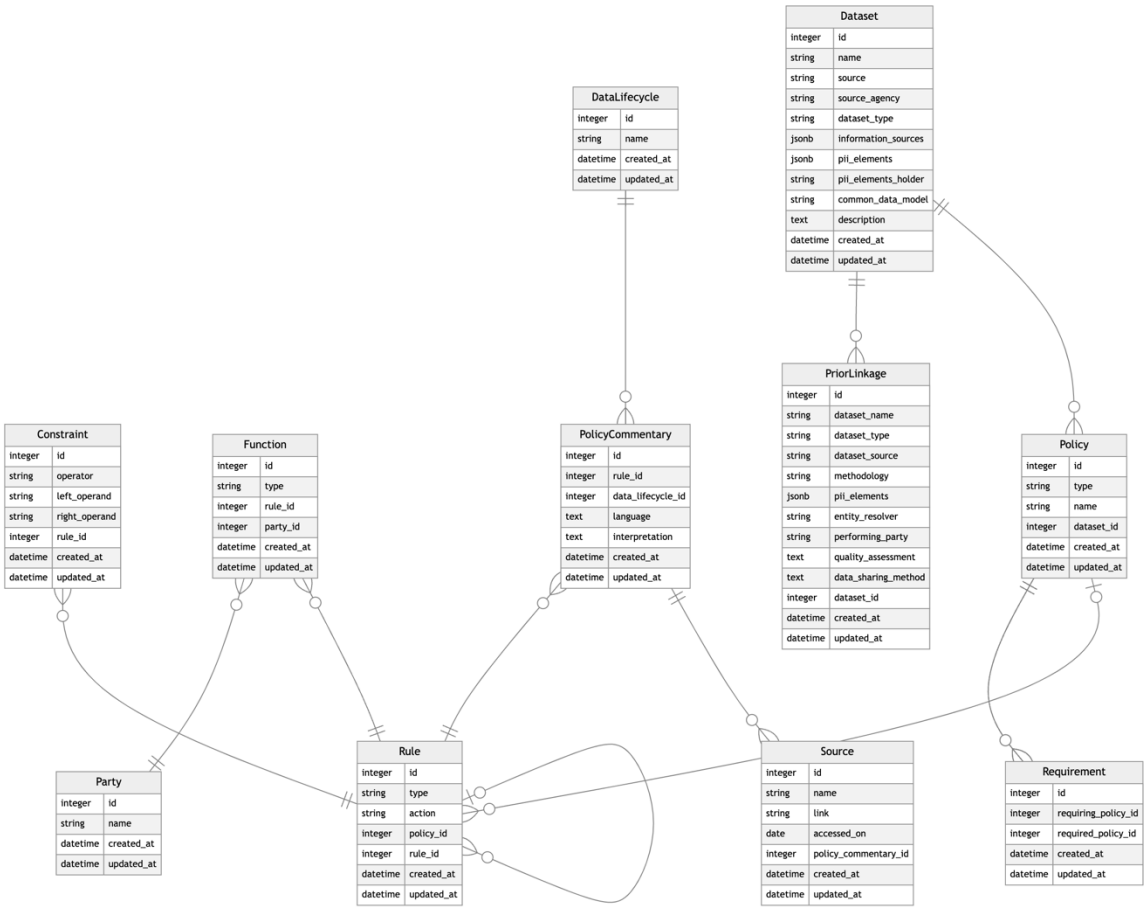


Figure 4: Data Governance Database Data Model

Transformation of Governance Information into Metadata

Two members of the project team annotated the Excel file that was structured in the Governance Information Framework and revised annotations to improve accuracy and completeness through three rounds of revision.

Dataset information and history of linkage contained varying levels of standardization and detail and were imported as exactly as documented into the database. No standardization was performed; therefore, database attributes may contain similar non-standard values (e.g., patient name vs. last name).

Policy information required the project team to annotate policies, rules, parties, and constraints. The project team annotated policies with a policy name and policy type. For some policies, the policy name

was obvious such as laws or DUAs, where the law name or agreement name was annotated as the policy name. Other policies' names were unclear, and the project team created a policy name based on the function or application of the policy or rules within that policy. One example is the N3C External Dataset Classification policy (Figure 4). Similarly, some policies had an obvious policy type (e.g., DUA, Determination, Contract, Certificate, Law) while other policies were less clear. The project team established that when unclear, any policy could be set to policy type=policy.

Rule annotation included the policy's permissions, prohibitions, or obligations and the action and any duties associated with the rule. Within each policy, the project team created one or multiple rules based on raw policy language and policy interpretation. Examples are permission to link, prohibition to reidentify, obligation to deidentify, or permission to use with a duty to obtain approval from the IRB.

Constraint annotation means annotation of a condition on a rule using three parts: left operand, operator, and right operand. A permission to link with a condition that linkage method must be privacy preserving record linkage, or PPRL, is an example of a rule with a constraint where left operand is linkage method, the operator is eq for equals, and the right operand is PPRL.

Party annotation includes the party and function that are assigned to a rule. Parties include principal investigator, IRB, privacy board, minor participant, and a review committee. Functions include assigner, assignee, consenting party, consented party, approving party, and approved party. A permission to link may have a consenting party of participant and a consented party of principal investigator because a participant in a study provides a permission to use their data to the principal investigator by signing the consent form.

Each row in the source governance information Excel sheet may contain one or multiple policies; thus, a delimiter (####) was used to mark the completion of one policy annotation and the initiation of a second policy annotation and to segment the raw policy language and policy interpretation accordingly. Table 4 presents example annotations with one policy and multiple policies.

Table 4: Example Governance Information Annotation

Raw Policy Language	Policy Interpretation	Annotation
Consent template for "Home Interview Consent": "Health research using NHANES can be enhanced by combining your survey records with other data sources. The data gathered are used to link your answers to vital statistics, health, nutrition, and other related records. "We can do additional health research by linking the interview and exam data of everyone listed under 'SP NAME' in the gray box below to vital statistics, health, nutrition, and other related records. May we try to link these survey records with other records? Yes, No, N/A."	Consent from adults authorizes data linkage	policy: NHANES Consent policy_type: Consent assigner: Guardian assignee: PrincipalInvestigator permission: link
N3C does not contain direct identifying information, and additional measures have been put in place to protect patient privacy. As a result, the National Center for Advancing Translational Science (NCATS) received a	Two IRB determinations authorize data sharing:	##### policy: N3C IRB Determination Policy policy_type: Determination

Raw Policy Language	Policy Interpretation	Annotation
waiver of consent from an NIH Institutional Review Board, conforming to the Federal Policy for the Protection of Human Subjects (“Common Rule”). Transfer of data from participating institutions to the NCATS N3C platform covered under a cIRB, unless an institution chose to utilize its own IRB for data transfer to NCATS.	1. Institutional IRB or external central IRB approval for transfer of data from participating institutions to the NCATS N3C Platform 2. NIH IRB waiver of consent for sharing through the NCATS N3C Platform	assigner: IRB assignee: PrincipalInvestigator permission: share ##### policy: N3C NIH Review Board Determination and Waiver of Consent policy_type: Determination assigner: IRB assignee: PrincipalInvestigator permission: share

Policy commentary and raw policy language were imported into the database as unstructured text. The source governance information Excel sheet notes many policies that are silent on dataset linkage, access, sharing, and use. The database annotation methods adopted the schema-recommended Open World Assumption, which meant that only policies with explicit governance rules were annotated. Policies with no rules about dataset linkage, sharing, access, or use were not annotated and were therefore excluded from the database and subsequently the visualization prototype.

Annotation occurred in parallel with schema development. As the schema design and vocabulary evolved, annotations were updated with new terms from the schema. Iteration between the database and data governance metadata schema was intentional so that real-world data governance information would inform schema development. Annotations were refined multiple times to improve the digital representation of governance metadata by incorporating additional schema terms, aligning annotation methods between annotators, and enriching annotation with additional metadata to improve accuracy.

The project team completed a first annotation of the governance information using the policy interpretation text as the sole information source. However, annotation revisions incorporated additional detail from raw policy language to improve annotation accuracy and completeness. For example, policy interpretation was often missing party information that was more completely articulated in the raw policy language.

For the prototype, perfect metadata fidelity was not required. The project team refined the annotation approach when they noted a pattern of imperfect annotation due to limitations imposed by the annotation approach. For example, initially the database extrapolated the action in a rule based on what section of the spreadsheet the annotation originated from, as the source governance information was organized by data lifecycle phases (e.g., any rule originally contained in the Data Sharing section of the spreadsheet was extrapolated to apply to the action “share”). The project team identified repeated instances where the extrapolated action was inaccurate, and annotations were expanded to include actions for every rule entry. Additionally, the annotation methods could not accommodate one instance where two constraints need to be connected by an or, which is supported by the schema design.

Query Tool

The query tool was developed as a precursor to the visualization prototype to enable the project team to explore and demonstrate what types of queries for governance information users may be interested in. The query tool presented coded metadata in a Yet Another Markup Language format (Figure 5). Users were able to select one or multiple datasets, and the query tool presented all governance metadata for those datasets (Figure 6). Users were able to filter governance metadata by data collection, data linkage, data sharing, data access, and data use. Because the query tool had multiple options in how to organize governance information for viewing (by lifecycle action or by policy), query tool users were able to explore and select which presentation of governance information best suited their needs.

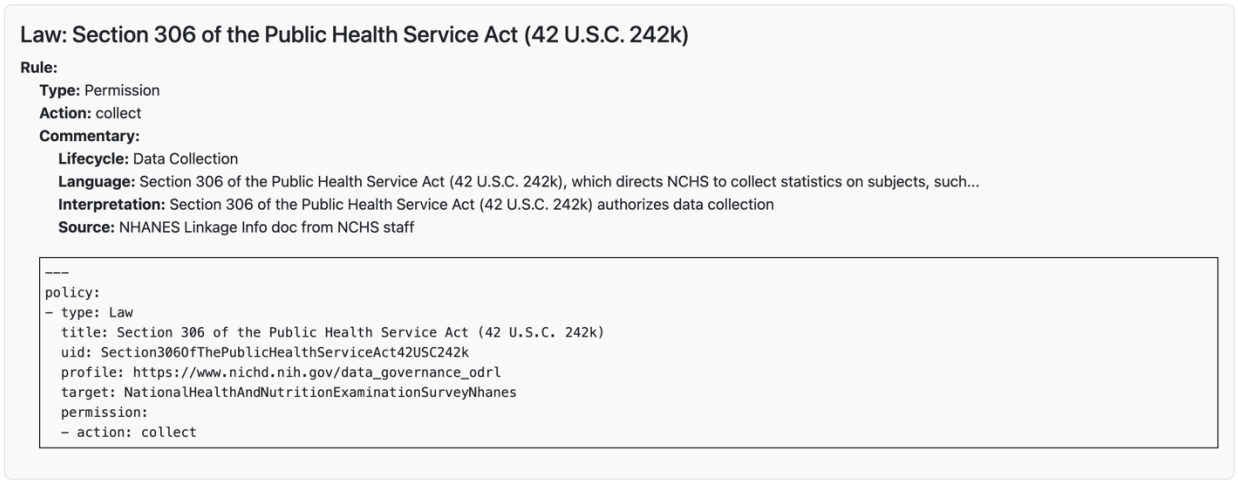


Figure 5: Screen Capture: Database Query Tool, Single Dataset View

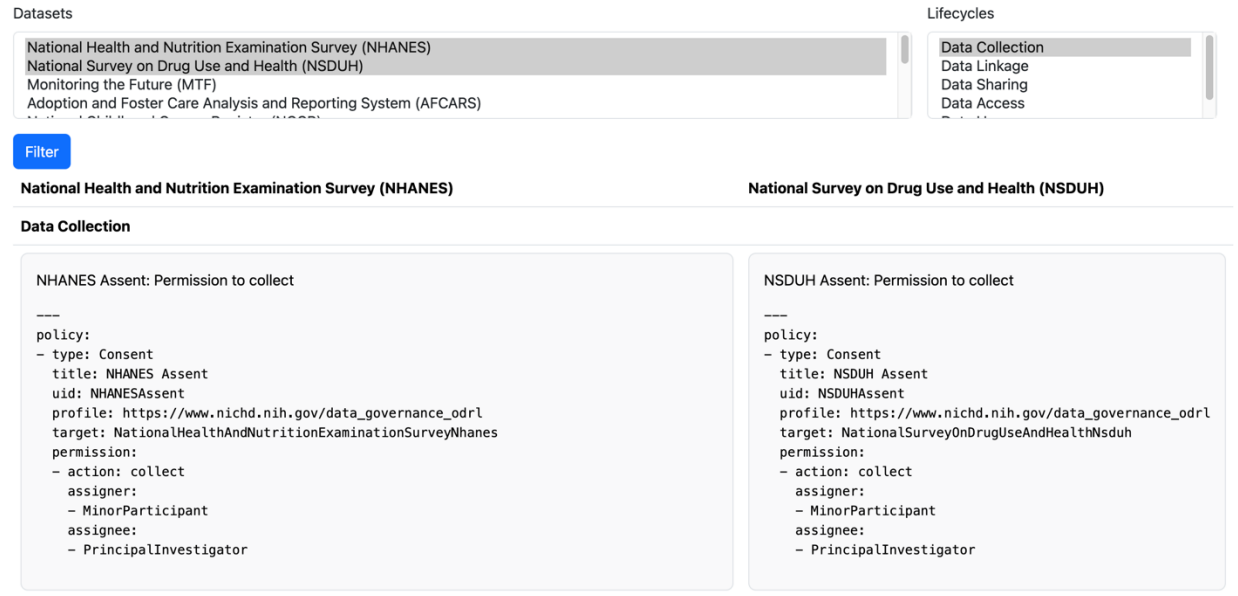


Figure 6: Screen Capture: Database Basic Query Tool, Multiple Dataset and Filter View

Deviations from the Schema

The governance metadata relational database and data governance metadata schema are approximately 95%, but not 100%, aligned. In particular, the source governance information (from the 11 selected datasets) contained two concepts that were loaded into the database but cannot be represented by a schema class: raw policy excerpts and policy interpretation. Co-designers and TEP members shared feedback that these types of information are valuable for decision making about linkage. Therefore, the project team annotated and load raw policy excerpts and policy interpretations from the source governance information into the database and include them in the visualization prototype, even though they had no mappings to the schema.

3.2 Visualization Prototype

Prototype Development

Figure 7 displays the Home page. The home page's purpose is to convey that users of this prototype have already chosen datasets to link for research so that users do not expect the prototype to serve as a dataset search tool, facilitating de novo dataset discovery. The prototype is designed to complement or extend existing search tools. The home page was added late in the prototype development in response to confusion about whether the prototype was also a dataset search tool. This page prepares the user to take the first action of selecting datasets by clicking on the Select Dataset banner or button at the bottom, which reads *Get Started By Selecting Datasets*. A searchable glossary is available on all pages and is visible in the upper right corner of all prototype pages.

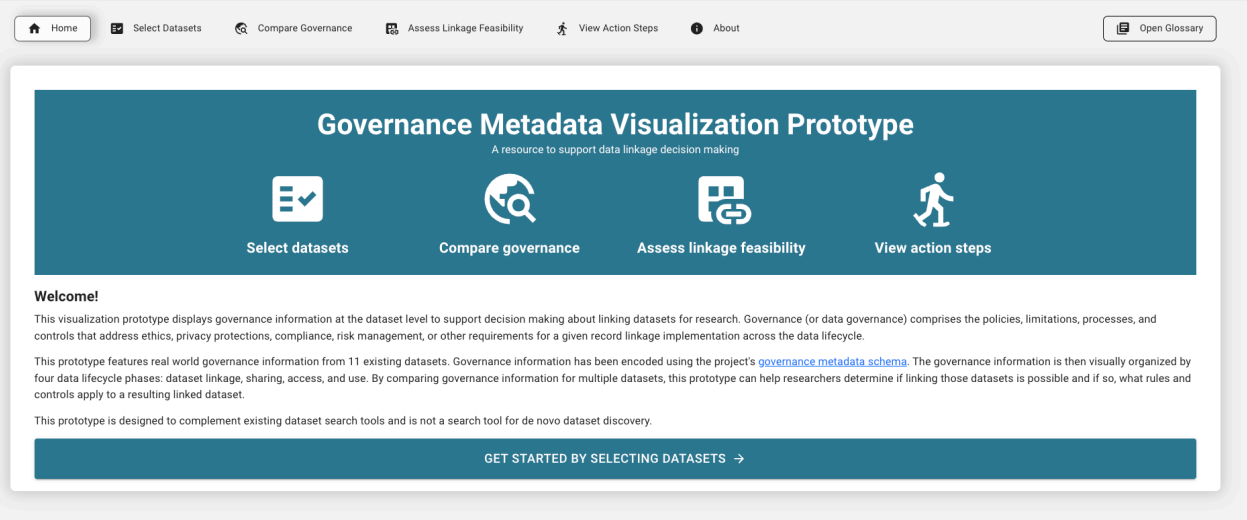


Figure 7: Screen Capture: Home Page

Figure 8 displays the select datasets page. Dataset information is presented to the user; each dataset has a tile that previews the dataset description and color-coded policy chips that show counts of policies by policy type. The Select Dataset page supports the user in selecting up to five datasets that they would like to link. If scrolling through datasets is challenging, a search box allows the user to filter datasets on key words in the dataset name or description. Select Dataset defaults to no dataset selections, and because Compare Governance, Assess Linkage Feasibility, and View Action Steps require at least one or two datasets to be selected, those visuals are inactive in the toolbar until selections are made. Dataset selections persist to compare governance, assess linkage feasibility, and view action steps.

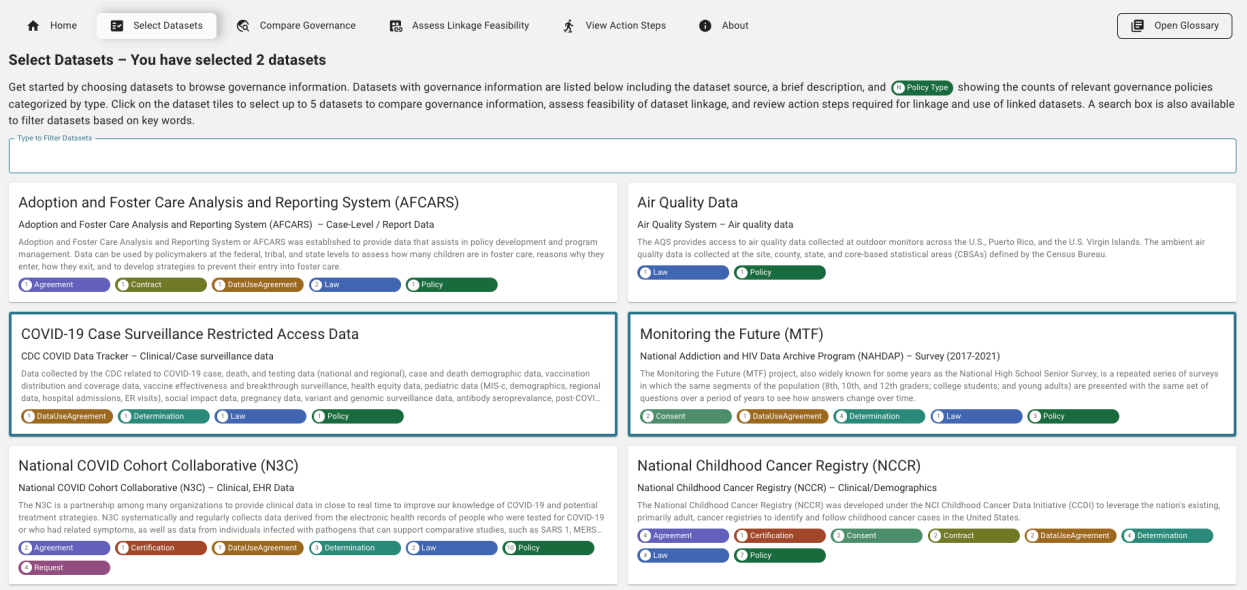


Figure 8: Screen Capture: Select Datasets Page

Upon selecting datasets, the user is guided next to Compare Governance. The purpose of Compare Governance is to review the policies that apply to one or multiple datasets, organized by policy type. Compare governance presents information in three layers: a collapsed view, expanded view, and policy detail view. Figure 9 presents the collapsed view for four datasets, which is the default view.

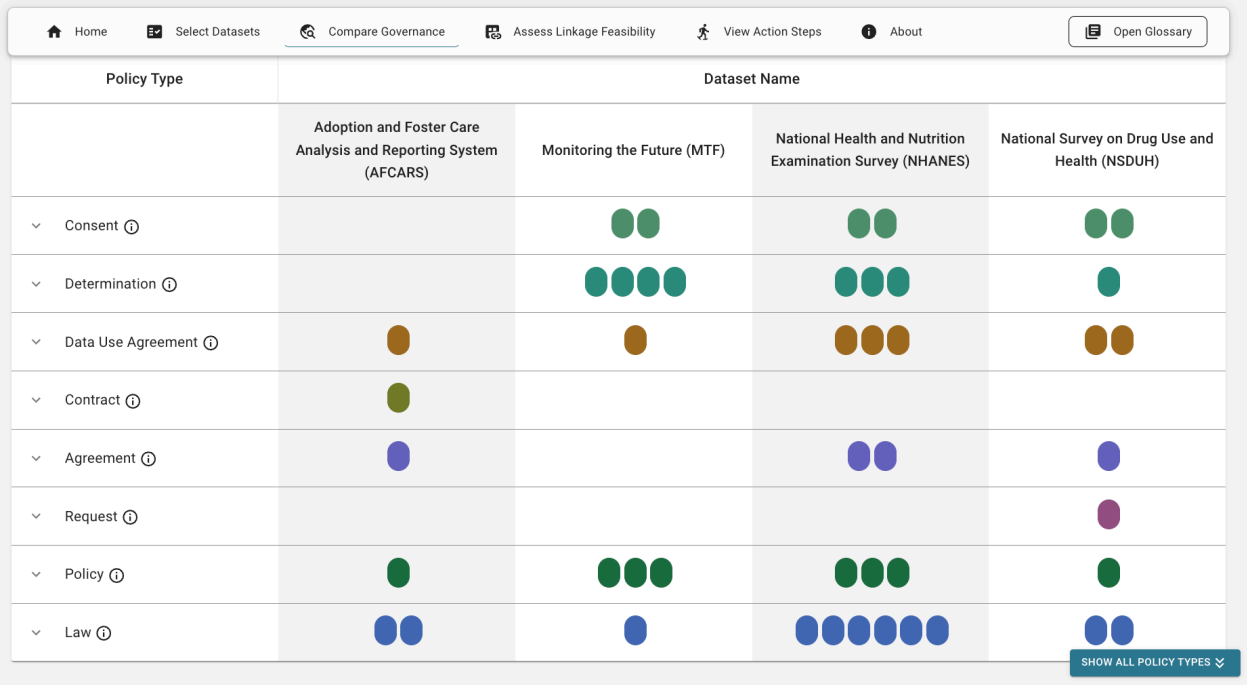


Figure 9: Screen Capture: Compare Governance Page, Collapsed View

The collapsed view aims to orient users to the volume and diversity of governance policies for the selected datasets, categorized and color-coded by policy type. Each policy type is linked to a definition in

the glossary; when a user clicks on the information button next to a policy type, the glossary will open on the right-hand column and provide a definition. Each pill-shaped icon ("pill") in the collapsed view is one governance policy. Figure 10 displays the expanded view. Users navigate from the collapsed view to the expanded view by clicking Show All Policy Types in the bottom right corner. Alternatively, the user can expand one or multiple policy type rows manually with an up and down expansion arrow (^) function.

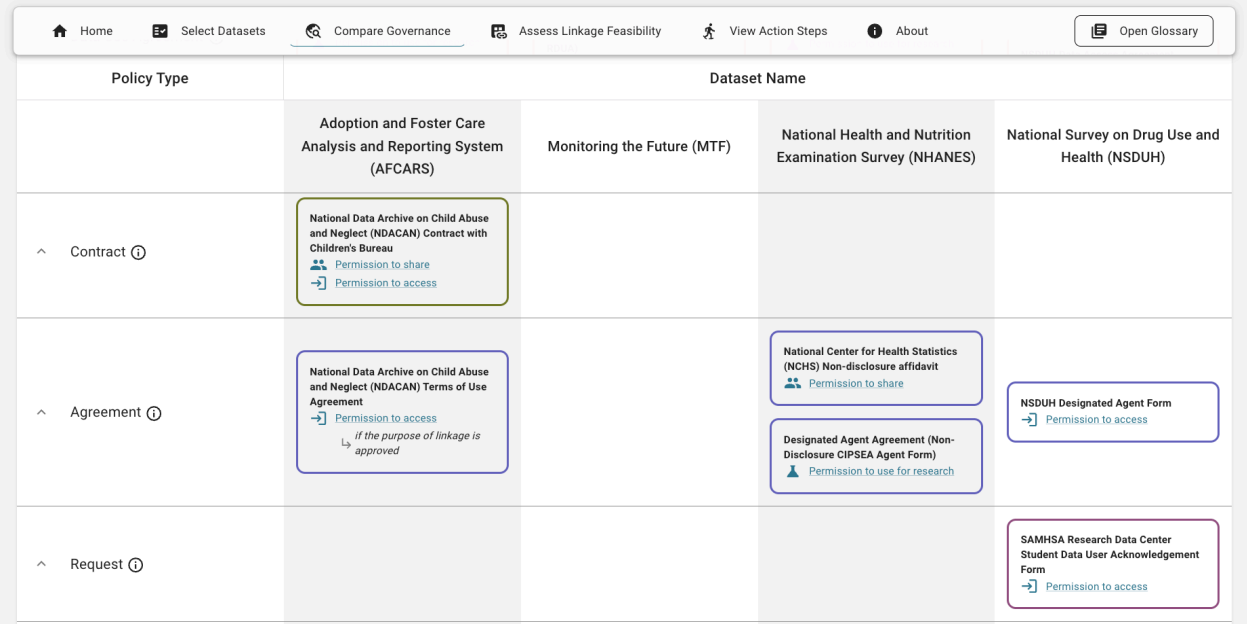


Figure 10: Screen Capture: Compare Governance Page, Expanded View

The expanded view presents the rule or rules from each policy pill; rules are permissions, prohibitions, and obligations to perform specific actions on the dataset (which is a type of asset). Policies may contain a mix of rule types, for example both permissions and prohibitions. For datasets with many policies in multiple policy types, the extended page length led to extensive scrolling that was difficult to navigate. To ease navigation, the user may collapse any policy type section using a ^ function in the row header that collapses the rule information back into pills for that policy type. Clicking on a rule takes the user to the policy detail view. Figure 11 displays the policy detail view, which includes the duties, assigned parties, policy language, interpretation, and source information.

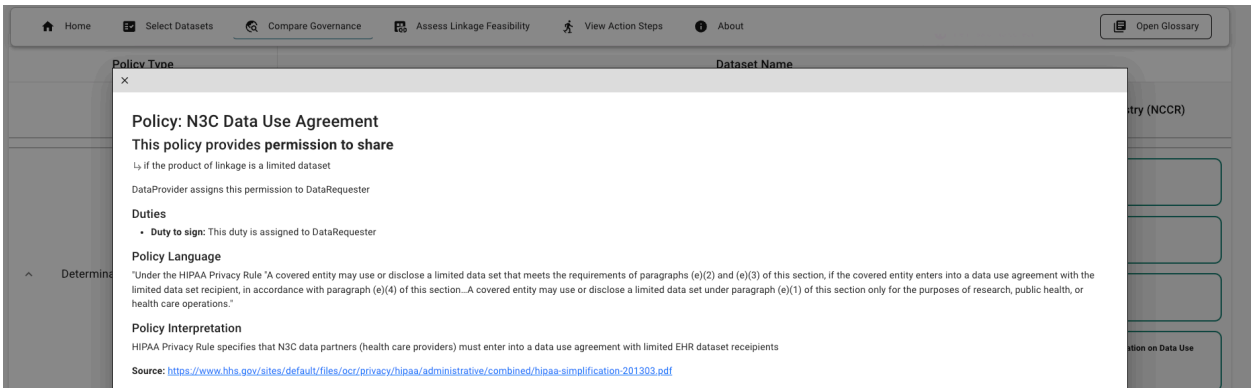


Figure 11: Screen Capture: Compare Governance Page, Policy Detail View

The desired outcome of Compare Governance is that by exploring governance policies and rules for datasets of interest, users are poised to assess linkage feasibility. Figure 12 presents the Assess Linkage Feasibility page that aims to support user decision making about whether selected datasets can be linked for research. Assess Linkage Feasibility presents the same information as Compare Governance but organizes governance information around actions including dataset linkage, sharing, access, and use.

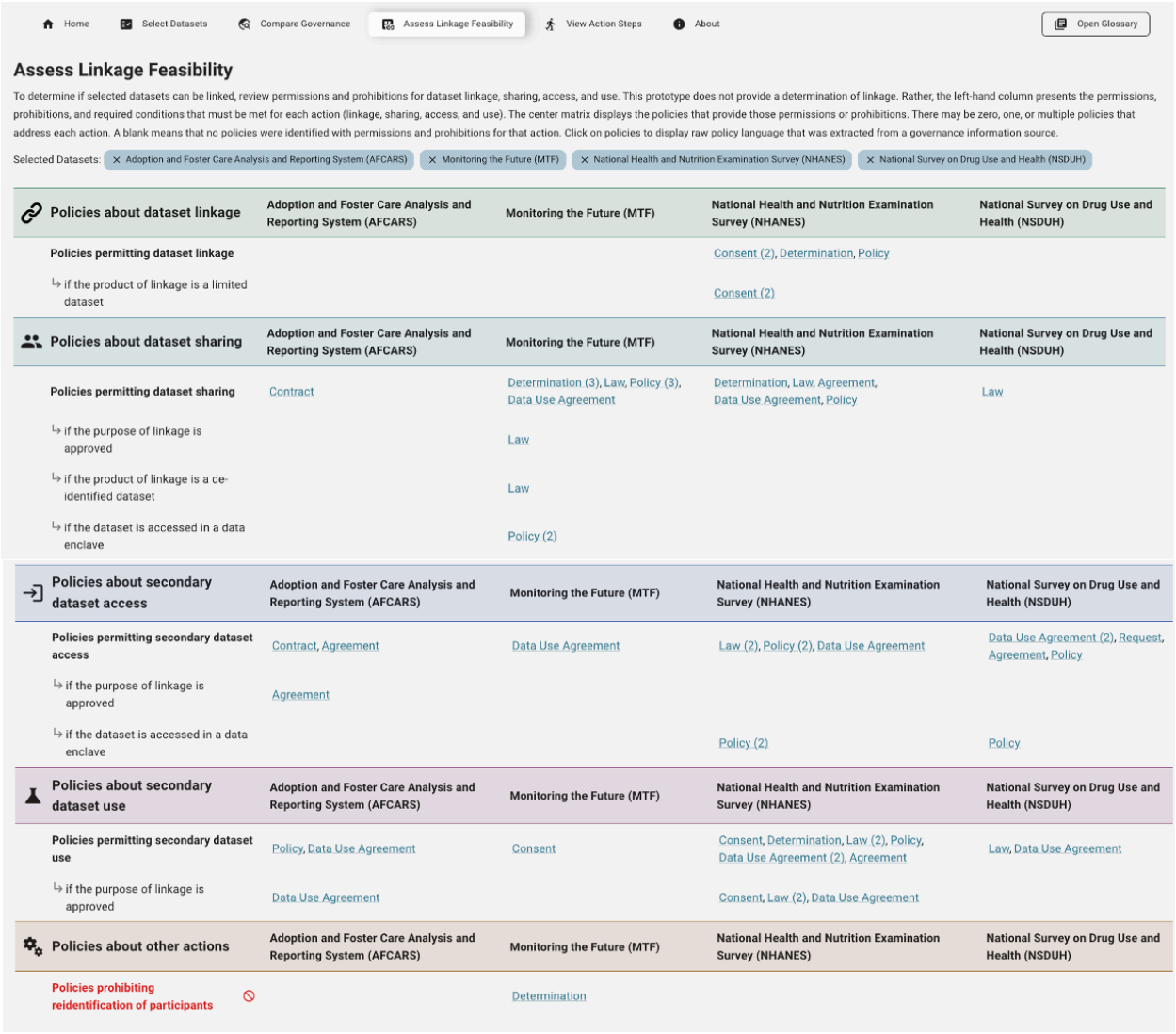


Figure 12: Screen Capture: Assess Linkage Feasibility Page

The left column synthesizes policies across all selected datasets for a single action (e.g., linkage) and lists rules and associated conditions underneath. Using the governance information in Figure 12 as an example, this page conveys that linking these four datasets is permitted only if the linked dataset is a limited dataset (inherited from the NHANES consent form). Note that a clash between rules related to other actions could also interfere with linkage (e.g., requiring access in two different enclaves). Assess linkage feasibility is designed to be read from left to right with the body of the matrix presenting the policy or policies that the represented rules and conditions originate from.

This Assess Linkage Feasibility page does not issue a determination. Linkage feasibility must be determined through a collaborative review of all data governance information and technical linkage considerations by data contributors, researchers, repositories, funders, and policy/legal experts, who can then assess their willingness and capacity to design an approach that meets requirements of linkage, sharing, access, and use. The prototype cannot and does not affirm that linkage of selected datasets is or is not possible. An additional piece of governance information that would likely be helpful to users

assessing the feasibility of linkage is information about prior linkages involving the selected datasets. Although this information was annotated and loaded in the database, the project team did not identify a way to visualize this information in the current prototype because of the level of complexity inherent in this information and its mismatch with the organization of the pages, including the Assess Linkage Feasibility page.

The desired outcome of Assess Linkage Feasibility is that a user understands if linking datasets may or may not be feasible, based on their ability to meet identified rules, conditions and prohibitions. For potentially feasible linkages, the next step is for the user to view the duties (i.e., governance action steps) that various parties are responsible for. Figure 13 displays View Action Steps, which presents action steps organized by assigned party.

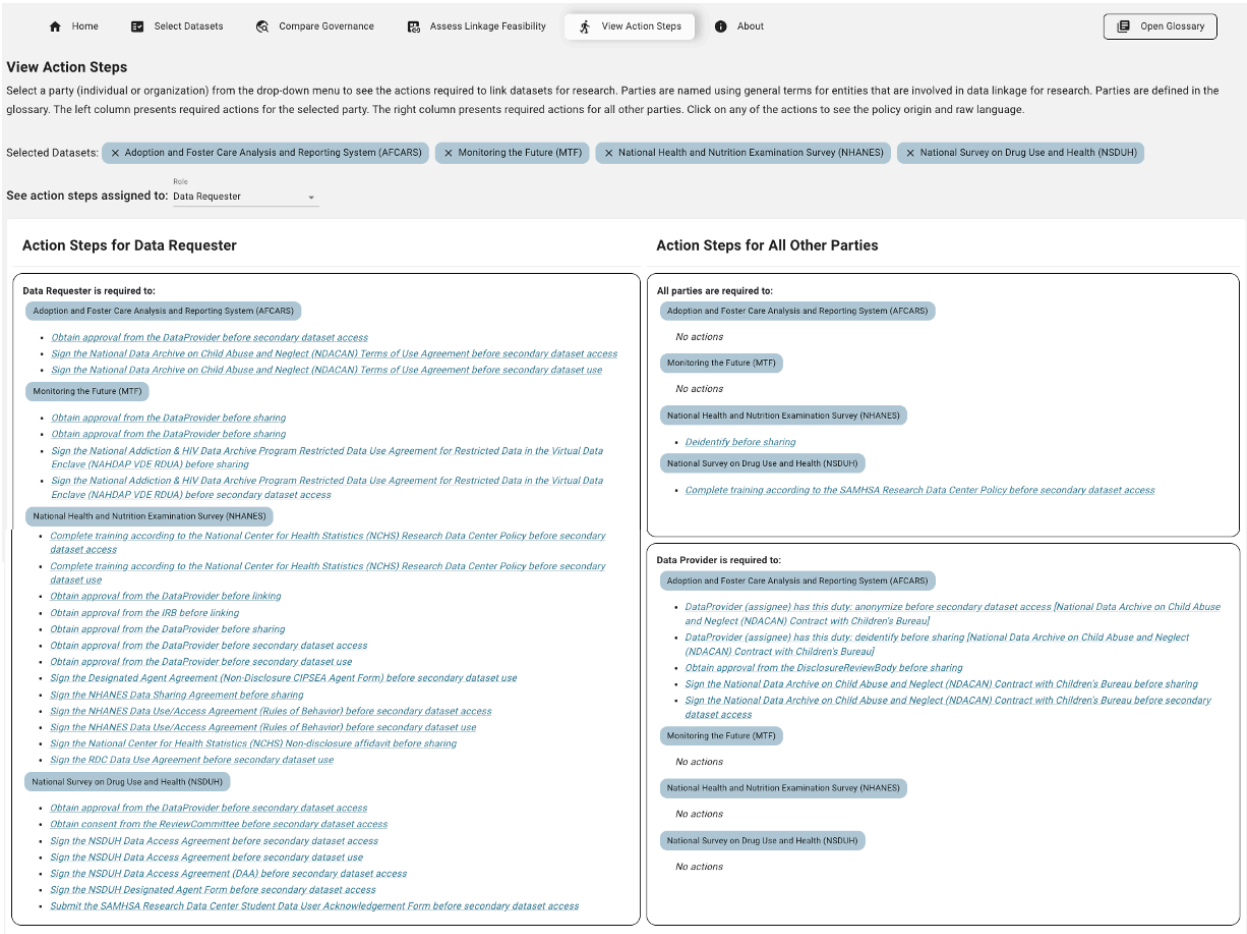


Figure 13: Screen Capture: View Action Steps Page

Action steps for the user-selected party are shown on the left, organized by each dataset. Action steps for all other parties are displayed in the right column, organized by assigned parties and dataset. Duties and obligations with no assigned party are applicable to everyone and are shown at the top under "*All parties are required to.*" While the user can select a variety of parties, the majority of duties are assigned to a data requester.

For View Action Steps, the project team applied the custom script to "humanize" the metadata, converting constraints, duties, and obligations into a human readable text. The schema conceptualizes a constraint as a left operand, operator, and a right operand (e.g., *Output EQ LimitedDataset*) where the left and right operand could be one or multiple concatenated words and the constraint is a condition on a rule. In the 11 datasets, constraints were only identified on permission rules and the degree to which a human could easily read and understand a constraint varies. The conversion generally entailed adding words for clarity. For example, *Output EQ LimitedDataset* was converted to "if the product of linkage is a limited dataset", Purpose EQ ApprovedPurpose was converted to "if the purpose of linkage is approved", and *AccessTier EQ ControlledAccess* was converted to "if the dataset is accessed in a controlled environment". Duties and obligations are rules that are composed of the rule type and an action such as a duty + sign (a duty to sign an agreement) or obligation + deidentify (an obligation to deidentify) where a duty hangs off another rule and an obligation exists at the policy level (not as a requirement from another rule). To humanize duties and obligations, the project team added policy names or parties, and in the case of duties, the rule that the duty originated from. For example, a duty to complete training from a permission to access was humanized as "*Complete training according to the National Center for Health Statistics (NCHS) Research Data Center Policy before secondary dataset access."* The humanization script included a series of "if [constraint] then [humanized text]" that was created for each constraint, duty, and obligation that appeared in the 11 datasets.

Co-design

The project team made updates to the prototype based on co-designer suggestions if the suggestion would improve a user's experience with the prototype and was aligned with the data governance metadata schema goals. The following co-designer suggestions for prototype modifications were implemented:

- Adding a home page that describes the context of this prototype
- Refining instructions at the top of each page for clarity
- Adding a glossary to provide definitions from any page
- Color-coding pills on Compare Governance to align with policy type color chips on Select Datasets
- Making Select Datasets clickable on the home page

Some co-designer suggestions were not implemented. For example, co-designers wanted the View Action Steps page to be deduplicated by policy so that only one action step for signing a given agreement would be presented. This change would simplify the visual and potentially improve usability but required a custom script that would be difficult to automate or scale in future visualization tools. Additionally, this feature may introduce new risks, in that the user cannot further explore underlying policy information for a given action step, which may differ for different sources of the same action. If the action steps were collated by policy, there would be further loss of how the duty is relevant to each stage of the data lifecycle.

3.3 Usability Evaluation

Qualitative analyses by the project team identified 10 themes from the user testers' feedback and observations, including a list of benefits and challenges related to determining the feasibility of linkage

implementations. Three of the themes align with concepts in the UTAUT model, including Effort Expectancy, Performance Expectancy, and Social Influence.[15] The numbers in parentheses in the subsequent section indicate which user tester(s) discussed each given theme.

1. Prototype is easy to navigate, but the core task of understanding governance information within and across datasets is difficult (Effort Expectancy)

The UTAUT model defines effort expectancy as the degree of ease associated with consumers' use of technology. Testers thought that the prototype itself was easy to use and navigate, but that the task of reviewing and reconciling policies was difficult. Five testers reported that the prototype was "easy to use" (1) and "accessible" (4). Testers appreciated the ease of navigating between pages (6), the layout of the pages (1, 2, 3, 4), and that "there wasn't too much on any of the pages" (6). Three testers noted that using the prototype would still be challenging due to the difficulty of absorbing the governance information presented, including the raw policy language. Testers verbalized that it was "just a lot of information" (1), was "complicated" (1), and would take a lot of time to "digest" (3, 5). Reconciling many different examples of governance information is a complex task that requires background knowledge and experience with the tool to perform effectively.

Across the six usability evaluation sessions, the time required to navigate through the prototype ranged from 45 minutes to 1 hour and 3 minutes. During the sessions, however, multiple testers noted that in a real-world setting they would require "a little bit more time to review fully" (3), and that successfully using the prototype was "a matter of spending enough time to take in the information" (5). Two testers (3, 5) thought that users would need some experience with the prototype before they could use it effectively.

2. Testers exhibited substantial variation in their interpretation of the governance information visualizations (Performance Expectancy)

The UTAUT model defines performance expectancy as the degree to which using a technology will provide benefits to consumers in performing certain activities. All testers were able to use the prototype to answer questions and make their own determination as to whether linking the three datasets in the scenario might be feasible. Testers' consistency in answering questions and assessing linkage feasibility, however, varied substantially across sessions. The accuracy rate across responses to testing questions varied from 20% to 100%, and appeared to be related to differences in how testers interpreted the questions and presented governance information. Testers' ratings of the feasibility of linking the three datasets (on a scale of 1 to 10, with 1 being trivially easy and 10 being not possible) varied from 3 to 9 with an average score of 7.2 and depended on whether the tester focused on the possibility or the practicality of linkage. Testers acknowledged the potentially subjective nature of interpreting a collection of policies: "It's not providing the determination of linkage, so that's really up to the user then to determine if the linkage is feasible." One tester (1) suggested that the initial description associated with a dataset (in the Select Dataset page) might provide a clearer picture of that dataset's governance and next steps for implementing a linkage:

> *Probably all of these places have really good resources and help desks and places where you can ask questions, so that might even be a helpful thing on next steps as well ... we have a technical*

assistance contractor who assists researchers, and there's a lot of information on the request process as well as all of the forms and documentation that need to be taken into account. (1)

3. User testers' institutions would generally support use of a tool like this prototype but may be cautious due to the potential negative consequences of misinterpretation of governance information (Social Influence)

The UTAUT defines social influence as the extent to which consumers perceive that important others believe they should use a particular technology. Three user testers thought that research institutions would generally support the use of a prototype like this (1, 3, 5). One tester noted:

I think my research institution would like to see this as a tool that people can readily use. I mean, we all know how hard it is to deal with this kind of issue. You know, trying to find things through Google is not a very easy process and is error prone. So yeah, I think the tool has a huge amount of value. (3)

Some testers did note potential hesitation about using the prototype due to legal consequences for incorrectly interpreting governance information. One tester noted that researchers may be cautious about using features in the prototype as they are "aware that if things go south, I'm going to have to hold the bag for some of it" (4). Two testers (2, 6) felt that research institutions may be wary of the accuracy of the information presented in the visualization prototype, stating:

The only thing that might make me or make my institution not support it is if they don't trust the interpretations that are extracted across the board. So, I could see other research institutions and their lawyers saying "no, we're never going to use a tool like this because we need to review everything ourselves." (6)

4. The value of organizing policies by type was unclear

The Compare Governance Information page organizes applicable policies by type, using the categories "Consent," "Determination," "Data Use Agreement," "Contract," "Certification," "Agreement," "Request," "Policy," and "Law." User testers did not find this organization to be helpful, unanimously finding this page to be the least helpful with decision making. Four testers (1, 2, 3, 6) thought that policy type definitions were not consistent across the field and consulted with the glossary to review the definitions used in the prototype. As stated by one tester: "I'm just curious as someone who thinks a lot about consent—there's a lot of definitions of consent, which makes sense because it can be a complex concept" (6).

Two testers (2, 5) felt that presenting policy names and types could help to set user expectations, such as recognizing that data from administrative data sources and data collected as part of routine care delivery may not have a "Consent" type policy (5). One tester also used the web browser "find" functionality (Ctrl+F) to search for specific keywords like "linkage" in the policy titles (4). Four testers, however, noted that a policy's name or type does not necessarily inform the user of what the policy contains. One tester noted that this is due to a lack of standardization in the field about what types of policies contain which types of rules:

In real life there is sometimes a Data Use Agreement, and it will talk about restrictions on use that you thought would have gone somewhere else, like in the IRB statement or research submission ... sometimes all the information's not put in the right bins, as it should be. (3)

This tester additionally hoped that the introduction of a governance metadata schema would help to address this issue if it was used widely (3). Another tester felt that the policy type categorizations could be stripped out because "it's the content of the cells that really matters" (4). Regardless, the testers felt the need to review all the listed policies in order to understand governance information about specific parts of the dataset lifecycle such as secondary access and secondary use (1, 3, 4). Testers noted potential negative consequences of relying on policy type categorizations to review governance information. Testers feared that users may "selectively jump to" certain policies without exploring all the content (3) or only focus on types of governance information they are familiar with such as Consents or Data Use Agreements (4).

5. The use and interpretation of the prototype may vary depending on whether the user is research-oriented or policy-oriented

User testers anticipated that research-oriented and policy-oriented users would have different goals and interpret data in different ways. All testers suggested that researchers would be very practical, goal oriented, and focused on "what they would need to do to actually just conduct their research" (1). Testers thought that researchers would be most interested in "an executive summary" (1) and would look for ways to avoid reading all the policy information. Two testers (1, 4) felt that a research-oriented user would want to look at governance specifically about linkage first to determine whether a review of subsequent governance information was necessary. As noted by one tester:

So, as the researcher, this is the thing I'm going to be looking for probably first. Because if I can identify that there's a prohibition on linkage, then I can just stop right there, and I don't need to read anything else. Whatever all has to be done doesn't really matter to me until I know that linkage is permitted. Then I can get into "oh, but what kind and how much," and are there other requirements, and things like that. (4)

In contrast, testers who took a more policy-oriented approach to reviewing the information wanted to read all policies before coming to conclusions (1, 3, 4, 5). These testers would want to "explore every single piece" (5) and "read them at length to confirm my understanding from what I've seen displayed here" (4). Policy-oriented testers were more likely to use external links to review original policy documents, which they felt "you always need to investigate" (5). One tester was confused when they didn't see an external link for a policy they were exploring as they weren't sure whether they were "missing something" (3).

Five testers (2, 3, 4, 5, 6) indicated that a research-oriented viewpoint was more interested in the practical aspects of linkage than pure feasibility. Testers thought that factors such as grant timelines, gaining the attention of officials at their institution, and engaging with external research partners could limit a researcher's ability to implement a linkage. As noted by one tester assessing the feasibility of linkage in the test scenario:

I think it's possible. Do I wanna go through the process of doing it? I don't know. Because it seems like a heavy lift, because of just how many lines are here. I think if I really wanna do it, if the research is really compelling, I can do it. It's allowed. (6)

Research-oriented users were most interested in the View Action Steps page when considering the potential for linkage feasibility (3, 4, 5, 6), feeling that it has "discrete steps" (4) that "direct me on what to do" (3). They felt that Compare Governance and Assess Linkage Feasibility pages were not as helpful to support their decision making (6) and did not tell the user what could potentially block a linkage implementation (2). Testers also thought that researchers may lack a background in governance concepts and terminology (4, 5, 6) to effectively use the prototype, and may lack the time to read through all of the policies listed (6). Additionally, user testers did not appear to recognize that the View Action Steps page is about steps various parties need to take across the lifecycle of a given dataset. Rather they appeared to think incorrectly that it was about steps to take to implement a new record linkage implementation.

6. Trust in system information and features varied substantially between testers

User testers often discussed the trustworthiness of the prototype, questioning whether the governance information was an accurate reflection of all relevant policies. Even though the orientation for the user testing sessions reminded testers that the governance information derived from the previous OS-PCORTF Pediatric Record Linkage Governance Assessment project, five testers (1, 3, 4, 5, 6) questioned whether the information in the prototype was complete, thinking that the user "might need to do some more digging" (1) or "try to Google search the governance policies and read through them independently" (3). Three testers (1, 2, 3) thought that the owners of the datasets would provide the most accurate governance information. As noted by one tester:

> *There's a trust issue. And the trust issue comes down to ... do I trust the tool has all the data in it that I need? And I think that that comes down to a little bit as this rolls out. How well are you able to get places to provide the data that you need? (3)*

Testers exhibited different understandings of features meant to aid policy interpretation such as icons, coloring, or categories. Testers appreciated prototype features that aid interpretation, such as the "Policy Interpretation" section of the policy text pop-up boxes (2, 6) and the red text associated with prohibitions (4, 5, 6). Some testers recommended that the prototype have additional features to "boil down" (1) the presented information and present a "hard and fast 'no' response" (6) to questions about linkage feasibility. There was, however, inconsistency with how the existing features were interpreted. On the Assess Linkage Feasibility page, the terms "permission," "prohibition," and "condition" are not clearly defined, and testers interpreted them in different ways. Two testers (2, 6) initially thought that if a policy did not have any prohibitions listed, as indicated by its red text and the icon, that there were no "showstoppers" (2) that would bar linkage. After further exploration, though, they found that most prohibitions were tied to specific conditions and then felt that the use of red was "a little contradictory" (2) since linkage could still be permitted. As stated by one tester:

> *I talked a lot about the red circles with the arrows [prohibition icon]. It seemed like they should be some of the most helpful visual cues for me. But then when I personally dug into them, I didn't think that they were really accurate reflections of a full-on prohibition, it felt like that was a misalignment between the visual cue and the text embedded. (6)*

Another tester skipped over header lines indicating permitted actions "because the end summary take away is that these are things that are in the governance structure that allow me to link the data" and

they therefore have no conditions to fulfill or actions to take (6). One tester was unclear as to what the header language regarding permissions and prohibitions meant on the Assess Linkage Feasibility page:

> *I'm seeing already there's permission to link in the consent. I would have to look at that to figure out what that actually means, because as a first-time user, I don't know right away whether if something says "permission to link," it means "there is information about the permission to link in here" or whether it means there is explicitly permission to link. (4)*

When faced with multiple conditions being documented in one policy, this tester also wondered whether all the conditions are "alternatives," indicating that only one condition must be met, or whether they are "additive," meaning that all conditions must be met (4).

In the Compare Governance and Assess Linkage Feasibility pages, blank spaces in the table also confused testers. Two testers (1, 4) interpreted a blank space as an "unknown" where "you still have to do digging to determine their requirements" (1). Another tester interpreted blanks as a blanket permission since no prohibitions were listed (2). Seeing blank spaces could also cause confusion, leaving testers wondering "why are there no determinations" for datasets with blank spaces (4). One tester interpreted the existence of actions on the Action Steps page as a sign that linkage was ultimately allowed for a dataset (6), misunderstanding that this view is not steps for linkage but rather duties for each lifecycle phase. Testers were not aware of the Open World Assumption that missing information is neither a prohibition nor a permission.

7. Testers thought that some information was irrelevant for decision making, and requested new information to be displayed or made more obvious

User testers indicated that some information provided in the prototype seemed irrelevant to their decision-making process and could make dataset linkage implementations seem more complicated than it should. Three testers (1, 4, 6) thought that information related to primary data collection and sharing was irrelevant since those were actions that would be taken by another party. As noted by one tester:

> *I perceive a high possibility of the fact that this system is pulling in information that is relevant for primary researchers who submitted the data, that this is going to trip up secondary researchers and will possibly pollute a little bit like the instructions provided to them. So, I just think one way to address that could be not including information like that ... like a sharing expectation from the [policy] on primary researchers has essentially no bearing for secondary researchers. (4)*

One tester questioned whether actions such as deidentifying a dataset would be relevant to a secondary researcher because "the data should just be deidentified already before I get to it" (6).

Three testers (1, 2, 4) thought that the prototype should emphasize situations where policies across datasets conflicted as these would require more effort to reconcile. One tester thought that "we just have to know if the prohibition is across the board, and if it's across the board, it actually means that all three of them have the same policy" (2). A possible conflict commonly identified by testers was whether or not PPRL is required for linkage since a requirement to use PPRL for one dataset means that "presumably you just want to use that everywhere" (4).

Two testers (2, 4) wanted to see more dataset information in the Select Datasets screen to help with data discovery and convey potential restrictions on usage. One tester wanted information indicating whether participants in the listed datasets were likely to overlap:

> *If I went to something like [the tool], what I really would like to see is do these things actually have overlapping subjects ... and then I can come here for the governance information, and it would be a whole lot nicer. If the tool was both, obviously that's a big thing because determining if two studies have overlapping participants is itself a huge challenge. (3)*

While testers appreciated the organization and clear sequencing of the prototype's pages (1, 2, 3, 4, 6), testers also faced challenges with some of the features and may have missed information relevant to understanding full intention of governance information presented. Three testers (1, 2, 5) used the zoom feature in their browsers to increase font size and subsequently were unable to see the "About" and "Open Glossary" buttons on the top of the screen. Testers also suggested more guidance on which visual features on each page are interactive and clickable.

Testers were confused by the lack of interactivity of the "pill" visual icons in the Select Datasets (1, 2, 5, 6) and Compare Governance (5, 6) screens. Two testers also did not immediately understand that the policy language pop-ups in the Assess Linkage Feasibility page could contain information on multiple policies (2, 5), suggesting that the tabs that toggle between policies be more prominently visible. While testers thought that the Compare Governance page was the least helpful with decision making, testers appreciated the visual structure of the page (1, 2, 3) and suggested that it be incorporated into the other pages.

8. Testers wanted more clarity on the roles defined in the prototype, and how those roles relate to policy and rules

User testers were confused at some of the terminology used to define roles and actions in the prototype. One tester thought that the Action Steps page was the only one that adequately communicated roles and responsibilities: "there's part of it earlier where there weren't nouns or names being used, and so I didn't know which parties we were talking about all the time" (4). Throughout the prototype, four testers (1, 3, 4, 6) questioned what the terms "primary" and "secondary" meant in the context of the roles. Some testers incorrectly thought that "secondary use" referred to the user redistributing their linked dataset to other researchers. When navigating the Action Steps page, four testers (1, 2, 3, 4) thought that users of the prototype might naturally identify themselves as the "Primary Investigator" instead of the "Data Requester." One tester was confused by the term "Data Provider" and thought that the definition in the glossary is "kind of a catch-all term for everyone who may be involved in funding and doing the data collection and submission" (6). In the Action Steps page, four testers (1, 2, 3, 4) were also confused by roles such as "IRB" and "Data Coordinating Center" and respective duties for a given dataset across its lifecycle since multiple parties may be involved in a linkage implementation that involves multiple institutions:

> *I'm guessing that this means the data coordinating center that currently holds the data that I'm trying to get access to. Or it potentially means the one that I'm part of, so I would be investigating this further and looking down at this to figure out whether this is the thing that I want. (4)*

9. The volume of presented information may discourage researchers from pursuing linkage implementations

All testers suggested that the volume of information presented in the prototype, especially in the Action Steps page, could discourage researchers from pursuing a linkage implementation. As stated by one tester: ″You know, this is supposed to be something that's gonna help the investigator be like 'I can do this', but that's not the case, at least looking at those action steps" (2). Testers noted that the prototype contains "just a lot of information" (1) and that the scenario in the testing session listed "a crazy amount" (6) of action steps. One tester felt that listing all the associated policies for multiple datasets may "lead to probably limited use of some of the datasets" that have long lists of associated policies and action steps (5).

Splitting Action Steps by the dataset lifecycle inflated the number of steps listed and testers thought that repetition of information in the prototype could make the task of linkage seem more daunting than it actually is. Four testers (1, 2, 3, 6) thought that repeating actions on the Action Steps page for each step of the dataset lifecycle made the task of linking datasets seem daunting:

> *I think for the view action steps, I didn't like how it broke things down so discretely at that point. I think it was maybe overly broken out and that major steps should really be reflected on that page rather than each individual section of an agreement, for instance, being called out separately. (6)*

One tester also questioned whether multiple steps assigned to different parties could possibly represent one action step to accomplish:

> *As a researcher, I'm looking at this and saying, really? I have to obtain approval from the data provider before secondary data set access, AND I have to obtain approval from the Review Committee before secondary data set access? Is it possible those could be the same thing? (4)*

Testers also noted cases where policies that are listed separately in the metadata database are grouped together in practice, meaning that one action is represented as multiple action steps

> *What do I need to sign in order to link the [dataset] to something? I need to do the Data Use Certification agreement, which is one agreement, the [dataset] User Code of Conduct, which in my role I know that's kind of embedded in other agreements. Typically, it's referenced in other things, so I don't know that that's a separate signature—it might be a check box or something. (6)*

Testers also thought that the terms used to describe different actions were sometimes inaccurate and made tasks seem more difficult to complete. One tester (2) identified a case where a “sign” action in the prototype was only a clickthrough on a website, indicating that obtaining a signature implied more required time and effort. Another tester (3) wondered whether “submit” actions would also require a signature.

10. Directions for Future Development

Testers suggested future directions for development of tools like this prototype. Multiple testers suggested that this prototype would be useful for policy analysis outside of a research context. One tester thought that research consortia would find this prototype useful: "I can think of where consortia

would find a prototype like this very helpful when they're bringing Investigator A through Z together to bring all their data together, that they gathered from their different studies" (2).

One policy-oriented tester remarked:

> *I was wearing both hats ... to approach it as a researcher, but also as a policy person. I actually am a user of this stuff all the time. So, when things go wrong or go weird, we are often looking up "well, what did they say in the institutional certification? What was in the consent, or in the submission agreement?" There are users who are not just researchers that will probably use these things for sometimes rather high stakes investigations as well. (4)*

One tester suggested that users would want an exportable version of the Action Steps page to act as a list of rules that need to be respected in the design of a new linkage strategy:

> *I will probably break them out and be doing color coding and stuff and making sure that all this stuff is getting done so I can really push on my institutional officials to sign these dang things before my grant runs up. (4)*

One tester suggested that tools like this "may also lead to some requests ... for improving the process or changing the process in a way that is more easily accessible by researchers" (5). Three testers (3, 5, 6) discussed the importance of scaling this prototype to add datasets and encourage widespread usage. Two testers (3, 6) suggested that it be hosted on the National Center for Biotechnology Information or another NIH webpage for widespread use.

4 Discussion

The discussion section addresses the two aims of this proof-of-concept implementation project. The first aim is to test the data governance metadata schema by examining how the schema structure and design can support researchers to make decisions about a proposed linkage implementation. The second is to demonstrate how governance information collected from real-world datasets can be visualized to inform users how to appropriately link and handle datasets of interest across the lifecycle.

The discussion presents themes guided by six interrelated questions:

- Does the approach to annotating governance metadata follow the schema?
- Can governance metadata be transformed into human readable governance information?
- Can governance information (from metadata) be consumed and interpreted accurately by a user?
- Can a user make an assessment about the feasibility of dataset linkage for research?
- Does a user understand the conditions that must be met to link datasets for research?
- Does a user understand what actions are required to link datasets for research?

4.1 Themes

Visualization of governance information in and across the data lifecycle is complementary to the Record Linkage Implementation Checklist: Perhaps the most powerful outcome of this proof-of-concept implementation project is the novel approach to organizing and visualizing governance information in

and across the data lifecycle, which is intended as a companion resource to the NICHD ODSS developed Record Linkage Implementation Checklist. Users had an incorrect understanding of the View Action steps page, misconstruing that these were steps for "linking these datasets" when they were, by intention, duties that various parties must do *across the lifecycle* of *each individual* dataset. These duties are rules that would need to be considered when designing a linkage implementation, but the steps for linkage are much more complex and require going back to the Record Linkage Implementation Checklist and collaborating with multiple parties. Many of the findings in the usability evaluation were driven by user testers having wrong impressions of this page and the context for how they would determine if a linkage implementation is feasible or practical. User testers also seemed to not understand that linkage is not just about data access; for example, it requires going back to the data originator who has PII that can be used to link participants. While linkage is a discrete phase of the data lifecycle, typically someone at each of the other lifecycle phases plays a role in making linkage possible.

Manual annotation approach may not scale but is necessary for schema refinement: The approach to annotating governance information from the spreadsheet to relational database with governance metadata aligned to the schema was time consuming and requires new approaches for scalability. This project translated governance information to metadata and back-translated metadata into information. First, the project team annotated the Excel file that was structured in the Governance Information Framework and loaded it into a governance metadata relational database. The metadata was then reconstituted as governance information for presentation in the visualization prototype. The visualization prototype should help the user to consume and interpret the information, make an assessment about the feasibility of a linkage implementation, understand what conditions must be met, and understand which actions they and other parties need to take. Visualizing governance information from annotations in the prototype was a powerful data quality check, to identify how annotations could be improved. Manual annotation across these tools provided a learning experience for the team and informed improvements to the schema.

- **Converting governance metadata into human readable governance information**: The project team created governance metadata for the 11 datasets by manually annotating information from an Excel spreadsheet and then loading it into a relational governance metadata database for presentation, in a human readable format.

- **Gaps in governance terms**: The project team discovered that some concepts in the source information could not be represented in the schema vocabulary (e.g., there was not a term for PPRL or limited dataset). For gaps identified before the schema was finalized, over 70 terms were added to the schema vocabulary. Terms that were not discovered in time for inclusion in the schema are noted in recommendations for future development.

- **Schema limitations**: The team also discovered concepts in governance scenarios that could not be represented with schema classes and were difficult to annotate (e.g., three parties on an agreement, 15-20 parties in a network agreement, rules that come from groups of similar laws such as state cancer registries laws).

- **Variation in annotations:** The project team realized that the schema's current specification allows multiple ways to annotate some governance concepts. For example, one data source requires users to complete an IT training, attest to a data user code of conduct, and complete a

human subjects' research protection training in order to access data. This could be annotated as three rules within one data access policy or three policies, each with one rule. Manual annotation varied between two members of the project team, which suggests the schema's specificity could improve. If a long-term goal is for automated tools to translate governance information into metadata and back-translate metadata into information, the schema's specificity or implementation materials require refinement.

- **Reaching completeness in policy annotation was challenging:** The project team conducted four rounds of annotation, an initial pass and three subsequent refinements. Each time annotations were revisited; policies and their rules were expanded with more details. For example, duties to sign every data use agreement and contract were added. Policy associations with other policies were added through the schema "relies on" function and assigning and assigned parties were added. Even after four rounds of annotation, the governance metadata could still be enriched with finer detail. The project team even found instances where the source governance information, which was on the whole very detailed, had gaps on some key policy details. This observation suggests that reaching completeness in policy annotation is difficult and raises a consideration about what level of detail in annotation is reasonable or realistic.

- **Annotation of parties require further consideration:** Annotation of parties was particularly challenging. It was difficult to determine when to annotate one or both parties (e.g., assigner and assignee). When selecting from the parties value set, an organization often fulfilled multiple roles such that many values would be accurate; for example, when the government organization was also the certifying organization, or the government organization was also the data repository. Modifications to how parties are conceptualized and function in the schema may need to be considered.

Visualization tools cannot render definitive decisions about feasibility of linkage implementation, as these decisions require a human determination: The visualization prototype was never meant to result in a yes/no decision about the feasibility of pursuing a linkage implementation, even though users wanted the prototype to render a definitive decision about linkage feasibility. User testers were frustrated to find that the prototype supported their decision making rather than making a decision for them. On the other hand, some users were unconvinced that the Assess Linkage Feasibility page provided enough information or the right information to make a decision and related the sentiment that linkage feasibility in their experience is often not their decision, but the decision of dataset owners. Navigating data governance is critical for protecting research participant privacy, addressing ethics, managing risks, facilitating compliance, and respecting participant trust but it can be challenging. Linkage implementers should be prepared to put in the multidisciplinary effort necessary to identify and understand data governance and develop a strategy for how to respect/adhere to the associated rules. A more thorough policy analysis, engaging legal and/or policy experts, will be required to verify governance information, and collaboration will be required to make decisions. It is not possible for one group to make decisions on their own when it comes to linkage.

Capturing and visualizing policy inter-relationships using the schema would require a level of legal verification that is likely impractical: While users suggested visualizing representation of inter-relationships between policies, policy inter-relatedness would require a level of legal verification that is

likely impractical. User testers articulated how many policies are related and flow into each other and voiced that those relationships are not well visualized in the tool or documented by the schema. Indeed, many governance policies within the 11 datasets were interrelated. For example, the consent reflects the guidelines of the IRB and the requirements of the Common Rule. Processing governance based on which action rules and policies applied to, rather than where they originated from added complexity for the user. The fact that multiple policies can offer the same rule but different conditions for the same dataset added further complexity. Users, particularly researchers, tend to mentally organize governance based on origin (i.e., rules from consent and rules from the IRB) rather than what action governance rules apply to. It is possible that visualizing the relationships between policies could aid user interpretation. However, it is often impossible to determine whether a rule came from another policy or whether a given policy adopted that rule on its own. Additionally, seeing policy relationships could add confusion. This topic requires more exploration.

Data governance literacy will impact future development of visualization tools like this prototype: Common data governance terms and their relation to the data lifecycle may not be widely understood in the same manner. User testers, co-designers, and TEP members grappled with the meaning of many common governance terms like dataset, linkage, sharing, secondary access, and secondary use. Researchers affirmed that these terms are used widely across the research community, and yet linkage implementation means different things to different groups in the context of the constituencies they collaborate with. Despite the definitions and instructions provided in Home Page, the Glossary, and the About Page, unclear terms made exploration of the data visualization prototype take longer and shook some users' confidence in their own understanding of governance and their assessment of if linkage implementation would be feasible.

Visualization of data governance as demonstrated in the prototype could be used to inform the development of solutions to other data governance challenges, beyond linkage: The schema can be used to define a minimum or ideal set of governance metadata that travels with data across its lifecycle to help adhere to a dataset's rules in many settings such as determining whether certain AI/ML applications or other emerging technologies are appropriate for the dataset. Additionally, the schema could be adopted to develop solutions for integrating data across multiple repositories in a federated data ecosystem like NIH's, extending the application of the schema from participant-level linkage into system-level interoperability. These efforts are policy-relevant for NIH and have potential to be applied to any federation project where there is intention to bring together datasets for a new activity. This application aligns with the forthcoming 2025-2030 NIH Strategic Plan for Data Science Goal 4: Support for a Federated Biomedical Research Data Infrastructure.

4.2 Limitations

The database and prototype tested the schema using structured governance information from only 11 datasets. The 11 datasets used here were diverse and represent a wide diversity of types of data, governance policies, technical architectures, and applicable laws. However, governance information is varied across the vast continuum of existing health-related datasets with many bespoke governance concepts within the research and health subspecialties. The schema and visualization tools like this one will benefit from adoption and testing with additional data types and domains.

The project team engaged only six co-designers, which represents only a small sample of biomedical research perspectives and only six user testers. Furthermore, user testers were selected based on experience with NICHD ODSS, data sharing, and linkage implementation, with varied background and experience working in research and policy analysis roles. The usability evaluation did not capture the experience of a linkage-naïve researcher using this prototype. It also did not capture the perspectives of institutional representatives, data repository stewards, legal experts, and other parties who also play an important role in making data linkage possible. Engaging these groups as co-designers and user testers could have yielded different results.

5 Recommendations

Recommendations for future data governance visualization efforts and the data governance metadata schema are based on the project teams' outcomes and findings from database and prototype development as well as the prototype usability evaluation.

This project produced a prototype visualization tool not intended for production use. While the tool was successfully used by a handful of user testers, this experiment surfaced areas where visualization of governance metadata can be refined. Developing and testing the governance metadata visualization prototype highlighted ways that the data governance metadata schema could be improved to support data linkage determinations and the design of linkage implementations that ensure appropriate sharing, access, use or handling of linked datasets.

1. **Based on the annotation exercise and database creation, the data governance metadata schema should consider the following:**

    - **Providing user guidance regarding policy annotation**: Important clarification is needed about when to enter a new policy versus when to add to an existing policy. During annotation, project team members were often uncertain whether to enter a new policy or add a rule to an existing policy. While rules were usually clear, the policy container that rules originated from was often vague. A strategy for handling clear rules with an unclear policy origin will be required to guide production metadata visualization applications.
    - **Selecting an approach to creating duties and obligations:** The View Action Steps visual highlighted the importance of duties in understanding next steps for governance. Duties on policies or rules are often implied (e.g., a duty to sign a DUA) and were not well specified in the source governance information. The project team annotated many duties based on assumptions (e.g., if there is a DUA required on the data requester, then the data requester has a duty to sign it). A set of assumptions could be developed for encoding policy types, actions, and constraints in the schema so that duties could be generated without having to be manually entered. If the View Action Steps page continues to be important in a tool like this prototype, the schema could incorporate a standard set of rules that are created anytime someone creates a rule to take a training (e.g., adding a duty created on top of it, specifying that the data requester needs to take this training).
    - **Expanding the policy type value set**: Policy types are the prototype's primary strategy for categorizing policies. For some datasets, users were challenged to navigate the "policy" type category that held upwards of 20 policies. The schema could expand the policy type value set to

further disaggregate policy into more discrete types. However, any value set expansion should first affirm that a "type" value set is the best way to sort policies and that such a value set can be maintained over time. Note, this recommendation may also apply to future visualization tools and how tool developers choose to organize the governance information, whether by policy type as in the current tool or another sorting method.

- **Adding raw policy documentation and functional links to all source information in policy annotations:** Some users were keen to navigate the raw policy language and review excerpts from policy documentation to confirm their understanding of policies and rules (e.g., the language in the Data Use Agreement or consent form) and to address any concerns about the comprehensiveness of the information presented in the prototype. However, policy excerpts imported from the source were not thoroughly hyperlinked with included references. The schema captures source, which is sometimes populated by the link where the policy documentation is available, but the schema has no class for raw policy language. A schema class could be created to capture policy excerpts from policy documentation, including content that is not publicly available, and support easy navigation to the relevant policy section for long policy documents. Future visualizations could prioritize linking and citing relevant resources within the raw policy language.
- **Adding a class to hold a human readable version of a constraint:** Depending on a user's knowledge and experience, some of the constraints are readable, but many are not. The project team wrote a custom script to convert constraints into human readable statements in the prototype. The schema could add a free text field to store a human readable version of constraints – diverging from the schema's intent to hold structured metadata. Alternatively, the schema could alter how it captures constraints to be more readable.

2. **Standards making bodies should maintain the governance vocabularies used by the schema**: Many of the recommendations in this report would require adding terms to the schema vocabulary to represent additional governance concepts. For example, a term for PPRL was used in the database and should be added to the schema vocabulary; but the annotation of 11 datasets suggests that the schema will always be expanding its terminology and vocabulary to represent additional governance concepts. As new linkage technologies emerge and more linkage implementations occur, novel governance concepts will be created that require a corresponding governance terminology. Terminologies will continue to exist outside of the schema, and standards making bodies should maintain these terminologies through SME engagement. The schema can be updated as needed to map to relevant terminology concepts.

3. **There is a broad need for data governance literacy education:** Researcher teams would benefit from educational resources on data governance for linkage and how it relates to all phases of the dataset lifecycle. For example, data linkage likely involves working with original data providers when the available shared datasets are deidentified. Researchers generally receive minimal education on the concepts of data governance across the data lifecycle including linking datasets for research. Some user testers struggled to use the visualization tool in part because they have limited experience and training on the underlying task. Moving forward, metadata visualization will be more successful as education and training on the topics of data sharing and linkage are integrated across the research community. The schema and tools like this prototype, along with their documentation,

when combined with the NICHD Record Implementation Checklist, could serve as education materials to improve literacy. Future tools should consider adding or linking to relevant educational modules to inform governance novices and research teams on how to most effectively use the visualization tool. Consensus language for the governance of linkage, access, sharing and use across these and other governance visualization tools should align to NIH and HHS recommended terms and include a comprehensive glossary.

4. **Researchers should engage policy/legal experts to interpret governance information for a new linkage implementation**. The schema and prototype documentation we created (including user guides) could serve as frameworks to help guide what information needs to be collected (data collection prototype) and how to organize the information for interpretation (visualization prototype) to guide more thorough policy/governance analyses and facilitate conversations with legal/policy experts. Implementing new linkage requires a bigger picture strategy guided by the governance and technical considerations enumerated in the Record Implementation Linkage Checklist, combined with detailed input of policy/legal experts.

5. **Future visualization tools should build on this prototype (the schema, database, code, and user interface design) and consider the following**:

    - **Expanding dataset information to support dataset selection:** Users wanted more information as part of their dataset selection process including uniform data types, indication of datasets that cannot be linked, and history of linkage as evidence that dataset linkage is possible. A number of ongoing efforts are working to define an ideal set of metadata for describing datasets to enhance search and reuse[16], and those standards could be incorporated for future tools The schema in this project attempted to capture information regarding history of linkage, but displaying that information was too complex for this prototype.
    - **Adopting a uniform value set for data type:** Users wanted more searchable dataset descriptors (i.e. types) when selecting datasets. The prototype is intentionally not a dataset search tool, and the schema could be implemented to extend an existing search tool that offers extensive dataset information, following other existing dataset-level metadata models. Whether additional searchable data type field(s) or an ontology for data types could improve usability is worth exploration, but tool developers should be sure not to duplicate existing search functions or create new data type ontologies. Note, this recommendation could also apply at the level of capturing data type in the schema.
    - **Incorporating dates for governance information:** As governance information evolves over time (e.g., agreements and consents are updated), users will benefit from a last updated date that reflect when governance information was generated, accessed, or updated.
    - **Visualizing or explaining policy silence:** this prototype presented a number of policy 'blanks' where an action like linkage is not addressed by a policy, as the schema adopted the open world assumption. This concept was difficult for users to understand and led to concerns about the trustworthiness of the information presented in the prototype. When presenting policy 'blanks', consider how to guide the user's interpretations (e.g., the blank cell/missing information is neither a permission nor prohibition).

6 Conclusion

The project team completed a proof-of-concept implementation project to develop and test a governance metadata visualization prototype, based on the metadata schema. Development work reflected an agile design, progressing in stages from examining the source governance information from 11 HHS and other federally funded datasets, building a back-end governance metadata relational database populated by real-world governance information, and then developing a visualization tool to render governance information for linkage implementation decision making. The project team collaborated with researchers and policy experts as co-designers and testers, conducted usability testing, and developed open-source documentation to support others to innovate further. The prototype includes six pages that support users to make an assessment about the feasibility of a linkage implementation, based on governance metadata from the selected datasets:

- Welcome and Home page
- Select Datasets page
- Compare Governance Information page
- Assess Linkage Feasibility page
- View Action Steps page
- About page
- Glossary

Within these pages, the prototype is primarily focused on policies, rules, and duties. Six user testers successfully navigated the governance information in the prototype validating that the schema can describe data governance information in a standard way to facilitate human interpretation and machine readability. Usability evaluation demonstrated that governance metadata can be visualized to support researchers to make an assessment about the feasibility of linkage implementations. The usability evaluation generated recommendations for future schema improvements and data visualization tools evolution. User testers emphasized that the task of linkage implementation decision making is inherently challenging and subjective. Developing the Assess Linkage Feasibility page, in the context of the stages of the dataset lifecycle, was the most challenging and resource-intensive aspect of this project. Feedback from user testers confirmed that the determination of linkage feasibility is not binary. Determinations on whether linkage is not only feasible but practical are made based on whether a multiple parties can collaborate to design a linkage implementation that adheres to and complies with the conditions and duties as outlined. All those who interacted with the prototype were supportive about its value for visualization of structured governance metadata and its potential to advance linkage implementations for research.

Future work to evolve governance metadata visualization tools and other related resources to support policy analysis for data linkage will require collaboration among data providers, repositories, funders, and policy/legal experts to bring multiple perspectives about decision making for linkage implementations.

If widely adopted, this work would contribute to making data more findable, accessible, interoperable, and reusable and promote trust and appropriate oversight in linking individual-level participant data when combined from different sources for research. A refined metadata governance schema and

production-level governance information visualization tools could be leveraged throughout the HHS and NIH research ecosystem, supporting innovative and responsible research to improve health outcomes for all Americans.

7 Visualization Prototype Glossary of Terms

Academic Research: Purposes associated with conducting or assisting with research conducted in an academic context; for example, within universities.

Access: To acquire data from a data repository or other data sharing system.

Access Type: The approach for making data available for use by others; for example, open access or controlled access.

Account Numbers: The presence of the personally identifiable information element of an account number.

Agreement: A Policy that grants the assignee a Rule over an Asset from an assigner.

Anonymize: To anonymize all or parts of an Asset.

Any Other Unique Identifying Numbers or Codes: The presence of any other personally identifiable information elements.

Approve: To provide permission to perform the requested action.

Approved Party: The party receiving approval.

Approved Purpose: The allowed activity or purpose for which the operator wishes to use the asset.

Approving Party: The party providing approval.

Assent: An assent form is used to express willingness to participate in research by persons who are, by definition, too young to give informed consent but are old enough to understand the proposed research in general, its expected risks and possible benefits, and the activities expected of them as subjects. An assent form is a type of consent form.

Assignee: The Party that is the recipient of the Rule.

Assigner: The Party that is the issuer of the Rule.

Biometric Identifiers: The presence of the personally identifiable information element of a biometric identifier.

Certificate License Numbers: The presence of the personally identifiable information element of a certificate license number.

Certification: An attestation that an official status has been earned by satisfying defined requirements, or the act of providing such a status, as proof that something has happened or defined standards have been met or will be upheld in the future.

Certification Organization: An organization that grants or approves certifications or certificates.

Classify: To organize data by relevant categories so that it may be used and protected more efficiently.

Collect: The process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.

Complete Training: To complete training.

Consent: The IRB-approved written record that is in compliance with the Common Rule (45 CFR 46) and, as applicable, Protection of Human Subjects rules (21 CFR 50), and is used to demonstrate the consent by a participant or guardian to participate in research.

Consent Requirement: A requirement to obtain consent from a participant or organization.

Consented Party: The party receiving consent on an asset.

Consenting Party: The party granting consent on an asset.

Constraint: A boolean expression that refines the semantics of an Action and Party/Asset Collection or declares the conditions applicable to a Rule.

Contract: A contract is an agreement between parties, creating mutual obligations that are enforceable by law. (source: https://www.law.cornell.edu/wex/contract)

Contracted Party: The Party that is being contracted.

Contracting Party: The Party that is offering the contract.

Controlled Access: Established processes for verifying appropriate use of shared data, such as requiring verification of requestor identity, committee approval of proposed research use, and signing a data use agreement to access protected data. (modified from source: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-22-213.html)

Data Access Committee: A group of individuals who review all requests for access to datasets from external requestors and is composed of individuals with expertise in science, policy, or bioinformatics resources.

Data Coordinating Center: Academic and commercial data management and study oversight organizations, including contract research organizations, whose responsibilities include providing some or all the following services: study administration, financial management, protocol administration, data management, stakeholder communication and coordination, quality assurance, site monitoring, safety, regulatory, document management, post-study management, and close-out activities.
(source: https://www.niddk.nih.gov/-/media/Files/Research-Funding/Process/NIDDK-Guidance_DCC-Management-of-CCAs_Final_Version-1,-d-,1_OD-Approved_External-Website.pdf)

Data Enclave: A secure network through which confidential data, such as identifiable information from census data, can be stored and disseminated. In a virtual data enclave a researcher can access the data from their own computer but cannot download or remove it from the remote server. Higher security data can be accessed through a physical data enclave where a researcher is required to access the data from a monitored room where the data is stored on non-network computers.
(source: https://nnlm.gov/guides/data-thesaurus/data-enclave)

Data Provider: Institutions, organizations, and researchers that collect data from research participants or that collect administrative data and may also submit the data to a repository for sharing.

Data Repository: A physical location or virtual system for preserving, maintaining, and providing access to data. (modified from source: https://www.whitehouse.gov/wp-content/uploads/2022/05/05-2022-Desirable-Characteristics-of-Data-Repositories.pdf)

Data Requester: An individual or organization that requests access to a dataset that is not the participant.

Data Use Agreement: A document that establishes who is permitted to use and receive data, and the permitted uses and disclosures of such information by the recipient. (modified from source: https://www.hhs.gov/hipaa/for-professionals/special-topics/emergency-preparedness/data-use-agreement/index.html)

Dataset: A collection of related data records.

Dates: The presence of the personally identifiable information element of dates. Includes birth dates and death dates.

Deidentification Method: The method to remove the identifiers or any information that could directly identify a participant from a dataset to mitigate privacy risks to that participant.

Deidentified Dataset: A set of information that has had personally identifiable information (PII; e.g., a person's name, email address, or social security number), including identifying protected health information (PHI; e.g., medical history, test results, and insurance information), removed.

Deidentify: To remove identifying information. Deidentified data is participant information that has had personally identifiable information (PII; e.g., a participant's name, email address, or social security number), including protected health information (PHI; e.g., medical history, test results, and insurance information), removed. This is normally performed when sharing the data from a registry or clinical study to prevent a participant from being directly or indirectly identified. (modified from source: https://toolkit.ncats.nih.gov/glossary/de-identified-patient-data/)

Determination: Outcome of a ruling or legal decision.

Device Identifiers and Serial Numbers: The presence of the personally identifiable information element of device identifiers and serial numbers.

Disclosure Review Body: A group of individuals that establishes and operates processes and policies to ensure the public release of data products that do not reveal any information about the participants included in those datasets.

Email Addresses: The presence of the personally identifiable information element of an email address.

eq: Equals.

Fax Numbers: The presence of the personally identifiable information element of a fax number.

Full Face Photos and Comparable Images: The presence of the personally identifiable information element of a full face photo or comparable image.

Geographic Data: The presence of the personally identifiable information elements of geographic data. Includes street address, city, state, ZIP code, or elements of a geocode.

Government Organization: An international, federal, state, tribal, or local government organization. Inclusive of departments or divisions within a larger government organization.

Guardian: An individual who is authorized under applicable state or local law to consent on behalf of a child to general medical care. Inclusive of a child's biological or adoptive parent. (source: https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46/subpart-D/section-46.402)

Health Plan Beneficiary Numbers: The presence of the personally identifiable information element of a health plan beneficiary or member number.

Institutional Review Board: The institutional entity charged with providing ethical and regulatory oversight of research involving human subjects, typically at the site of the research study. (source: https://orwh.od.nih.gov/toolkit/human-subjects-protections/institutional-review-board)

IP Addresses: The presence of the personally identifiable information element of an IP address.

Law (regulations and statutes): A system of rules created and enforced by governmental bodies that regulate the behavior of individuals, organizations, and governmental entities. Inclusive of statutes and regulations.

Limited Dataset: A set of information about a participant that excludes 16 direct identifiers specified in the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule and may be used or disclosed, for purposes of research, public health, or healthcare operations, without obtaining either a participant's authorization or a waiver or an alteration of authorization for its use and disclosure, with a data use agreement. (modified from source: https://privacyruleandresearch.nih.gov/pr_08.asp)

Link: To combine information from a variety of data sources for the same participant. (source: https://hcup-us.ahrq.gov/datainnovations/raceethnicitytoolkit/or19.jsp) Synonymous with record linkage.

Linkage Method: The method used for linkage. Inclusive of PPRL and non-PPRL methods.

Make Determination: To reach a decision or draw conclusions.

Medical Record Numbers: The presence of the personally identifiable information element of a medical record number.

Metadata: A set of data that describes and gives information about other data.

Minor Participant: Refers to children or participants who have not attained the legal age for consent to treatments or procedures involved in the research, under the applicable law of the jurisdiction in which the research will be conducted. (modified from source: https://www.hhs.gov/ohrp/regulations-and-policy/guidance/faq/children-research/index.html)

Names: The presence of the personally identifiable information element of a name. Includes first, middle, and last name.

Obtain Approval: To obtain verifiable approval to perform the requested action.

Obtain Consent: To obtain verifiable consent to perform the requested action in relation to the Asset.

Output: The result of a process.

Participant: A healthy human or a patient who is or becomes a participant in research. (modified from source: https://policymanual.nih.gov/3014-001#C)

Phone Numbers: The presence of the personally identifiable information element of a phone number.

Policy: A formal statement of intent or plan of action that is adopted by an organization and defines specific procedures, rules, or regulations that individuals are expected to adhere to or follow.

Policy (ODRL): A non-empty group of Permissions and/or Prohibitions.

Principal Investigator: The investigator with the overall responsibility for the designing, conducting, and reporting of the research, and ensuring both the protocol and the research team's actions are compliant with law, regulation, and NIH policy, even when certain aspects of the research are delegated to other investigators. (modified from source: https://ohsrp.nih.gov/confluence/display/ohsrp/Chapter+2+-+Roles+and+Responsibilities+of+the+Principal+Investigator#:~:text=The NIH Principal Investigator (PI,research to appropriately qualified individuals)

Prior Linkage: A set of information about prior linkages of a dataset.

Privacy Board: A group of individuals who review and approve research uses and disclosures of data to ensure that the privacy rights of research participants are protected.

Process: A procedure that individuals are expected to adhere to or follow.

Purpose: A defined purpose for exercising the action of the Rule.

Recipient: The party receiving the result/outcome of exercising the action of the Rule.

Reidentify: A general term for any process that re-establishes the relationship between identifying data and a participant. (modified from source: https://csrc.nist.gov/glossary/term/re_identification)

Request: A Policy that proposes a Rule over an Asset from an assignee.

Research Use: Working with data for scientific research or other analytical purposes.

Review Committee: A group of individuals convened to review materials, provide approvals, and issue determinations on requested actions.

Review Policy: To review the Policy applicable to the asset.

Rule: An abstract concept that represents the common characteristics of Permissions, Prohibitions, and Duties.

Safe Harbor Method: A method for deidentification by removal of 18 HIPAA identifiers. (modified from source: https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#safeharborguidance)

Secondary Link: To link a dataset for a secondary purpose distinct from the original or primary intended purpose.

Sell: To transfer the ownership of the Asset to a third party with compensation and while deleting the original asset.

Share: The act of making an asset such as a dataset available for use by others.

Sign: Signing a document or agreement.

Social Security Numbers: The presence of the personally identifiable information element of a social security number.

Submit: To submit a document such as a form, application, or protocol.

Transfer: To transfer the ownership of the Asset in perpetuity.

URLs: The presence of the personally identifiable information element of a URL.

Vehicle Identifiers and Serial Numbers: The presence of the personally identifiable information element of vehicle identifiers and serial numbers.

Virtual Location: An identified location of the IT communication space that is relevant for exercising the action of the Rule.

8 Glossary

Term	Definition
Accessibility (data)	To be accessible, metadata and data should be readable by humans and machines, and must reside in a trusted repository (NIH NLM)
Aggregate data	Summary statistics compiled from multiple sources of individual-level data (NIH aggregate data)
Authorization	Permission provided by a law/regulation/policy or an authority, or an agreement to perform data lifecycle activities, including collecting, linking, sharing, accessing, or using the data
Common data model (CDM)	A CDM standardizes the definition, format, and model content of data across participating data partners so that standardized applications, tools, and methods can be applied (PCORnet CDM)
Controlled access	Application and eligibility requirements need to be met and approved (e.g., by a data access committee) to gain access (NIH controlled access A) "Controlled access" and "access controls" refer to measures such as requiring data requesters to verify their identity and the appropriateness of their proposed research use to access protected data (NIH controlled access B)
Controls	Processes established to ensure compliance with governance for data sharing, access, and use (e.g., user must access data in a physical enclave, user must sign data use agreement, user must receive data access committee approval)
Data access	Acquiring data from a data repository or other data sharing system
Database/data repository	Virtual data storage that stores, organizes, and validates data, and makes the data accessible for use by others
Data collection	Obtaining data from participants for research, clinical, or administrative purposes
Data governance	As defined in this report, governance or data governance comprises the collective set of rules and controls that define and enforce how data are handled across the data lifecycle including: appropriate data collection, sharing, linking, access, and use. Data governance addresses privacy protections, ethics, compliance, risk management, and other requirements and derives from a variety of sources such as participant consent, IRB determinations, laws, agreements, and policy documents.
Data linkage/record linkage	Combining information from a variety of data sources for the same individual (AHRQ record linkage); in the context of this report, it is synonymous with individual level dataset linkage. This concept is complicated because new record linkage implementation requires effort from each of the other phases of the data lifecycle.
Data masking	The process of systematically removing a field or replacing it with a value in a way that does not preserve the analytic utility of the value, such as replacing a phone number with asterisks or a randomly generated pseudonym. (NIST masking)

Term	Definition
Data originator/ contributor/submitter	Institutions/organizations/researchers that collect data from patients or study participants or that collect administrative data; they may also be the party to submit the data to a repository for sharing
Data pseudonymization	De-identification technique that replaces an identifier (or identifiers) for a data principal with a pseudonym in order to hide the identity of that data principal (NIST pseudonymization)
Data science	Interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data
Dataset	Collection of related sets of information composed of separate elements that can be manipulated computationally as a unit
Data sharing[p]	Making data available to the broader data user community; for example, by submitting the data to a data repository for dissemination
Data standards	Documented agreements on representation, format, definition, structuring, tagging, transmission, manipulation, use, and management of data
Data steward	A formal position or an assigned accountability with responsibility for the following areas (HHS data steward): • Adherence to an appropriately determined set of privacy and confidentiality principles and practices • Appropriate use of information from the standpoint of good statistical practices (such as by not implying cause and effect when the data only point to correlation) • Limits on use, disclosure, and retention • Identification of the purpose for a specific use of the data • Application of "minimum necessary" principles • Verification of receipt by the correct recipient, wherever possible • Data deidentification (HIPAA-defined and beyond) • Data quality, including integrity, accuracy, timeliness, and completeness (NCVHS data steward)
Data use	Working with data for secondary research or other analytical purposes
Data use agreement	A document that establishes who is permitted to use and receive data, and the permitted uses and disclosures of such information by the recipient (modified from HHS data use agreement)
Data user (or secondary data user)	A person who accesses and uses data collected by another party for new research purposes
Deductive disclosure	Disclosure is revealing information that relates to the identity of a data subject, or some sensitive information about a data subject through the release of either tables or microdata (HHS deductive disclosure)
De-duplication	The process of removing redundant patient records from a database (CDC de-duplication)

[p] The act of data sharing, which we generally define as making data accessible to the broader data use community, often encompasses multiple steps and parties.

Term	Definition
De-identification	De-identified patient data is patient information that has had personally identifiable information (PII; e.g., a person's name, email address, or social security number), including protected health information (PHI; e.g., medical history, test results, and insurance information), removed. This is normally performed when sharing the data from a registry or clinical study to prevent a participant from being directly or indirectly identified. (NIH de-identification)
Electronic health records (EHRs)	EHRs are electronic versions of the paper charts in a doctor's or other healthcare provider's office. An EHR may include medical history, notes, and other information about the patient's health including symptoms, diagnoses, medications, lab results, vital signs, immunizations, and reports from diagnostic tests such as x-rays. (HHS EHR)
Enclave	A data enclave is a secure network through which confidential data, such as identifiable information from census data, can be stored and disseminated. In a virtual data enclave, a researcher can access the data from their own computer but cannot download or remove it from the remote server. Higher security data can be accessed through a physical data enclave where a researcher is required to access the data from a monitored room where the data is stored on non-network computers. (NLM enclave)
Entity resolution	Process of joining or matching records from one data source with another that describes the same entity (Census Bureau entity resolution) In PPRL, hash codes/tokens are used to match individual records without using PII/PHI (N3C entity resolution)
FAIR	Findable, Accessible, Interoperable, Reusable
FAIR Guiding Principles	A set of guiding principles for scientific data management and stewardship that describe distinct considerations for contemporary data publishing environments with respect to supporting both manual and automated deposition, exploration, sharing, and reuse
Findable (data)	For data to be findable there must be sufficient metadata and a unique and persistent identifier, and the data must be registered and indexed in a searchable resource (NIH NLM)
Governance	Governance or data governance, as defined in this report, comprises the policies, limitations, processes, and controls that address ethics, privacy protections, compliance, risk management, or other requirements for a given record linkage implementation across the data lifecycle

Term	Definition
HIPAA Privacy Rule	The Standards for Privacy of Individually Identifiable Health Information are codified in 45 CFR Parts 160 and 164 promulgated by the U.S. Department of Health and Human Services under the Health Insurance Portability and Accountability Act (HIPAA) of 1996. The HIPAA Privacy Rule establishes national standards to protect individuals' medical records and other individually identifiable health information (collectively defined as "protected health information") and applies to health plans, healthcare clearinghouses, and those healthcare providers that conduct certain healthcare transactions electronically. The Rule requires appropriate safeguards to protect the privacy of protected health information and sets limits and conditions on the uses and disclosures that may be made of such information without an individual's authorization. The Rule also gives individuals rights over their protected health information, including rights to examine and obtain a copy of their health records, to direct a covered entity to transmit to a third party an electronic copy of their protected health information in an electronic health record, and to request corrections. (HHS Health Information Privacy)
Honest broker	A party that holds deidentified tokens ("hashes") and operates a service that matches tokens generated across disparate datasets to formulate a single Match ID for a specific use case (N3C honest broker)
Institutional Review Board (IRB)	An IRB is the institutional entity charged with providing ethical and regulatory oversight of research involving human subjects, typically at the site of the research study (NIH IRB) An Institutional Review Board is an appropriately constituted group that has been formally designated to review and monitor biomedical research involving human subjects. An IRB has the authority to approve, require modifications in (to secure approval), or disapprove research. This group review serves an important role in the protection of the rights and welfare of human research subjects. (FDA IRB)
Interoperability	According to section 4003 of the 21st Century Cures Act, the term "interoperability," with respect to health information technology, means such health information technology that—"(A) enables the secure exchange of electronic health information with, and use of electronic health information from, other health information technology without special effort on the part of the user; (B) allows for complete access, exchange, and use of all electronically accessible health information for authorized use under applicable State or Federal law; and (C) does not constitute information blocking as defined in section 3022(a)" (HIT interoperability)
Interoperability (data) in computer systems	The ability of data or tools from non-cooperating resources to integrate or work together with minimal effort (the FAIR Guiding Principles for scientific data management and stewardship) Data must share a common structure, and metadata must use recognized, formal terminologies for description (NLM interoperable)
Letter of determination	A letter of determination documents an IRB decision on the status of research (HHS letter of determination)
Limitations	Restrictions on data linkage and use (e.g., dataset must only be linked with other disease-relevant data, dataset must be used in a physical enclave)

Term	Definition
Machine learning	A field of computer science that gives computers the ability to learn without being explicitly programmed by humans
Metadata	Information describing the characteristics of data including, for example, structural metadata describing data structures (e.g., data format, syntax, and semantics) and descriptive metadata describing data contents (e.g., information security labels) (NIST metadata)
Metadata schema	A metadata schema is a structured set of metadata elements and attributes, together with their associated semantics, that are designed to support a specific set of user tasks and types of resources in a particular domain. A metadata schema formally defines the structure of a database at the conceptual, logical, and physical levels. (Taylor, A. G. (2004). Introduction to cataloging and classification (10th ed.))
Ontology	A set of terms or concepts defining the properties or identities of subjects (e.g., genes, proteins, conditions) and relationships between them; similar to a standardized vocabulary
Open access	Data within this category presents minimal risk of participant identification. Access to these data does not require user certification, and researchers may explore data content without restriction. (NCI open access) No access restrictions or registration required to access (NIH open access) [see also data access model]
Patient identifier	Unique data used to represent a person's identity and associated attributes (NIST patient identifier)
Personally identifiable information (PII)	Any information that can be used to distinguish or trace an individual's identity, either alone or when combined with other information that is linked or linkable to a specific individual (NIST PII) and (CODI PII)
Privacy preserving record linkage (PPRL)	A technique identifying and linking records that correspond to the same entity across several data sources held by different parties without revealing any sensitive information about these entities (UK Office for National Statistics)
Protected health information (PHI)	Individually identifiable health information that is transmitted or maintained in any form or medium (electronic, oral, or paper) by a covered entity or its business associates, excluding certain educational and employment records (NIH PHI)
Provenance	The documented trail that accounts for the origin of a piece of data and where it has moved from to where it is presently (NLM provenance)
Reusable (data)	Data and collections must have clear usage licenses and clear provenance, and must meet relevant community standards for the domain (NLM reusable)

9 Abbreviations and Acronyms

Acronym	Definition
AFCARS	Adoption and Foster Care Analysis and Reporting System
AHRQ	Agency for Healthcare Research and Quality
AWS	Amazon Web Services
CDAC	Controlled Data Access Coordination
CMS	Centers for Medicare & Medicaid Services
COVID	Coronavirus Disease
DUA	Data Use Agreement
EC2	Enterprise Cloud Computing
EHR	Electronic Health Record
FAIR	Findable, Accessible, Interoperable, and Reusable
FDA	Food and Drug Administration
FFRDC	Federally Funded Research and Development Center
HHS	Department of Health and Human Services
HIPAA	Health Insurance Portability and Accountability Act
IRB	Institutional Review Board
JSON	JavaScript Object Notation
MTF	Monitoring the Future
N3C	National Clinical Cohort Collaborative
NCI	National Cancer Institute
NICHD	National Institute of Child Health and Human Development
NHANES	National Health and Nutrition Examination Survey
NIH	National Institutes of Health
NLM	National Library of Medicine
NSDUH	National Survey of Drug Use and Health
ODRL	Open Digital Rights Language
ODSS	Office of Data Science and Sharing

Acronym	Definition
OS-PCORTF	Office of the Secretary Patient-Centered Outcomes Research Trust Fund
PHI	Protected Health Information
PI	Principal Investigator
PII	Personally Identifiable Information
PPRL	Privacy Preserving Record Linkage
RADx	Rapid Acceleration of Diagnostics
RE-AIM	Reach Effectiveness Adoption Implementation Maintenance Reach
TEP	Technical Experts Panel
T-MSIS	Transformed Medicaid Statistical Information System
U.S.	United States
UTAUT	Unified Theory of Acceptance and Use of Technology
WARP	Web Application Rapid Prototyping

Appendix A: Technical Experts Panel Membership

Table 5: Technical Experts Panel Membership

Name	Affiliation
Age Chapman, PhD	Professor of Computer Science, University of South Hampton
Mike Conway, MSc	Data Systems Architect/Engineer, Office of Data Science, National Institute of Environmental Health Sciences
Kerry Goetz, PhDc, MS	Senior Advisor for Data Science, National Eye Institute
Brian Gugerty, DNS, MS	Healthcare Data Standards Specialist, All of Us Research Program (NIH)
Ryan Harrison, PhD	Presidential Innovation Fellow, Centers for Disease Control and Prevention, Data Modernization Initiative
Rui Li, PhD, MS	Director, Division of Research, Office of Epidemiology and Research, Maternal and Child Health Bureau, Health Resource and Services Administration
Frank Manion, PhD, MS	Vice President for Innovations at Melax Technologies, Intelligent Medical Objects (IMO) Health
S. Trent Rosenbloom, MD, MPH	Vice Chair for Faculty Affairs, the Director of Patient Engagement, and a Professor of Biomedical Informatics, Vanderbilt University Medical Center
Elizabeth E. Umberfield, PhD, RN, NI-BC	Nurse Scientist, Division of Nursing Research and Department of Artificial Intelligence and Informatics, Mayo Clinic

Appendix B: Co-designers

Table 6: Co-designers

Researcher	Affiliation
Brian Gugerty, DNS, MS	Healthcare Data Standards Specialist All of Us Research Program (NIH)
Dr. Chun-Ju (Janey) Hsiao, PhD	Program Officer for Data Search and Discovery in the Integrated Infrastructure and Emerging Technology (IIET), NIH ODSS
Jacob Kean, PhD	Associate Professor in Health System Innovation and Research Department of Population Health Sciences, University of Utah, and Research Scientist, Salt Lake City VA Health Care System
Frank Manion, PhD, MS	VP for Innovations at Melax Technologies, IMO Health
Heidi J. Sofia, PhD	Deputy Director National Institutes of Health National Center for Biotechnology Information, National Library of Medicine (NLM)

Appendix C: Usability Evaluation Session Script

Introduction

Thank you for joining us today! We appreciate you giving us this time to help us evaluate the usability of the governance metadata visualization prototype that we have developed.

We will start with brief introductions. I will outline the tasks we will be performing to evaluate the usability of the prototype, and then we can get started.

On the call today we have me as the primary facilitator for this session, and two others from the MITRE team who you may have met at the orientation session. They will be observing the session and taking notes.

Purpose of the Usability Evaluation

As a quick refresher, we have worked with the NIH Office of Data Science and Sharing, National Institute of Child Health and Human Development, to develop a robust metadata schema for data governance information relevant to linking individual-level participant data and sharing and using linked datasets.

We are now testing the use of the metadata schema in a proof-of-concept data visualization prototype that displays governance information for datasets that have been entered into a relational database of data governance.

Structured governance metadata can facilitate the determination of whether a dataset can be linked (combined with data relating to the same person from other sources) and if so, what rules flow down to the linked dataset.

The prototype enables research study teams to explore data governance information for a collection of pre-selected datasets that they would like to link together to better understand the rules around linking those datasets.

In this session, we will be testing the usability of this governance visualization prototype to better understand the governance of research datasets throughout the data lifecycle.

Governance is the policies, limitations, processes, and controls that address ethics, privacy protections, compliance, risk management, or any other requirements necessary to implement record linkage across the data lifecycle.

When we talk about the **data lifecycle**, we are talking about five phases: dataset collection, dataset linkage, dataset sharing, dataset secondary access, and dataset secondary use. I have the definitions for all of these phases on the screen.

- **Dataset collection** means a primary study collects the data and initiates sharing.
- **Dataset linkage** means combining information from a variety of data sources for the same individual.
- **Dataset sharing** means making data available to the broader data user community; for example, by submitting the data to a data repository for dissemination. The act of data sharing, which we

generally define as making data accessible to the broader data use community, often encompasses multiple steps and parties.
- **Secondary dataset access** means acquiring data from a data repository or other data sharing system for secondary research purposes.
- **Secondary dataset use** means working with data for secondary research or other analytical purposes.

You will be taking on the **role of a researcher**, with this user story:

- As a small business innovation researcher, I want to study the success of algorithms developed for diagnosing COVID in children by linking/combining research data on these diagnostics (RADx) with electronic health records (PEDsnet) and claims data (CMS), so I can extend selected algorithms for the diagnosis of other pediatric infectious diseases.
- 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.

You have selected three datasets: PEDSnet, the Rapid Acceleration of Diagnostics Data Hub (RADx), and the Transformed Medicaid Statistical Information System (T-MSIS).

This prototype assumes that you have already gone through the process of selecting these datasets by reviewing their contents using a separate dataset cataloging tool, and your goal with this prototype is to look at the governance information for these datasets and determine whether dataset linkage is feasible, and how to accomplish a linkage if you decide to move forward.

The goal of this session is not to test your knowledge of data governance or the selected datasets. The goal is to see if the prototype can be easily used by researchers, and if it's helpful for reviewing governance information.

Session Flow

This session will start with some basic questions about your research and experience with dataset linkage for individual-level data.

Then we will ask you to interact with the prototype and answer questions related to the governance information of the three selected datasets, and we will observe those interactions.

At the end, we will have questions about your impression of the prototype's usability and usefulness, challenges you faced using the prototype, and suggestions for improvements.

Agreement to Participate

As a user tester, this evaluation session is completely voluntary. If you need a break, or want to stop for whatever reason, please let me know and we can take a break or end the session if needed.

This usability evaluation was reviewed by the MITRE IRB and was deemed exempt from human subjects' review.

We have set up Microsoft Teams to audio record the session and generate a transcript. Once the transcript is generated, we will only use the transcript for analyses. If needed, we will use the audio file for clarification, then delete the audio file two weeks after your session.

Your responses will be kept confidential, and your responses will be anonymized before we incorporate them into our final report, summarized along with the input we receive from other usability evaluation participants.

- Do you agree to participate as a user tester in this usability evaluation?
    - Yes / No
- Do you have any questions at this point?

Thank you again for participating - I will now start recording through Teams.

Initial Questions

I have a few initial questions related to your experience as a researcher and experience with dataset linkage.

- Approximately how many years have you conducted biomedical research?
- How would you describe your main field of biomedical research? What is your research domain?
- Do you have experience with linking individual-level data from two or more datasets together for research purposes?
    - Yes / No
- [If "Yes"] How many research studies have you conducted where individual-level data from multiple datasets were linked?
- [If "Yes"] In the past when you have linked data at an individual level across multiple datasets, have you done this in cases where:
    - All of the datasets were owned by my research institution
    - Datasets were owned by multiple different research institutions
    - All of the datasets were publicly available datasets
- Do you have experience with oversight of data governance in your role?
    - Yes / No
- Do you have experience with the management or governance of a clinical data network or consortium?
    - Yes / No
- [If "Yes"] How many years have you been involved with clinical data networks or consortia?

Governance Metadata Visualization Interaction

Data Entry Instructions

Thank you for answering those initial questions! Now we will ask you to explore data governance information in the prototype.

I want you to navigate to the data visualization prototype using the link that I have placed in the chat: https://warp.mitre.org/data-linkage-governance/query-prototype/vis#/. After I am done with the instructions, I will ask you to share your screen so that we can observe your experience in using the prototype.

I have a **few instructions for you before you begin**.

- During the evaluation session, I will be asking you to perform tasks in the prototype and will also be asking you questions about the information that is displayed in the prototype. For each task I will give you a verbal description, and I may also leave you a description in the chat window as well that you can reference as you navigate the prototype. Please keep the chat window open if you can.
- Part of our evaluation considers the time required to complete different tasks. Please wait until I have finished explaining a task before interacting with the prototype. After I finish explaining the task, and when you feel that you have completed the task, please give me a verbal cue that you are done with the task. If you get to a point where you are unsure of how to complete the task, please tell us and we will move on to the next task.

As you are navigating this prototype, I'd like you to "think aloud" or verbalize your thought process as you navigate through the different screens, and your impressions on what you're seeing.

- Anything that you are thinking related to the prototype, or any questions that you may have as you work through the prototype, please talk through them out loud, including anything that is unclear or confusing.
- I may not answer all of the questions that you ask during the session, but please speak them out loud so that we can collect them for our discussion at the end of the session.

To manage time for this session, at points I may interrupt as you are navigating the prototype and present you with a new task or question to address.

Do you have any questions before we begin? Please share your screen, and I will start you out with your first task!

Prototype Navigation Procedures

- **Home Page**: Please take as much time as you need to read through the introduction text and click around to familiarize yourself with the functions on this page. If you have any general thoughts or questions as you review any of this, please speak to them aloud so that we can document them. Go ahead and get started and let me know when you are finished exploring!

- **Select Datasets page**: Now we will start with the first task, which is to select the three datasets. Based on what you see in this screen, where would you navigate to select the datasets you want to explore? Please navigate there now.
    - PEDSnet
    - Rapid Acceleration of Diagnostics Data Hub (RADx)
    - Transformed Medicaid Statistical Information System (T-MSIS)
- **Select Datasets page**: Continuing with the first task, please take as much time as you need to read through the instructions at the top of the page and review the contents to get a feel for what's on this page. If you have any general thoughts or questions as you review, please speak to them aloud so that we can document them. Go ahead and get started and let me know when you're finished exploring!
- **Select Datasets page**: Now I will ask you to select the three datasets described in the user scenario. I will put them in the chat window. Please start your selection now, remembering to speak aloud as you navigate the prototype.
    - PEDSnet
    - Rapid Acceleration of Diagnostics Data Hub (RADx)
    - Transformed Medicaid Statistical Information System (T-MSIS)
- **Select Datasets page**: Look at the three datasets that you have selected. Which dataset has the largest number of relevant “Policies,” and how many Policies does it have?
- **Compare Governance page**: Now we will start with the second task, which is to look at more details on the governance for all of our datasets. Based on what you see in this screen, where would you navigate to look at more details on the governance for all of the selected datasets? Please navigate there now.
- **Compare Governance page:** Please take as much time as you need to read through the instructions at the top here, click around, and review the content to get a feel for what's on this page. If you have any general thoughts or questions as you review any of this, please speak to them aloud so that we can document them. Go ahead and get started and let me know when you're done exploring!
- **Compare Governance page:** If you want to link these three datasets together, how many "Consent" type policies do you have to consider?
- **Compare Governance page:** How many Policies prohibit linkage, and what are the names of these policies?
- **Compare Governance page:** Looking at one of the policies that prohibits linkage—let's look at the T-MSIS Secondary Linkage Policy—under what conditions is linkage allowed?
- **Assess Linkage Feasibility page**: Now we will move on to the third task, which is to assess whether linkage is feasible, considering all of the governance information related to these datasets. Based on what you see in this screen, where would you navigate to help you assess this feasibility? Please navigate there now.
- **Assess Linkage Feasibility page**: Continuing with the third task, please take as much time as you need to read through the instructions at the top here, click around, and review the content to

get a feel for what's on this page. If you have any general thoughts or questions as you review any of this, please speak to them aloud so that we can document them. Go ahead and get started and let me know when you're done exploring!

- **Assess Linkage Feasibility page**: Looking at policies around dataset sharing, how many policies permit sharing if the product of a linkage is a deidentified dataset?
- **Assess Linkage Feasibility page**: Looking at policies about dataset linkage, how many policies permit dataset linkage?
- **Assess Linkage Feasibility page:** Looking at one of the policies around secondary dataset use, what does the PEDSnet Responsible Use of Data Agreement say about secondary use? I'll put the name of that policy in the chat. What does the "PEDSnet Responsible Use of Data Agreement" say about secondary dataset use?
- **View Action Steps page:** Now we will move on to the final task, which is to define a plan for implementing a linkage between these datasets. Based on what you see in this screen, where would you navigate to help you figure out next steps? Please navigate there now.
- **View Action Steps page**: Continuing with the last task, please take as much time as you need to read through the instructions at the top here, click around, and review the content to get a feel for what's on this page. If you have any general thoughts or questions as you review any of this, please speak to them aloud so that we can document them. Go ahead and get started and let me know when you're done exploring!
- **View Action Steps page**: Looking at the Action Steps to link these three datasets, how many actions would a Data Coordinating Center need to take?
- **View Action Steps page**: Looking at the Action Steps for the Data Requestor, how many documents need to be signed to link the RADx database?
- **Is Linkage Feasible?** Based on the information you have seen in this prototype, I want to get a sense for how feasible you think it would be to link these three datasets for a linkage implementation. Let's rank feasibility on a 10-point scale here, where 1 is trivial to link, and 10 is completely infeasible to link.

Discussion Questions

Thank you for the helpful feedback you provided during the walkthrough!

Now we are going to debrief with a few questions about the usability of the prototype and your experience exploring governance information for our three datasets.

I will show some statements here on these slides and ask you how much you agree or disagree with them on a 4-point Likert scale. Four will be most agreement, 1 will be the most disagreement. I will also have open ended discussion questions as well.

I'll take this time to emphasize that we are interested in both the "good and the bad" of your experience with this data visualization prototype. This is your opportunity to influence future work related to the collection and use of dataset governance information, and your honest feedback is critical.

- This prototype provides me with the information I need to determine whether dataset linkage is feasible from a governance information standpoint. [RE-AIM: Effectiveness / UTAUT: Perceived Usefulness]

- This prototype is easy to use. [UTAUT: Perceived Ease of Use]
- I would be likely to use this prototype in my own work. [RE-AIM: Implementation / UTAUT: Attitude Toward Behavior]
- I believe that my research institution would support my use of a prototype like this.
- After using this prototype, I feel that I have a better understanding of data governance.
- What unintended negative consequences may come from using a prototype like this? [RE-AIM: Effectiveness]
- Please rank these three screens in terms of their ease of use: Compare Governance, Assess Linkage Feasibility, and View Action Steps. A rank of 1 is the easiest to use, and a rank of 3 is the most difficult to use.
- Please rank these three screens in terms of their usefulness in making decisions about dataset linkage: Compare Governance, Assess Linkage Feasibility, and View Action Steps.
- What aspects of the prototype were helpful as you were reviewing governance information?
- What aspects of the prototype were challenging to work with as you were reviewing governance information?
- This prototype does not provide extensive information on the contents of different datasets, only information on dataset governance.
- Are there existing data catalogs, tools, or systems that you can think of where it would be useful to integrate a visualization prototype like this?
- Do you have any suggestions to improve this prototype, or future prototypes like this?

Appendix D: Usability Evaluation Analysis Codebook

Table 7: Usability Evaluation Codebook

Code	Source	Definition
Reach	RE-AIM	Individual-level measure of participation. Example measures: the absolute number, proportion, and representativeness of individuals who are willing to participate in a given initiative, intervention, or program, and reasons why or why not.
Effectiveness	RE-AIM	The impact of an intervention on important individual outcomes, including potential negative effects, and broader impact including quality of life and economic outcomes; and variability across subgroups (generalizability or heterogeneity of effects).
Adoption	RE-AIM	The proportion and representativeness of settings that adopt a given policy or program. (Setting levels) The absolute number, proportion, and representativeness of settings and intervention agents (people who deliver the program) who are willing to initiate a program, and why. Note that adoption can have many (nested) levels, for example, staff under a supervisor under a clinic or school, under a system, and within a community.
Implementation	RE-AIM	The extent to which a program is delivered as intended. At the setting level, implementation refers to the intervention agents' fidelity to the various elements of an intervention's key functions or components, including consistency of delivery as intended and the time and cost of the intervention. Importantly, it also includes adaptations made to interventions and implementation strategies.
Maintenance	RE-AIM	At the setting level, the extent to which a program or policy becomes institutionalized or part of the routine organizational practices and policies. Within the RE-AIM framework, maintenance also applies at the individual level. At the individual level, maintenance has been defined as the long-term effects of a program on outcomes after a program is completed. The specific time frame for assessment of maintenance or sustainment varies across projects.
Performance Expectancy	UTAUT	The degree to which using a technology will provide benefits to consumers in performing certain activities.
Social Influence	UTAUT	The extent to which consumers perceive that important others believe they should use a particular technology.
Effort Expectancy	UTAUT	The degree of ease associated with consumers' use of technology.
Facilitating Conditions	UTAUT	Perceptions of the resources and support available to perform a behavior.
Hedonic Motivation	UTAUT	The fun or pleasure derived from using a technology.
Price Value	UTAUT	Cognitive tradeoff between the perceived benefits of an application and the monetary cost for using it.

Code	Source	Definition
Experience	UTAUT	Opportunity to use a technology; the passage of time from the initial use of a technology by an individual.
Habit	UTAUT	Extent to which people tend to perform behaviors because of learning.
Policy Type Value	Evaluation Data	Discussion of whether organizing policies by type would help users review governance information.
User Perspectives	Evaluation Data	Differences in tool use based on the user's background and experience.
System Trust	Evaluation Data	Differing perceptions of the trustworthiness of the information presented in the tool.
Interpretation Ambiguity	Evaluation Data	Situations where testers differed in their interpretations of the same information or visual features presented in the tool.
Information Relevance	Evaluation Data	Discussion of information that was relevant or irrelevant to the decision-making process for determining whether datasets can be linked.
Role Clarity	Evaluation Data	Confusion over role definitions and how roles relate to rules and actions governing dataset linkage.
Discouragement	Evaluation Data	Indication that the information presented in the tool could discourage researchers from pursuing a linkage implementation.

References

[1] NICHD GitHub Repository Project: https://github.com/NIH-NICHD/Data-Linkage-Governance

[2] Privacy Preserving Record Linkage (PPRL) for Pediatric COVID-19 Studies Report: https://www.nichd.nih.gov/sites/default/files/inline-files/NICHD_ODSS_PPRL_for_Pediatric_COVID-19_Studies_Public_Final_Report_508.pdf

[3] NICHD Record Linkage Implementation Checklist: https://www.nichd.nih.gov/sites/default/files/inline-files/Record_Linkage_Implementation_Checklist.pdf

[4] PCORTF Pediatric Record Linkage Governance Assessment: https://www.nichd.nih.gov/sites/default/files/inline-files/PCORTF_Pediatric_Record_Linkage_Governance_Assessment_Formatted120423.pdf

[5] CMS Alliance to Modernize Healthcare (The Health FFRDC). Data Governance Metadata Standards: Landscape and Gap Analysis. Prepared under Contract No. 75N94023F00171. February 2024. https://www.nichd.nih.gov/sites/default/files/inline-files/Governance_Metadata_Standards_FINAL_revb.pdf

[6] CMS Alliance to Modernize Healthcare (The Health FFRDC). Data Governance Metadata Standards: Landscape and Gap Analysis. Prepared under Contract No. 75N94023F00171. February 2024. https://www.nichd.nih.gov/sites/default/files/inline-files/Governance_Metadata_Standards_FINAL_revb.pdf

[7] Open Digital Rights Language website: https://www.w3.org/ns/odrl/2/ODRL20.html

[8] Metadata Schema Github website: https://github.com/NIH-NICHD/Data-Linkage-Governance

[9] The Agile Alliance website: https://www.agilealliance.org/agile101/

[10] Johnson, Constance M., Todd R. Johnson, and Jiajie Zhang. "A user-centered framework for redesigning health care interfaces." *Journal of biomedical informatics* 38.1 (2005): 75-87.

[11] Re-AIM website: https://re-aim.org

[12] Venkatesh, Viswanath, James Y.L. Thong, and Xin Xu, Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology (February 9, 2012). MIS Quarterly, Vol. 36, No. 1, pp. 157-178, 2012, Available at SSRN: https://ssrn.com/abstract=2002388

[13] PCORTF Pediatric Record Linkage Governance Assessment: https://www.nichd.nih.gov/sites/default/files/inline-files/PCORTF_Pediatric_Record_Linkage_Governance_Assessment_Formatted120423.pdf

[14] Kushniruk, A. W., and V. L. Patel, Cognitive and usability engineering methods for the evaluation of clinical information systems. *Journal of biomedical informatics*, 37.1 (2004): 56–76. https://doi.org/10.1016/j.jbi.2004.01.003

[15] Venkatesh, Viswanath, James Y.L. Thong, and Xin Xu , Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology (February 9, 2012). MIS Quarterly, Vol. 36, No. 1, pp. 157-178, 2012, Available at SSRN: https://ssrn.com/abstract=2002388

[16] CMS Alliance to Modernize Healthcare (The Health FFRDC). Data Governance Metadata Standards: Landscape and Gap Analysis. Prepared under Contract No. 75N94023F00171. February 2024. https://www.nichd.nih.gov/sites/default/files/inline-files/Governance_Metadata_Standards_FINAL_revb.pdf