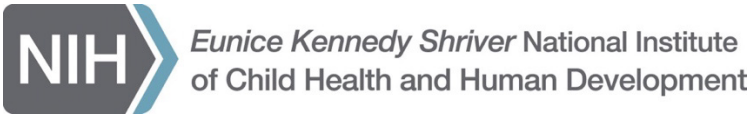
Implementation Report: Development of a Governance Metadata Collection Tool

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

October 14, 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 Sharing31 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 Collection Tool. Prepared under Contract No. 75N94023F00171. October 2024.

Authors

Emily Kraus, PhD, MPH
Susan C. Hull, MSN, RN, NI-BC, NEA-BC, FAMIA
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 into 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 standard exists. To progress the field, governance information 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), the *Digitizing Consent and Regulatory Metadata Towards Streamlining Governance of Pediatric COVID-19 Research Data Linkages* project supported NICHD ODSS and the Health FFRDC to develop the data governance metadata schema and build a data collection tool that tests the schema's capacity to accommodate real-world governance information about datasets. Learn more about the metadata schema on the NICHD GitHub Data Linkage Governance Repository.[1]

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 collection tool 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.

The Data Collection Tool prototype project has two aims. The first aim is to explore how governance information may be collected from researchers, as proxies for data providers, and ascertain which

[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.

governance information is the easiest and most challenging to collect. The second is focused on testing the data governance metadata schema to ascertain how well its structure and design perform in real-world data collection settings by focusing on these three lines of inquiry:

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

The Health FFRDC team collaborated with researchers as co-designers, conducted usability testing, developed open-source documentation to support others to innovate further on this proof-of-concept effort, and conducted a translation exercise to examine alignment with the data governance metadata schema. Community experts in the form of a Technical Experts Panel provided key guidance and feedback on this work. The result is a Governance Metadata Collection Tool designed around how biomedical researchers conduct research and manage data governance, rather than following the governance metadata organization based on the new schema alone. The tool includes 165 questions presented in 11 sections, organized by governance policies relevant to research (e.g., consent, Institutional Review Board (IRB), data use agreements (DUA), laws).

All those who interacted with the tool were unanimous about its value for collecting structured governance metadata and its potential for exchanging governance metadata to advance linkage implementations for research. The development and testing of the Data Collection Tool also highlighted ways that future iterations of the data governance metadata schema and future data collection tools could be improved. The governance metadata schema is ready for adoption, but it is a living product that will be enhanced over time, based the findings from this project and from future use. The Data Collection Tool was a prototype not meant for production, but the materials and lessons learned can be used to support new or improve existing workflows for collecting governance information. Future work to evolve governance metadata collection will require collaboration and input from data providers, institutional representatives, IRBs, and policy/legal experts to bring multiple perspectives for determining and communicating rules about a dataset.

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 production-level data collection 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 Department of Health and Human Services (HHS) Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) and the NIH Office of Data Science Strategy (NIH ODSS), the *Digitizing Consent and Regulatory Metadata Towards Streamlining Governance of Pediatric COVID-19 Research Data Linkages* 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.[2]

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: (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 sharing, and patient-centered outcomes research. The overall goal is to provide researchers, repositories, funders, policy/legal experts, and other community members involved in linking data for research 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 first proof-of-concept implementation project, the Governance Metadata Data Collection Tool.

[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.

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 tool development and subsequent efforts to evaluate the tool's usability. See Appendix A for TEP membership.

Alignment with OS-PCORTF Strategic Plan

In September 2022, the OS-PCORTF released its strategic plan for building data capacity for patient-centered outcomes research through coordinated, systematic efforts across federal agencies[3]. Data capacity, in a patient-centered outcomes (PCOR) context, refers to the availability and sustainability of data and analytic resources to address national health priorities. The strategic plan addresses a broad range of data sources, including clinical, clinical trial, social services, and administrative and claims data, and notes that issues of availability, quality, accessibility, and interoperability are significant hurdles to PCOR research. Varied, multi-modal data sources; data linkage; data analysis; and equitable access are the cornerstones of the PCOR data infrastructure.

The OS-PCORTF strategic plan articulates four interrelated goals and desired outcomes:

- Goal 1: Data Capacity for National Health Priorities
    - Outcome 1: Data, tools, and services to improve patient-centered outcomes research relevant to HHS priorities
- Goal 2: Data Standards and Linkages for Longitudinal Research
    - Outcome 2: Accessible, timely, interoperable, linkable, and longitudinal data
- Goal 3: Technology Solutions to Advance Research
    - Outcome 3: Robust real-world data across platforms and systems used to generate real-world evidence and expand data usage that informs patient, clinical, and policy decision making
- Goal 4: Person-Centeredness, Inclusion and Equity
    - Outcome 4: Accurate, relevant, and representative evidence is accessible to individuals; communities; and state, federal, and tribal programs when making health decisions

Goal 2 of the plan describes data standards and linkage for longitudinal research and includes activities to assess the impact of policies related to privacy, security, and consent on PCOR efforts and to build consensus-based linkage methodology. The aim of this project, to develop and test a generalizable metadata schema that facilitates decision making for PCOR dataset linkage and the subsequent sharing and use of linked datasets, aligns with Goal 2 of this plan. The project's goal to streamline decision making for record linkage should move the HHS community toward secure and appropriate data linkages and the responsible sharing and use of linked datasets for patient-centered outcomes research.

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[4]

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,[5] 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 data for research. 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. However, designing a new record linkage strategy also requires funders, researchers, data repositories, and other stakeholders to consider the other steps described in the checklist, such as considering additional controls to mitigate potential risks.

OS-PCORTF Pediatric Record Linkage Governance Assessment[6]

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 de-identified.

Governance Metadata Standards: Landscape and Gap Analysis Report[7,8]

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[9] (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[10] 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 (Figure 1).

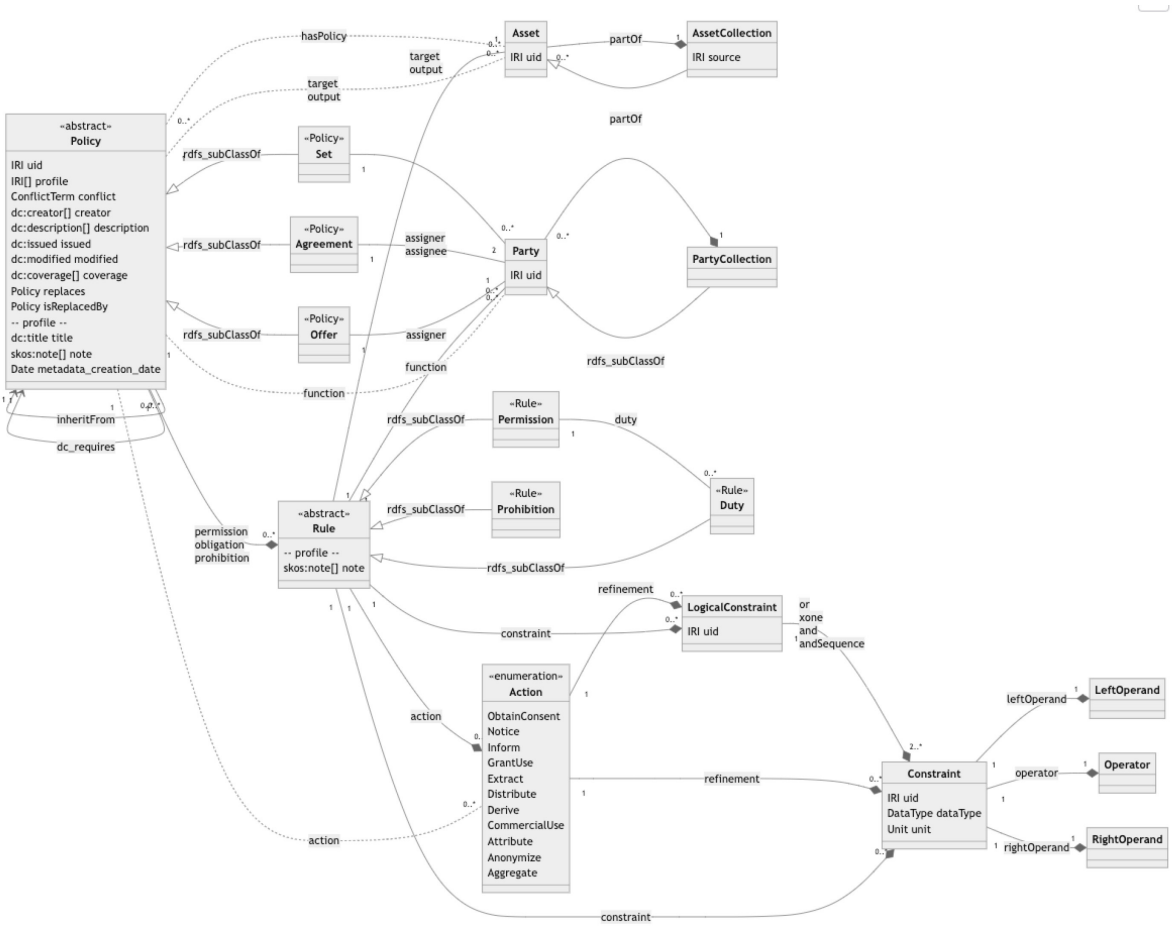


Figure 1: Data Governance Metadata Schema

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]. For example, a permission to use data with a duty to obtain approval from an institutional review board (IRB).

The data governance metadata schema extends the ODRL vocabulary by adding more than 70 additional terms and annotations required to accurately represent governance metadata. This new Data Governance Profile (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.

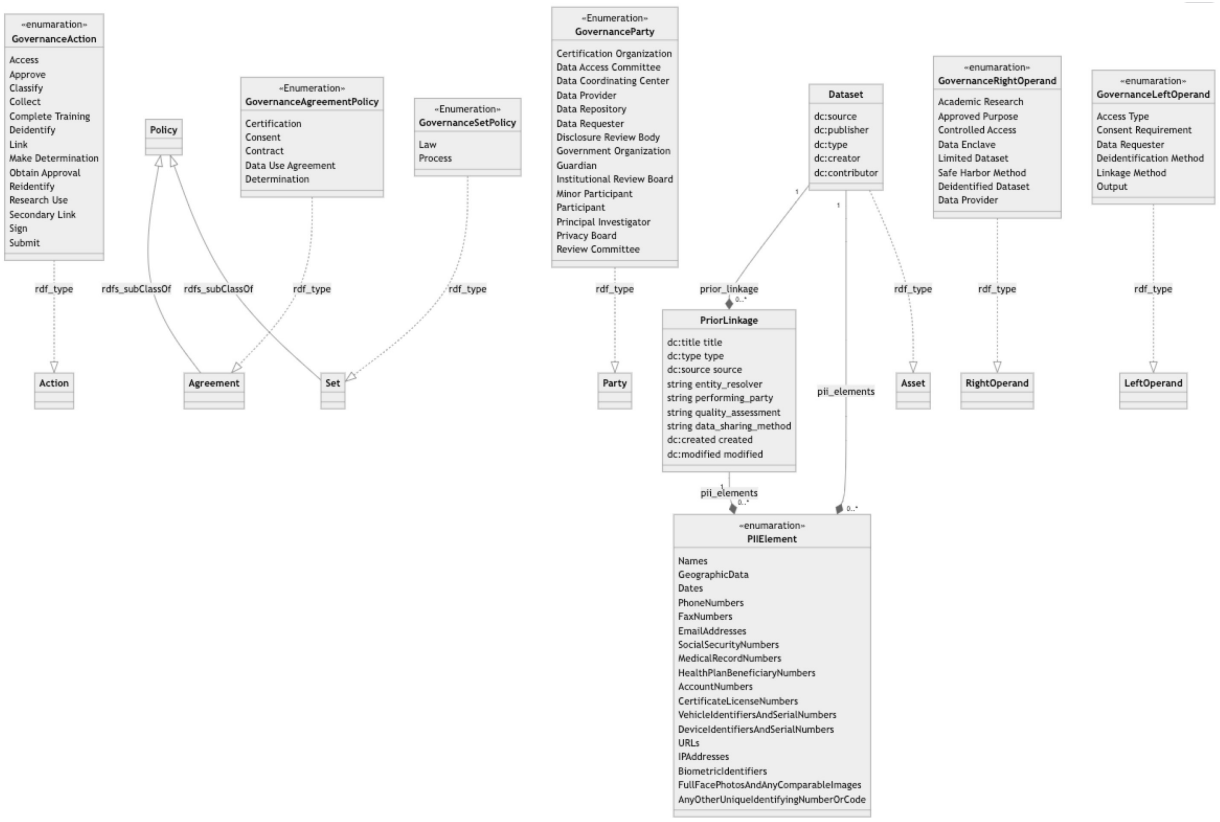


Figure 2: Data Governance Profile

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, OS-PCORTF, and HHS data ecosystems. To test the metadata schema, NICHD ODSS 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 that extends an existing data collection tool to test the data governance metadata schema. The report describes the data governance metadata schema, the methods and approach to develop the governance data collection tool, key findings from the development and usability evaluation, and recommendations for the metadata schema and future governance metadata collection work.

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

1. To explore how governance information may be collected from researchers, as proxies for data providers, and ascertain which governance information is the easiest and most challenging to collect.
2. To test how the data governance metadata schema performs in a real-world data collection setting. Testing the metadata schema is focused on three lines of inquiry:
    - What are the questions to solicit governance metadata?
    - Can a researcher answer questions about governance metadata?
    - Do the question responses generate metadata that fits within the schema? Do the value sets in the tool support governance metadata collection goals?

1.5 Audience

The intended audience of this public report includes: (1) researchers generating datasets from a study or program that are or could be linked, 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, sharing, 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 data collection tool through these steps:

1. Selecting an existing data collection tool, commonly used by researchers, that can be extended
2. Developing an early prototype using agile principles and methods[11]
3. Refining the tool with co-designers based on user-centered design principles and methods[12]
4. Completing tool development based on feedback from co-designers and TEP members
5. Conducting a usability evaluation using existing frameworks such as the Reach Effectiveness Adoption Implementation Maintenance Reach (RE-AIM) framework[13] and the extended unified theory of acceptance and use of technology (UTAUT)[14]

The project team worked in collaboration with NICHD ODSS and regularly sought guidance from the TEP and defined a user story (Table 1) to guide the governance metadata collection tool development:

Table 1: User Story for Tool

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 clinical or public health researcher, I want to share data with other researchers; for example, by submitting to a data repository.	I want to document how the dataset can be shared, linked, and used based on consent and other policies and regulations (and make updates as needed) so that potential secondary users have accurate information about these rules.	Today, there is no structured format for capturing requirements for how data can be shared, linked, and used based on consent or other requirements.

2.1 Existing Tool Selection

The rationale for extending an existing data collection tool was a desire to enhance one of the many data collection tools already used by researchers in the field, rather than building a new tool that would require promotion, adoption, integration, training, and maintenance.

The project team identified seven candidate tools including: Research Electronic Data Capture (REDCap),[15] Fast Healthcare Interoperability Resources® Structured Data Capture standard[16] (FHIR SDC), Open Data Kit[17] (ODK), Kobo Toolbox,[18] Hypertext Markup Language (HTML) Javascript,[19] S2O,[20] and African Partnership for Chronic Disease Research open-source electronic questionnaire[21] (APCDR EQ). NICHD ODSS and the TEP eliminated HTML Javascript, S2O, and APCDR EQ as options and selected REDCap, FHIR SDC standard, ODK, and Kobo Toolbox to undergo a tools readiness analysis. This analysis aimed to determine how candidate tools could be extended for governance data collection through a requirements analysis and estimation of the technical level of effort.

The project team defined requirements for a data collection tool, including:

- Be open source
- Have a robust user community
- Enable form business logic
- Provide a user interface
- Enable back-end data collection
- Support data exchange
- Leverage interoperability and standards
- Ensure adequate performance
- Require minimal load time
- Ensure privacy and security
- Enable integration with other software
- Be able to scale in an operational environment

The project team reviewed publicly available resources to determine how well candidate tools met each requirement.

For technical estimation, the project team defined the required steps for tool implementation and then used a consensus-based Agile technique known as “scrum poker” to estimate the level of effort for each component and assembled estimates into a total number of implementation hours. The project team enlisted expert developers to provide estimates based on the defined required technical steps for each candidate tool’s implementation.

The project team presented technical readiness analysis results to the TEP and NICHD ODSS for discussion. Appendix B presents detailed tools readiness analysis findings. No candidate tool met all requirements; REDCap and the FHIR SDC standard met most of the requirements. REDCap required an estimate of 38 hours to extend the tool for implementation. FHIR SDC, implemented as an instance of a FHIR Questionnaire embedded in a web application (e.g., the Lister Hill FHIR SDC Form Builder), required an estimate of 76 hours for implementation.

NICHD ODSS selected the FHIR SDC standard for the extension because its implementations inherently enable FHIR-based data exchange, it was not significantly more time consuming to develop, it is translatable to other FHIR users, and it is truly open source with potential for unlimited use, reuse, and adoption, with strong community adoption and extensive active user communities.

The U.S. HHS Office of the National Coordinator for Health Information Technology launched the FHIR SDC effort in 2015, as a collaborative initiative in partnership with federal health agency partners (including the National Library of Medicine [NLM], the Centers for Medicare & Medicaid Services [CMS], the Food and Drug Administration [FDA], the Agency for Healthcare Research and Quality [AHRQ], the Assistant Secretary for Planning and Evaluation, and the National Cancer Institute [NCI]) and 329 committed members to support standards-based forms and common data elements to improve the interoperability of the numerous ways in which forms capture administrative data, claims data, clinical information, research information, public health surveillance, and case reporting.

2.2 Questionnaire Template

The project team developed a questionnaire template with these components:

- **High-level question flow** is the sequence of governance information that will be collected.
- **Questions** are the text of the questions.
- **Responses** are free text or structured response options.
- For questions with structured responses, **response options** are the potential responses that a user may choose.
- **Response format** is whether the question allows the user to select one or multiple options and if the user can enter a custom free-text value.
- **Business logic** is the relationship between responses and subsequent questions (e.g., skip patterns).

- **Instruction text** is the instructions for the user presented for the overall questionnaire and for specific questions.
- **Help text** is the words that are displayed in a pop-up box when a user clicks on an information button.

The project team developed the initial questionnaire template in Microsoft Excel and draw.io. They drafted straightforward questions that mapped to each of the schema classes such as "Has this dataset previously been linked?" and "What is the name of this policy?". This method ensured that responses to each question would fit directly within a schema class.

Instruction text provided necessary context and guidance about what governance (and schema terms) meant, what governance information the tool aimed to collect, and how the user could expect to navigate the tool. For example: "In this questionnaire, a policy is defined as the source of rules that dictate how a dataset is handled across the data lifecycle. Examples of policies includes laws as well as documents like DUAs."

The questionnaire focused on five actions required for handling data across the data lifecycle:

- **Dataset collection** means a primary study collects the data and initiates sharing.
- **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 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.
- **Dataset linkage** means combining information from a variety of data sources for the same individual. This concept is complicated because implementing a new record linkage implementation often requires effort from each of the other phases of the data lifecycle.[22]

Data collection is listed as an action in the data lifecycle in the instructions that frame the data collection tool. However, when the tool poses questions about the permission rules within each policy type (e.g., "Does the consent permit dataset linkage?"), the tool does not include questions regarding the permission to collect the dataset. Rather, the questionnaire assumes that the dataset has previously been collected.

2.3 Early Prototype

The project team created an early prototype to serve as a minimum viable product for research co-designers to engage with and provide feedback on. The prototype was designed for data providers and their teams to enter governance information for a single dataset retrospectively (after the data collection has occurred). The questions came from the questionnaire template and were thus mapped to the schema classes. The questions and instructional text were also designed to inform the researcher entering governance information in this tool that linkage-relevant governance is sourced from two key

areas: (1) the primary study where data collection happened, and (2) the rules that apply across the data lifecycle for a given dataset that might be contributed to a hypothetical future linkage implementation.

FHIR Questionnaire Creation

The project team selected the Lister Hill FHIR SDC Form Builder,[23] an online resource developed by the Lister Hill Center (LHC) for Biomedical Communication at the National Library of Medicine, National Institutes of Health, to implement the questionnaire template and generate a FHIR Questionnaire. The LHC FHIR SDC Form Builder facilitates the creation and management of forms compliant with the FHIR SDC standards and is free to use and publicly available. Its intuitive and flexible interface helps users design, customize, and deploy digital FHIR Questionnaires that can seamlessly integrate with other systems. The form builder allows developers to download their FHIR Questionnaire in a JavaScript Object Notation (JSON) format and to upload an existing FHIR Questionnaire to continue editing.

The project team iterated on the questionnaire template by downloading and uploading the FHIR Questionnaire to the LHC FHIR SDC Form Builder and implementing semantic versioning to track changes. Once the questionnaire template had been implemented in FHIR, all subsequent modifications to the question text, responses, order and business logic were made in the LHC FHIR SDC Form Builder. The questions themselves were refined with feedback from the project team and co-designers and were evaluated based on their effectiveness at soliciting governance information. The team made decisions about how questions were worded and how they are implemented in the SDC tool, both to match the schema and to make it human understandable. Prepopulated response sets for questions came from the real-world examples uncovered in the collection of governance from the 11 datasets. The project team worked together to identify common values that appeared across the 11 datasets; for example, CIPSEA (Confidential Information Protection and Statistical Efficiency), FERPA (Family Educational Rights and Privacy Act), HIPAA Privacy Rule, The Common Rule: 45 CFR 46 Part A, and The Public Health Services Act were included as response options in the initial questionnaire for the question about Common Federal Laws. The response options, including those for the Common Federal Laws question, were further refined through the co-design and tool completion processes.

Front-End Application

The project team selected the LHC-Forms Widget[24] to provide a front-end application that renders the FHIR Questionnaire as a web form that respondents can use to answer the questions about governance information. The widget's front-end user interface was combined with HTML and JavaScript using the React[25] user interface library to provide basic infrastructure to allow the user to load the questionnaire, save a questionnaire response, load a previous questionnaire, and continue a questionnaire response.

The frontend displays the questionnaire but needs to work in conjunction with a backend to provide functionality like long-term data storage or data exchange. The project team included a download response function as a FHIR feature on the front-end application to demonstrate how a FHIR Questionnaire response can be generated from the tool.

Back-End Application

A data collection tool requires a service backend to track user sessions (i.e., the time spent completing a questionnaire) and store questionnaire responses. The data collection tool backend was developed as a

Ruby on Rails application that serves the questionnaire page to users, allows a completed questionnaire to be saved, and allows saved forms to be continued or reviewed. The interface between the frontend and the backend is a JSON Application Programming Interface (API). JSON is a lightweight data exchange format. Storage for the backend was provided using the PostgreSQL relational database.

Deployment

The data collection tool was deployed to the MITRE Web Application Rapid Prototyping (WARP) environment to allow access by co-designers, user testers, and all participants in the development and evaluation processes. WARP is a MITRE-hosted restricted environment that allowed access to the collection tool to be limited to a specified set of users and provided a pre-vetted secure environment that minimized deployment costs.

2.4 Co-design with Researchers

The project team engaged three pediatric biomedical researchers and one researcher from the TEP membership in a series of co-design sessions to support updates to the questions, instructional text, and overall order of the data collection tool (See Appendix C). Researchers were invited to attend four one-hour biweekly co-design sessions and to complete weekly interval assignments with email feedback, estimated to take up to one hour. The project team sent an agenda and discussion slides ahead of time, and co-design sessions were recorded. Co-designers were provided access to the prototype through the MITRE WARP environment and were asked to enter real-world dataset governance into the data collection tool as part of the interval assignments between co-design sessions and send feedback by email to the project team for consideration. In each session, the project team presented the most current version of the prototype and collected feedback from co-designers, and between co-design sessions, the project team revised the data collection tool as a next version.

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

- Labeling the concern or observation with a topic category (e.g., dataset type, history of data linkage, consent, IRB, data lifecycle)
- Noting specific proposals for change to the tool
- 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 tool during each co-design session, gaining additional feedback

2.5 Tool Completion

During co-design, the project team made significant changes to the Data Collection Tool, including the question order. Following co-design, the TEP and NICHD continued to review and provide feedback to the project team on questions, responses, instructions, and help text. Using this feedback, the project team finalized the data collection tool.

2.6 Documentation and User Guide

The project team developed documentation for the data collection tool, in the form of a code repository and a User Guide. The code repository includes basic documentation describing the software dependencies and steps for setting up the tool in a development environment as part of the project README file. The README also includes instructions for updating the version of the questionnaire used in the tool along with descriptions of some basic utilities, such as validating a questionnaire and listing out the contents of a questionnaire.

The project team developed a User Guide as a reference for potential users, in which the team explained that the data collection tool was designed to collect governance information about a dataset through a questionnaire and then transform questionnaire responses into structured metadata.

2.7 Usability Evaluation

The project team conducted a usability evaluation, with a small group of user testers to evaluate whether the governance metadata data collection tool was easy to understand and use and to gather feedback for future tool enhancements. For this evaluation, the project team designed a usability test guided by human-computer interaction methods.[26] During usability testing, testers representing the anticipated user population were observed entering real-world dataset governance information into the prototype. This allowed the project team to understand how the tool performs in a realistic setting. The MITRE IRB reviewed the procedures for this evaluation and deemed them exempt from human subjects' review.

The evaluation framework was guided by concepts from the RE-AIM framework and the UTAUT. These questions were asked as part of the usability evaluation sessions:

- Are the instructions in the prototype easy to understand?
- Is the terminology used to describe dataset governance understandable?
- Is the prototype organized in an intuitive way?
- Are the question prompts and response options comprehensive for gathering governance information?
- Does the prototype require substantial time and effort to use?
- Do users feel confident in the accuracy of the governance information they are providing?
- 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 four experienced biomedical researchers as user testers to participate in the usability evaluation (Table 2). NICHD ODSS selected researchers who were familiar with dataset linkage and who represent different domains of biomedical research.

Table 2: Experience of Participating Researchers in Usability Evaluation

Tester Number	Years Conducting Research	Biomedical Field of Research	Years Linking Datasets
1	23	Maternal and perinatal health	8
2	23	Behavioral health	4
3	20	Maternal health equity	1
4	15	Health economics	20

Usability Evaluation Orientation

The project team conducted an introductory virtual half-hour group meeting to introduce the user testers to the project and the governance metadata data collection tool. 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 data collection tool, and the expected sequence for the 1:1 user tester session. The orientation emphasized that the prepopulated response value sets in the data collection tool were driven by the three original use cases and 11 datasets in the foundational work and were intended to be illustrative, but not comprehensive.

At the end of the meeting, the project team provided user testers with the User Guide for optional review prior to the evaluation session, if needed. The project team also asked user testers to select a dataset that they had either generated or used for research to use for data entry during their evaluation session. The user testers did not view the data collection tool during this session.

Usability Evaluation 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 data collection tool via MITRE's hosted WARP environment to enter data into the questionnaire based on the governance information for their pre-selected dataset, starting at the first section and proceeding through each of the 11 sections. The facilitator instructed testers 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 tool. The facilitator requested that testers "save" their responses at the end of the questionnaire for further analysis. The facilitator prompted user testers to start each section of the data collection tool and then verbalize when they had completed the section or when they had determined that they could not

complete the section to gather timing information. Notetakers documented key observations to three questions for each of the 11 sections:

- Did the user tester fill out the entire section?
- [If "No"] Did the user tester stop early due to confusion or frustration?
- Did the user tester ask questions during the "think aloud?"

At the end of the session, the facilitator asked the user testers discussion questions about their experience entering governance information, perceived usefulness of the tool, perceived ease of use of the tool, the tool's ability to help the user tester perform their work, and whether they thought the tool would be adopted by researchers in their field. The facilitator asked which aspects of the tool aided data entry and what challenges the participant faced as they were entering information, and the facilitator elicited 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. See Appendix D for the session script and evaluation questions.

Analysis

After each usability evaluation session concluded, the project team conducted a 30-minute debrief, analyzed the session transcripts, notes, and discussion question responses, and then 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 based on concepts from RE-AIM and the UTAUT to support the analysis. 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 formed from the text as needed. Team members also applied structural codes to categorize data by the 11 questionnaire sections. After all the text was coded, three project team members met to review code applications for consistency and iteratively grouped code applications into overarching themes. See Appendix E for Usability Evaluation Analysis Codebook.

The project team estimated the time required to fill out each section and the overall questionnaire by calculating when the tester started and ended each section, and the project team counted the number of sections that user testers could not complete. Team members also reviewed the responses saved by the user testers to calculate a proportion of "I don't know," "It doesn't say," and blank responses to each question on the questionnaire.

2.8 Translation of Governance Information to Metadata Schema

Feedback on the early prototype from the co-designers lead to significant changes to the organization of the questions and responses in the tool. Although these changes greatly improved the user experience and ability to enter governance information, they also created a gap between the response values and schema values as they were no longer designed as a one-to-one match. Therefore, in order to test the

degree of alignment between the final data collection tool and the data governance metadata schema, the project team translated multiple responses (one synthetic and four real-world responses) from the data collection tool to the schema. As this proof-of-concept project was designed to explore collection of metadata for the schema, data collection tool responses should be able to be mapped to schema classes and represented by the schema vocabulary. The project team's methods for translation were:

1. Created a synthetic dataset entry with all possible options in the data collection tool selected for Section 1: User Information, Section 2: Dataset Information, Section 5: Consent, and Section 7: Data Use Agreement
2. Wrote a translation script that decomposes questionnaire responses from sections 1, 2, 5 and 7 into metadata elements, maps elements to schema classes and terms (referenced as *class::term* or *class::reference_standard:class)*, and applied mapping when loading metadata into a test relational database that reflects the schema specification.
3. Categorized each question response as:
    - Perfect Match: Decomposed into metadata and mapped to schema classes and terms accurately, which meant the information was completely loaded into a test database with accurate representation, of the meaning.
    - Imperfect Match: Decomposed into metadata and mapped to the schema classes with imperfect representation, which meant the information could be at least partially loaded into a test database but was imprecisely represented, e.g., specific details such as the name of an IRB or the approved purpose of use cannot be mapped to a schema class, or the meaning of the mapped schema value is an approximate but imperfect match for the response value, or
    - No Match: Decomposed into metadata and unable to map to a schema class, which meant the information could not be loaded into a test database at all.
4. Manually validated that metadata elements from the questionnaire response were loaded successfully into a test database for perfect and imperfect matches.
5. Synthesized question responses and metadata elements that could not be mapped/loaded into a test database.
6. Applied translation script to four user tester questionnaire responses reflecting real-world datasets.
7. Loaded metadata from four user tester questionnaire responses into an isolated test database for discussion and validation.

During the application of the script, the project team also noted schema classes and attributes for which no corresponding data collection tool question exists.

3 Outcomes and Findings

The project team generated findings based on the development of the data collection tool and the evaluation of the tool through usability testing.

3.1 Early Prototype

The project team designed the initial high-level question flow for collecting data governance information with six sections (User Information, Dataset Information & History of Linkage, Common Federal Laws, Policy Information, Process Information, and Other Governance Information; Figure 3).

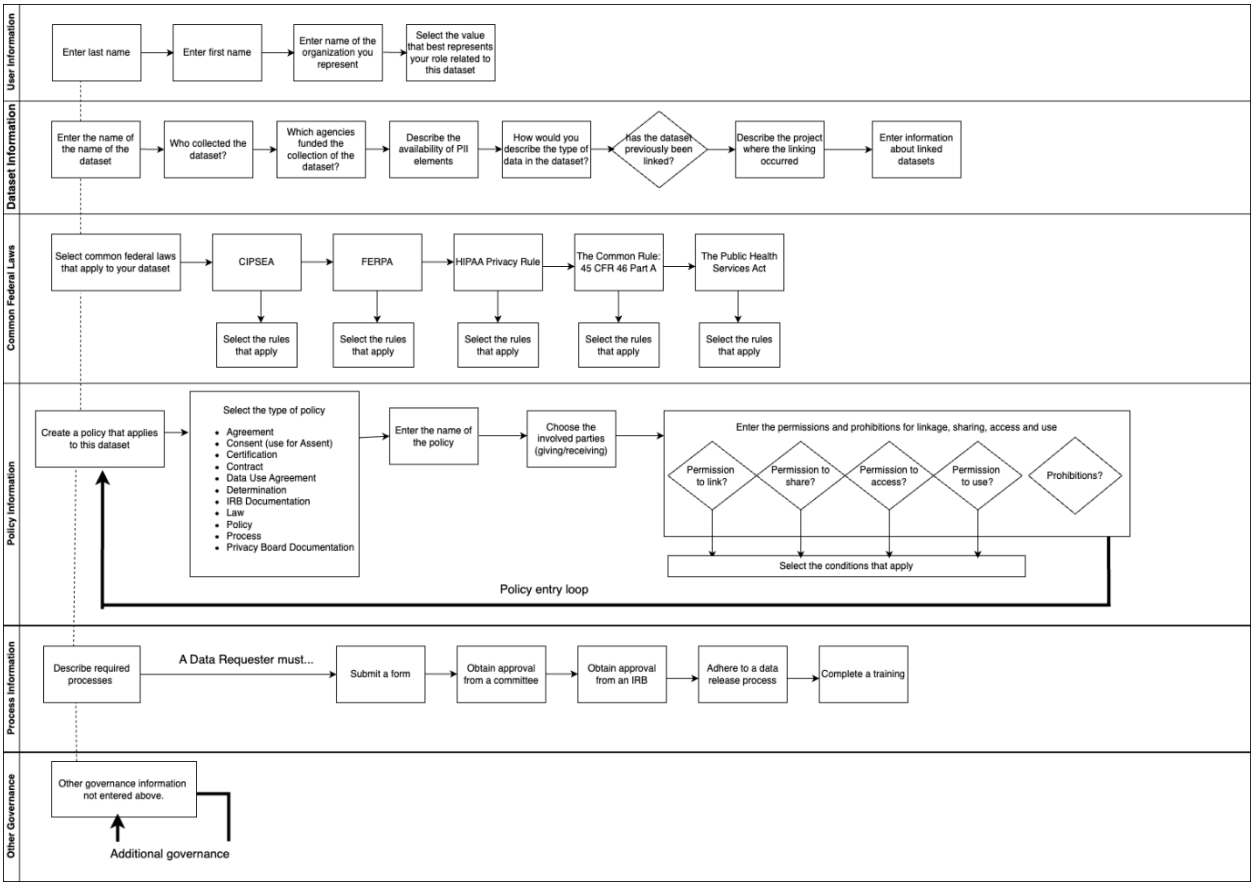


Figure 3: High-Level Question Flow, Early Prototype

The choice and order of these sections for the high-level question flow reflected the project team's goal of alignment to the metadata schema as well as initial assumptions of how users would approach entering data governance information. For example, beginning with dataset information and the history of linkage, users would consider common federal laws that apply to their dataset, and then consider multiple policy types. The Policy Information section was a flexible section enabling the user to enter multiple and different types of policies, naming each and then categorizing governance information by each policy type (e.g., consent, data use agreement). Within each policy, the user was presented with a series of questions about permissions for linkage, sharing, access, and use, and then a question about prohibitions on the dataset. Within each permission, the user could select a response of ***Yes, with***

conditions and then select one or more prepopulated conditions from a dropdown menu or enter a custom condition as free text. A looping feature allowed users to enter multiple policy entries. Conditional logic was configured to hide questions that were irrelevant based on previous answers.

3.2 Co-design

Refer to Figure 4 for the Version 1 (v1) of the tool, which was initially presented to the co-designers and subsequently significantly changed throughout the co-design process based on co-designer feedback.

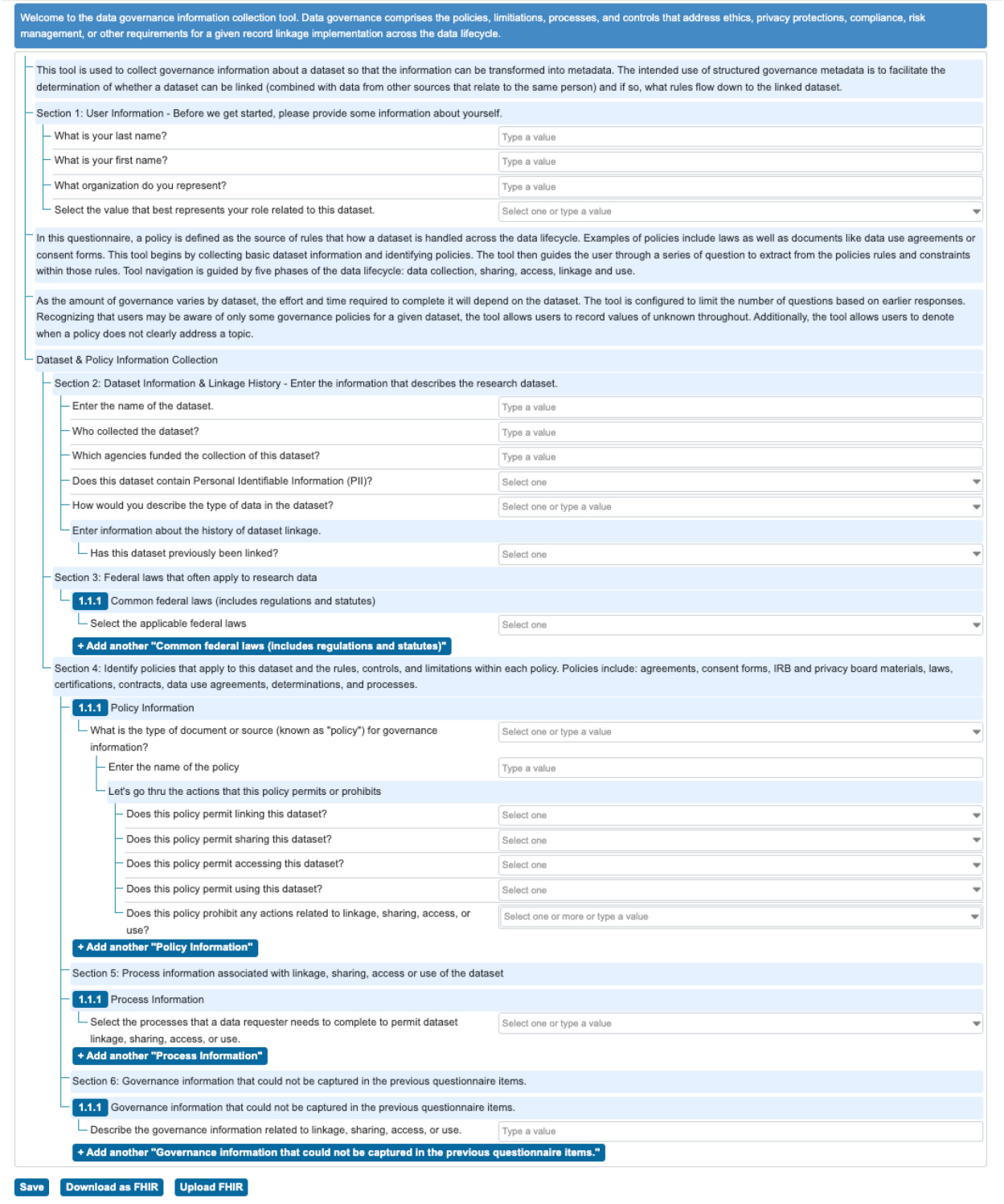


Figure 4: Co-design Session #1, Data Collection Tool, v1

During and between co-codesign session, the researchers offered over 80 comments, in the form of observations, concerns, and suggested changes with the goal of improving the user experience and

optimizing the questionnaire responses. Examples of suggested changes to the tool, based on co-designers' feedback, are included in Table 3.

Table 3: Examples of Suggested Changes to Data Collection Tool from Co-designers

Topic	Suggested Change	Implemented
Definitions	Define linkage early in questionnaire to clarify meaning.	Yes
User Experience	Make numbering more prominent. Add section numbers.	Yes
Dataset Type	Add response value for dataset type for data generated from biospecimens and from genomic studies.	Yes
Dataset Type	Add a question about dataset type including an option for deidentified with a hash token.	Partially
Dataset Type	Revise type of dataset question to allow multiple selections and update instructions to say: *select all that apply*	Yes
Dataset Information	Revise to state: *which organization collected the dataset*	Yes
Dataset Information	Add a question to collect grant number.	No
Dataset Information	Add field to capture IRB protocol number.	No
History of Linkage	Add a question about linkage methodology.	Yes
IRB	Ask questions about the IRB and permissions from the IRB at an earlier point in the questionnaire, perhaps after dataset information (before common federal laws).	Yes
IRB	Create a dedicated section for IRB and consent policy. Ask an early question about if participants were consented and then collect information there about the consent form and reconsenting.	Yes
IRB	Add field to capture the name of the IRB of record.	Yes
IRB	Add a question to capture if there is a single IRB or multiple IRBs.	Yes
IRB	Add a question to capture approved purpose.	No
Data Use Agreement	Create a DUA section; pull it out of the other policy section.	Yes
Policy	Add a question to collect a link to the form.	No
Permission Conditions	Add condition on permission to use dataset: *only for aggregate use*	No
Prohibitions	Take prohibitions out of permission constraints; instead, collect prohibitions in a separate question after permissions.	Yes
Prohibitions	Update permissions response values to separate *may not be linked* and *may not be used beyond explicit permissions*	Yes
Prohibitions	Add a prohibition for: *cannot share with commercial entities*	Yes

The project team recommended co-designer suggestions for implementation if the suggestion would improve a users' experience with the tool or lead to more robust responses and was aligned with the data governance metadata schema goals.

In summary, the data collection tool was updated in response to co-designer feedback through the following changes:

- Deconstructed the single policy section into separate sections dedicated to specific policy types.
- Reordered policy sections to align with the order suggested from practical research experience, e.g., start the questionnaire with IRB, Consent, then Privacy Board, and then DUA sections.
- Added genomic data, patient reported data, and data generated from biospecimens to types of data in the dataset.
- Revised the sequence of questions in the sections titled Availability of Identifiers Needed for Dataset Linkage and History of Dataset Linkage.
- Moved the questions about laws from early in the question section sequence to later in the sequence.
- Added a "save" feature to allow users to complete governance information entry over multiple engagements with the tool.
- Implemented a "transfer" functionality to allow multiple individuals to collaborate on a data governance information entry.
- Re-labeled dataset access and use as secondary throughout the tool.

Suggestions that gathered more detailed dataset information or specifics about a governance topic that extend beyond the schema's capabilities (e.g., IRB protocol number) were considered but not implemented. Some suggestions specific to a niche domain of research, targeting a different user, not possible using the FHIR SDC standard, or beyond the scope of the proof-of-concept project such as adding biospecimens to data types and material transfer agreement were not implemented. Other suggestions were not aligned to the definition of linkage that the schema adopts, such as adding mother-baby linkage into the history of linkage as a type of linkage. Because a dataset can be defined in multiple ways, co-designers recommended changing the unit of data collection from a dataset to an IRB protocol. This suggestion was also not implemented because it was not aligned with the parameters of the schema.

Broadly, the updates to the prototype suggested by the co-designers helped make the tool easier to use and more effective at capturing governance information from the users; however, the iterations during this phase also resulted in a questionnaire that no longer generated responses that all exactly aligned to schema classes. In order to assess the impact of this outcome, the project team conducted the translation exercise, the findings from which are described in section 3.6.

3.3 Tool Completion

The project team iterated on versions of the data collection tool, making updates through the LHC FHIR SDC Form Builder, downloading the JSON file, and loading the JSON file onto the WARP server. Twenty

releases were made over four months with version 2.8.6 as the final data collection tool release containing 165 questions organized in 11 sections. See Appendix E for the final 165 questions.

The 20 releases reflect tuning question wording, making formatting, value set order, and capitalization consistent, and revising the order of questions and associated business logic. For example, a ***YES/NO*** question was added at the start of sections 3–10 so that the section is collapsed at the outset and only presents questions about that topic when a user affirms (i.e., answers ***YES***) that the dataset has governance on that topic.

Figure 5 displays a screen capture of version 2.8.6 of the governance data collection tool with the sections collapsed and no responses populated. As users enter responses, the questionnaire expands to display other questions that are relevant based on that response. Question marks indicate where "help" text is available.

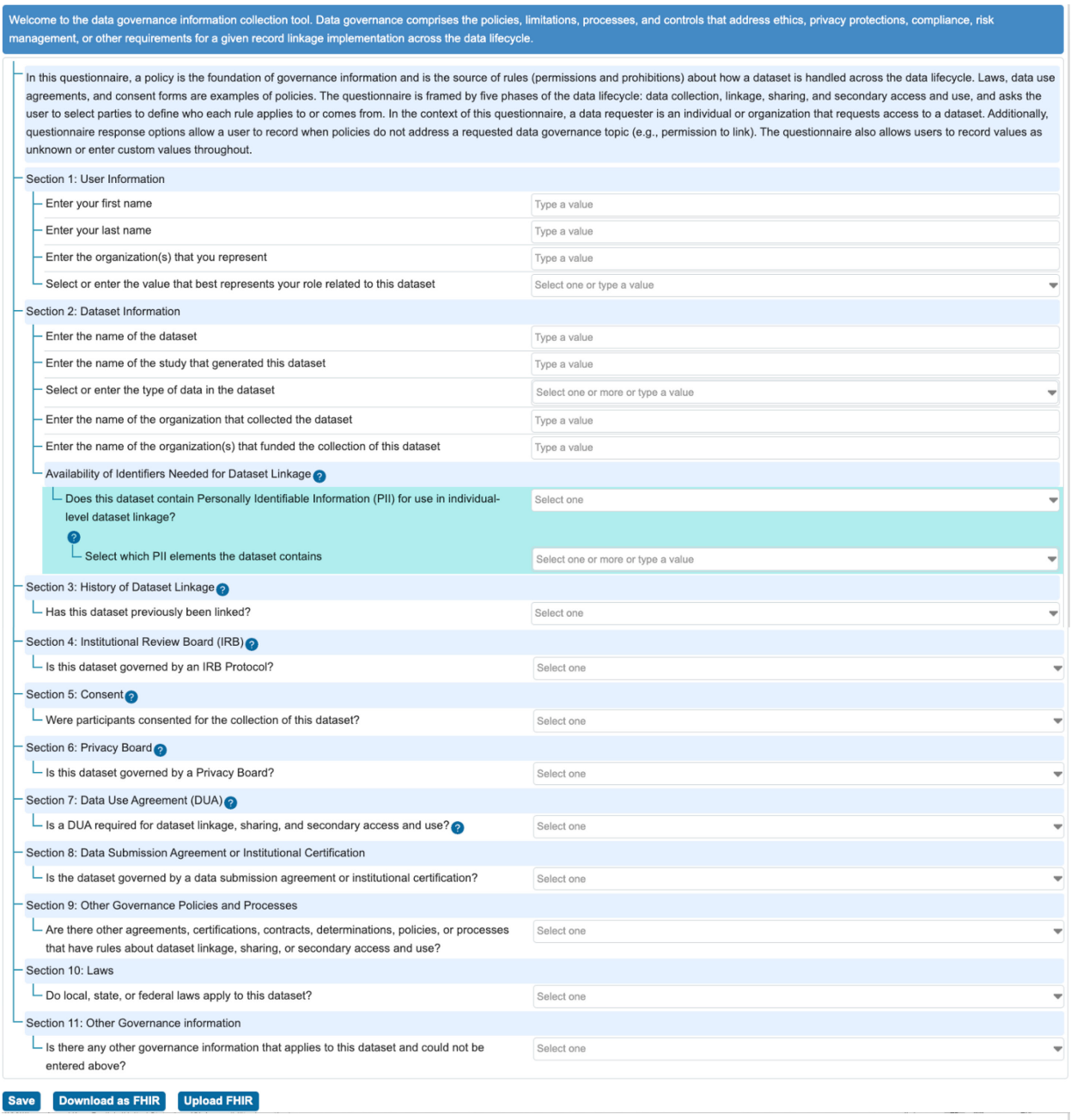


Figure 5: Governance Data Collection Tool, v2.8.6

The final questionnaire is organized in 11 sections:

1. User Information
2. Dataset Information
3. History of Dataset Linkage
4. Institutional Review Board
5. Consent

6. Privacy Board
7. Data Use Agreement
8. Data Submission Agreement or Institutional Certification
9. Other Governance Policies and Processes
10. Laws
11. Other Governance Information

Section 1 collects information about the individual entering governance information, including the user's first and last name, organization, and role related to the dataset.

Section 2 collects the name of the dataset, the study that generated the dataset, the type of data in the dataset, the name of the organization that collected/funded the dataset, and the availability of identifiers needed for dataset linkage.

Section 3 collects information about the history of dataset linkage, including if the dataset has been previously linked and the linkage method. Historical linkage projects can offer insights on the governance rules that may apply to future linkage projects with that dataset. History of linkage is placed after dataset information as a conceptual extension of dataset information with the goal of prompting the user to call up governance knowledge from previous linkages that could be relevant to the following policy sections.

Sections 4–10 are intended to help the user enter policies that apply to the dataset, parties for that policy, rules within each policy, and conditions for each rule. The questionnaire creates policies and rules by asking users whether common policies exist, and when they do, asking if rules about linkage, sharing, access, and use exist within each policy. Sections about the IRB and Consent are placed first because those are the most common sources of governance policies and rules and are considered to be the strongest governing forces in a research study (e.g., researchers look to the IRB to tell them what to do).

Sections 4–8 have a nearly identical question pattern. The section opens with a ***Yes/No*** question about if the dataset is governed by that type of policy or party. If ***No***, the user moves on to the next section. If ***Yes***, the questionnaire asks the user about permissions to link, share, access, and use the dataset based of the policy or party with the options of ***Yes, Yes with conditions, No, I don't know,*** and ***It doesn't say***. For responses of ***Yes with conditions***, the user is prompted to select or enter conditions from a prepopulated list. Finally, the user selects or enters prohibitions from a prepopulated list. Sections 4–8 include sporadic section-specific questions relevant only to that topic. For example, the Consent section includes the question, "*Will Minors in the dataset be reconsented?"* Section 9 uses the same question pattern but asks the user to first select a type for the policy they are entering.

The questionnaire asks the user to select parties from a curated list of research-relevant generic values for Data Use Agreement, Data Submission Agreements, Institutional Certification, and Other Governance Policies and Processes, when applicable. For Consent, IRB, and Privacy Board sections, parties are not collected and instead are assumed based on the section.

Section 10 addresses laws in two ways: allowing the user to select relevant federal laws and helping the user to enter policies for other relevant laws. Six common federal laws with rules about research and dataset linkage, sharing, access, and use are prepopulated into response sets that allow the user to select laws that apply and then select the rules from each law that are relevant.

Section 11 requests any additional governance that could not be entered in the previous sections with a free-text response. In total, the data collection tool holds 165 questions.

Knowledge of governance information can be fragmented across multiple individuals at a single institution and across institutions. The tool allows users to "save" data entries for later completion, and/or "transfer" a questionnaire to others for review or completion.

The data collection tool, as a proof-of-concept project, was not intended for long-term use or hosting by any organization after the completion of the project. However, it could serve as the basis for the development of future governance information collection tools and efforts. The project team generated source code and documentation and posted those materials to GitHub so that interested stakeholders could deploy and modify this tool. The project team posted data collection tool materials to GitHub as a subdirectory[e] within the Data-Linkage-Governance repository. The materials are shared under a permissive Apache 2 license to foster future adoption and adaptation.

3.4 Documentation and User Guide

The Guide assumes that the tool is easy to navigate, and the instructions and help text are intuitive and require little additional explanation. Therefore, the project team included these sections in the guide's contents:

- Overall approach to the questionnaire, including the common researcher user story, user roles, navigation framed by the data lifecycle, question structure and sequence, and general guidance on response types
- Terms and definitions (refer to Section 7 of this document)
- Guidance for each of the 11 sections of the questionnaire
- Resources and technical assistance, including for navigating the questionnaire, transferring a questionnaire response, exporting and importing questionnaire response data, and navigating multiple datasets

3.5 Usability Evaluation

The project team conducted four 90-minute usability evaluation sessions with researchers as user testers in July 2024.

Qualitative analyses identified **eight major themes** from the user testers' feedback and observations, as well as a list of benefits and challenges related to the use of the LHC Form Builder implementation of the FHIR SDC standard. Three of the themes align with concepts in the UTAUT model, including Effort

[e] https://github.com/NIH-NICHD/Data-Linkage-Governance/tree/main/MetadataCollectionTool

Expectancy, Performance Expectancy, and Social Influence.[27] The numbers in parentheses in the subsequent sections indicate which user tester(s) discussed each given theme.

1. Tool is easy to navigate, but the task of entering governance information is difficult (Effort Expectancy)

The UTAUT model defines effort expectancy as the degree of ease associated with consumers' use of technology. All user testers verbalized that the data collection tool was well organized and easy to navigate, providing feedback that it was "*very easy to click through*" (2) and "*straightforward*" (3). All testers, however, also thought that the tool would require deep background knowledge to fill it out correctly, with one tester noting "*my knowledge is the barrier, not the tool itself*" (1). One tester suggested that the effort required to fill in the questionnaire would vary substantially depending on the complexity of a dataset's governance and estimated that documenting governance for one of their standard research datasets would require only 10 minutes (3). Across the four usability evaluation sessions, the time required to complete the questionnaire ranged from 12 to 40 minutes.

2. Tool could be fundamental to supporting future linkage implementations, but governance information may be too complex to document it comprehensively (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. User testers recognized the data collection tool's ability to pull together disparate governance information for a dataset (1, 2, 3), and all thought it would be important to have this governance information before new linkage implementations are initiated. Testers also thought it would aid governance discussions with collaborators and IRBs (1), and that it could help to standardize how governance data are collected and represented across their field of research (2, 4). One tester noted that the tool could help with exploring published data:

> *"That's something where I think this tool could potentially help a lot, every time there is a publication, just having this tool required right on the dataset attached to the publication." (4)*

Despite the positive appraisal of the tool's usefulness, all testers thought that the tool may not be able to document all governance information comprehensively due to the overall complexity of governance:

> *"It has to be incredibly difficult to make a tool like this because every research project is a different flavor of research ... it's hard to get something that's going to be applicable across the board." (2)*

One tester noted that governance for a dataset is not static and may change over time as the research team is faced with unanticipated challenges (1).

3. User testers' institutions would support use of a tool like this, but some researchers may only use a tool if required for funding or other research-related activities (Social Influence)

The UTAUT defines social influence as the extent to which consumers perceive that important others believe they should use a particular technology. User testers thought that the research institutions they are affiliated with value data sharing and would support their use of this data collection tool (1, 2, 3) and

anticipated that government agencies may encourage or require the documentation of governance information in the future (2, 3). Additionally, testers anticipated how other researchers in their field would use the tool and suggested that researchers would only use the tool if required for funding or other research-related activities. Two testers thought that principal investigators (PIs) or others involved in research governance such as IRBs may default to choosing the most restrictive sharing and linkage options to mitigate potential risk of sharing data either due to an overabundance of caution or a lack of governance knowledge or to decrease time spent filling out the questionnaire (1, 2). One tester noted, "*researchers get pretty good at figuring out what is a check box that I just have to go through so I can get on with my day*" (2). As a result, this tester anticipated that "*you could end up with a dataset being labeled as much more restrictive than in reality*" (2).

4. User testers found it difficult to know the source of a rule, recognizing that a given source of governance information may base its rules on another source leading to duplicate rules

User testers noted cases where rules established in one type of policy are based on rules established in another type of policy and potential confusion over where to document rules when they overlap. Examples include rules established by an IRB being reflected in consent forms approved by that IRB (1, 2), IRB policy reflecting rules established by the Common Rule (1, 2), and DUA forms reflecting rules established by the IRB, privacy board, or consent forms (2, 3, 4). One tester noted that their privacy board was embedded within their IRB (2). Testers noted that governance policies could be defined by a number of rule-making entities and many different parties, and it could be difficult to determine where a rule originated from (2, 3). In cases where governance was interrelated, testers tended to document the overlapping rules in all sections where the rules were reflected to ensure that their responses were comprehensive (2, 4). This led to one tester feeling like they were filling out the tool incorrectly due to the repetition:

> *"It was tough for me, I didn't want to miss something, but I ended up just giving you the same answers again and again because I already answered that under IRB." (2)*

Two testers suggested that the tool should have to ability to copy governance information from one section to another when rules overlapped (2, 4). One tester suggested that the section that gathers information about applicable laws should be earlier in the questionnaire as many policies are based on those laws (2).

5. Researchers often do not have the authority to accurately interpret all policies and laws, and approval and guidance by an authoritative party is needed

User testers, as researchers, noted that a study's principal investigator is often not an authority on interpreting or documenting all of the governance information for a dataset, and that a governance authority should either enter or review governance information (1, 2, 3). All testers noted limitations to their own governance knowledge, suggesting that data will only be accurate if the "*right*" person fills out the questionnaire (2, 3, 4). Two testers thought that the IRB would be the best authority to review governance information (1, 2), though the ultimate authority on dataset governance may differ in cases where a dataset is generated or used by a research network, consortium, or government body (3). Two

testers suggested that the tool should have a way to pass governance information entry to another party for either data entry or review (2, 3), an existing feature in the tool that some user testers missed.

6. Governance rules depend on the context of sharing and whether a full dataset with PII or a deidentified subset is shared

User testers thought that the governance information about dataset sharing would depend on what was being shared and who the data would be shared with. Researchers often share only a deidentified or limited version of their dataset and not the full dataset with all of the gathered personally identifiable information (PII) (1, 2, 4). Sharing a subset of a dataset requires different governance than sharing the full dataset:

> *"Sharing of the deidentified dataset falls under a different category than sharing of a dataset with identifiers that could be used." (2)*

Depending on context, testers noted uncertainty about whether they should fill out the questionnaire from the perspective of sharing the full dataset, sharing a subset version, or whether their governance information entry should reflect both cases. For example, one tester was unsure of how to answer questions related to the PII contained in the dataset (2). Another tester attempted to provide answers related to multiple versions of their dataset but thought that they would be providing conflicting information if they simultaneously provided answers such as "*may only be shared as a de-identified dataset*" and "*may only be shared as a limited dataset*" for a question (3). On reflection, one tester thought that filling out the questionnaire for a deidentified version of their dataset was not aligned with the purpose of the data collection tool, since deidentified data cannot be used "*to make a linkable dataset*" (2).

Testers also noted that governance information about sharing would change based on who is requesting the data, such as whether the requester is a funding organization, government entity, or a member of a research network or consortium (1, 2, 3).

7. Certain instructions were too broad to effectively guide data entry

User testers discussed sources of governance information that they did not think were addressed directly by the sections defined in the data collection tool. These included tribal law and governance information about data for indigenous peoples (2), requirements for specific institutional affiliations such as the Veteran's Health Administration (1), community advisory boards (1), data policies for federally qualified health centers (1), use by researchers from countries of concern (1), international law (1), and policies and rules dictated by funding sources (1). While the final section in the tool (Section 11: Other Governance Information) is purposefully broad to collect data on less common forms of governance, such as those identified by the testers, testers did not enter information into this section. Testers thought that the wording of this section was too broad to provide effective guidance on what types of governance information should be entered. Testers suggested that more prompts be added to directly ask about anticipated forms of governance information that are not represented in the prior sections (2, 4).

Additionally, many questions throughout the data collection tool give users the instruction to *select or enter a value,* yet some user testers did not enter a value even when the prepopulated values did not

meet their needs. It may be that they did not notice this option due to the visual formatting of the LHC Form Builder or that the instructions did not highlight this option sufficiently.

8. Effective use of the tool requires knowledge of dataset governance and understanding of core governance concepts

User tester feedback and observations indicated that effective use of the data collection tool requires deep knowledge of both the governance of the specific dataset one is providing information about and core governance concepts in general. Some testers suggested that effective use would require more governance knowledge than the average PI would have.

Testers noted that PIs would more effectively answer questions about forms of governance they interact with regularly, such as the IRB and consent forms (2, 3). Two suggested that they would need to reference outside resources such as grant applications or executed DUAs to provide accurate information (1, 2). All testers reported that the section on laws was the most challenging to fill out. While testers often noted their lack of knowledge of certain governance topics, they rarely used the ***I don't know*** response option to answer governance questions. This pattern may indicate that users may feel pressured to provide definitive responses even when they are unsure of the accuracy of their answers.

Testers also noted known gaps in the governance information of their datasets that could lead to blank answers in the data collection tool. Testers' datasets did not have rules about permission to link with other datasets, prohibitions against selling datasets, and primary researcher control over the use of data once shared (2). Two testers noted that their institutions did not have template DUAs that they could draw on to answer questions, indicating that a new DUA would be established for every instance of sharing and linkage (2, 4).

Feedback and observations also indicated that testers struggled with the definitions of core governance concepts underlying the metadata schema, and general language used throughout the tool. Testers noted a lack of knowledge on the definitions or wording about core governance concepts such as institutional certification (1, 3), PPRL (3), the HIPAA (Health Insurance Portability and Accountability Act) Safe Harbor de-identification method (3), data enclaves (4), or the concept of participant reidentification versus recontacting (2). Testers struggled with core governance concepts, such as the dataset lifecycle. One tester wondered whether linking data at a geographic level counted as a "linkage" for the purposes of the tool (3). When providing information on the types of data in the dataset, one tester was unsure of the definitions of "Administrative" and "Electronic Health Record" data (2). Another tester noted that the questions were geared toward sharing governance information about quantitative research datasets and wondered whether the options would be the same for qualitative datasets (4). Two testers suggested that more information buttons for terminology or a glossary would help (1, 3). All testers appreciated the preset dropdown options that helped them answer questions using a standard set of response options.

Testers struggled with disentangling the concepts of sharing, secondary access, and secondary use (2, 3), which may have been related to their perspective as a PI. Based on their own role in the research process, one tester described not thinking of those parts of the lifecycle as being separate:

> *"I also had a hard time understanding the difference between sharing and reuse ... I guess from an administrative perspective, our opinion was always that once the dataset is out of my control, I can't control what somebody else does with it." (2)*

Testers also struggled with the perspective the user should take when filling out the questionnaire. One tester noted:

> *"PI's wear a lot of different hats and so just tell them which hat they're putting on so that they can give you the right answer – that would be helpful." (2)*

Testers suggested that the tool needs clearer guidance and education on the perspective or role that the user should inhabit when filling out the tool, such as when the tool should be used, the purpose of collecting the requested information, and what happens to the data that are entered (2, 3, 4).

Benefits and challenges of using the LHC implementation of the FHIR SDC standard

User testers provided insight into benefits and challenges in using the LHC Form Builder implementation of the FHIR SDC standard to create the questionnaire for the data collection tool. Testers appreciated that the questionnaire was all on one page, which provided an indication of its overall length and allowed the testers to easily see all sections and review prior answers (1, 2). One tester noted that there were not any required responses that would block progress in filling out the questionnaire if some sections were left blank (2).

Testers appreciated the ability to enter free-text answers when previously defined dropdown options did not adequately describe their dataset's governance (2, 3). However, this free-text option did lead to cases where the user did not see all the options in the dropdown list, and ended up entering free-text data that would have been better represented by a dropdown response (3, 4).

Some testers appreciated the business logic that collapsed most response options unless they were relevant, stating that this made the tool "*streamlined*" (3). However, users would also regularly choose different responses to questions to see all potential response options (1, 2, 4). For example, when selecting whether a source of governance allows dataset sharing or linkage, users would select the ***Yes, with conditions*** option to review potential responses before providing a final answer. In some cases, reviewing alternate responses prompted testers to change their answers, indicating that the responses helped to contextualize the questions:

> *"I think the dropdowns were great, not only because the information is right there and I didn't have to type it or figure out how to word it, but it also helped me understand what the question was looking for." (2)*

A drawback of using the LHC implementation of the FHIR SDC standard was that testers did not always understand based on the user interface when they could select multiple responses or enter free-text answers (1, 4).

3.6 Translation of Governance Information

The project team translated and loaded samples (sections 1, 2, 5 and 7 only) from five data collection tool responses into a test relational database to examine alignment between the responses generated from the data collection tool and the data governance metadata schema.

Findings from sections 1, 2, 5, and 7 are described below, with questions from the questionnaire referenced in italics.

Note that because Sections 4–10 feature a consistent and repeating design, the translation exercise performed on sections 5 and 7 would identify many issues that may have been identified in sections 4, 6, 8, 9, and 10.

Section 1: User Information and Section 2: Dataset Information

- All four user information questions (user first and last name, organization, and role) were mapped to *policy:creator*, which exists in an array and can hold multiple values and be mapped to the schema classes with imperfect representation (Figure 5).
    - Per ODRL, "creator" is meant to represent the "individual, agent, or organization that authored the Policy"[28], not the person providing the information, for example in a form. Therefore, a new schema class may be needed to more intentionally model the person entering the governance, the organization they represent, and their role.

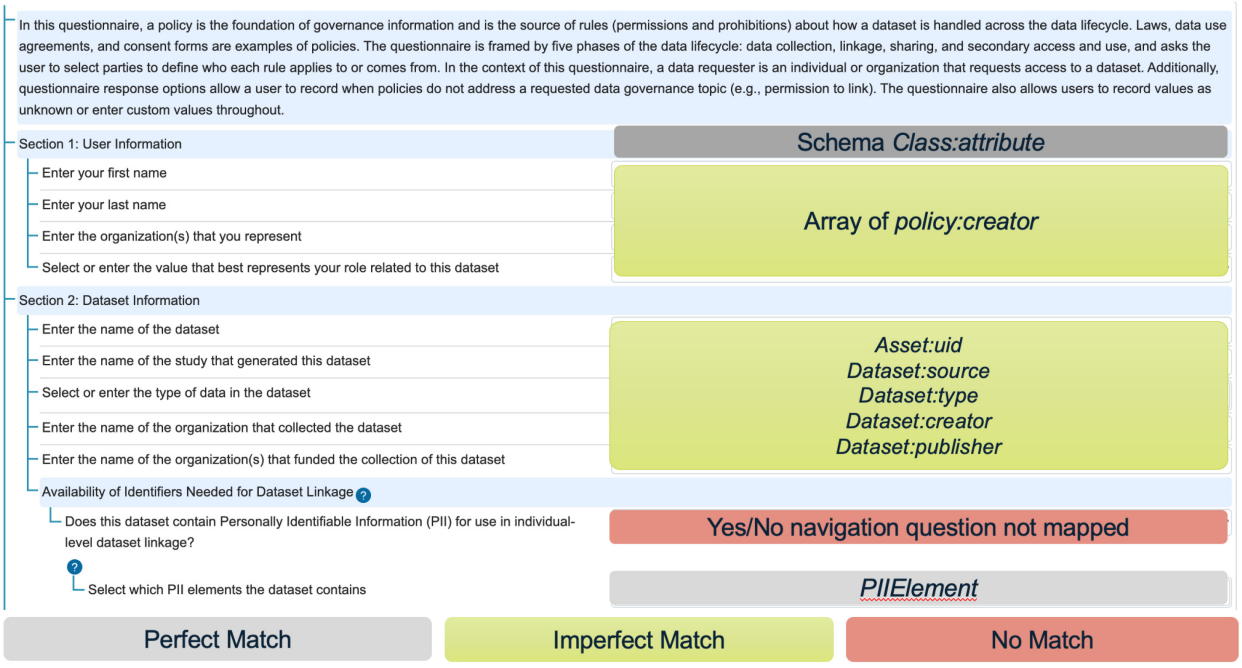


Figure 5: Translation of User Information and Dataset Information

- Dataset name, source, type, originating study, collecting organization, and funding organization can be mapped to asset and dataset classes in the schema with imperfect representation (Figure 5).

- Dataset name is mapped to *Asset::uid.*
- Name of the study that generated the dataset is mapped to *Dataset::dc:source* and the type of data in the dataset is mapped to *Dataset::dc:type.*
- Name of the organization that collected the dataset is mapped to *Dataset::dc:creator* and name of the organization that funded the dataset is mapped to *Dataset::dc:publisher*
- The mapping of source, creator, and publisher are imprecise. Future users of the schema could benefit from more detailed guidance on which data should be mapped to these fields.

- Responses to *"Are identifiers accessible outside of the dataset to generate a pseudo-identifier (e.g., hash or token)?"* were not able to be mapped to a corresponding schema class.
- Responses to "*Does this dataset contain personally identifiable information (PII) for use in individual-level dataset linkage?"* were not able to be mapped to a corresponding schema class.
  - A ***Yes*** response was not able to be mapped to a corresponding schema class because the schema class *PIIElement* stores specific PII elements that are available in the dataset rather than the presence of any PII in the dataset.
  - Responses of ***No*** or ***I don't know*** were not able to be mapped to a corresponding schema class. Per the schema's adoption of the Open World Assumption which indicates the schema can only represent explicit information and not missing information, the schema cannot represent the lack of PII elements or lack of knowledge about the presence of PII elements.
  - A lack of PII elements could indicate that this dataset is a deidentified dataset and such classification could be useful to capture. These responses hold meaningful governance knowledge that could be added to dataset information.
  - Responses to *"Enter the organization that holds these PII elements"* were not able to be mapped to a corresponding schema class.
  - PII elements present in a dataset are mapped to various *PIIElement* terms in the schema with accurate representation.

Section 5: Consent

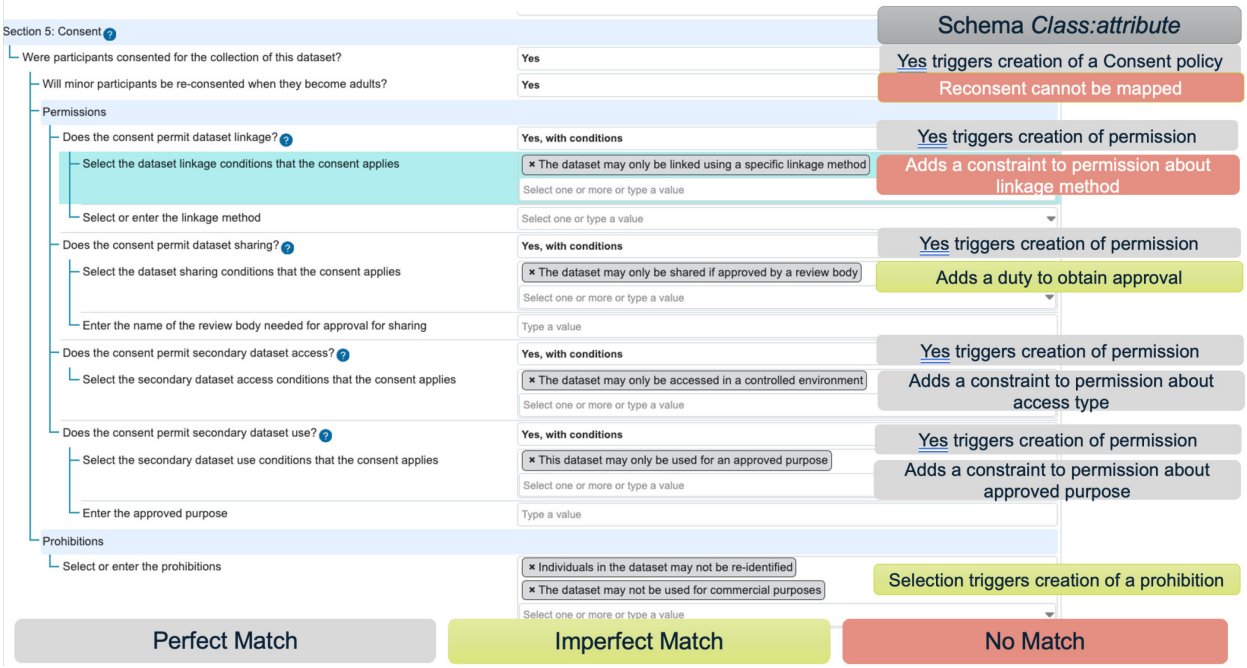


Figure 6: Translation of Consent

- Figure 6 shows findings of the translation exercise of consent.
- Responses of ***Yes*** to "*Were participants consented for the collection of this dataset?*" can be mapped to attribute classes in the schema with accurate representation.
    - A ***Yes*** response triggers the creation of a *policy* of policy type=consent.
    - Responses of a ***No human participants in the dataset***, ***No,*** or ***I don't know*** are not able to be mapped to a corresponding schema class.
    - A waiver of consent requirement or no human participants is meaningful governance knowledge that could be added to the schema.
- The consent section does not collect a consent name, which is desired for naming policies in the schema.
- The translation exercise highlighted that the data collection tool does not accommodate multiple consent forms, which is a plausible research situation for studies that include both minors and adults or for multi-site studies.
- The consent section does not collect information about the consenting and consented parties.
- Responses to "*Will minor participants be re-consented when they become adults?*" are not able to be mapped to a corresponding schema class.
    - The schema does not have an *action* term for reconsent.
    - If the reconsent action were added to the schema vocabulary, this concept could be created as a rule of a type of obligation to reconsent assigned to the principal investigator (party).

- Responses of ***Yes*** to "*Does the consent permit dataset linkage? sharing? access? use?"* can be mapped to attribute classes in the schema with accurate representation.
    - The ***Yes*** and ***Yes, with conditions*** responses trigger the creation of a rule of type=permission.
    - The ***No, I don't know,*** and ***It doesn't say*** responses are not mapped to schema classes because they do not contain governance metadata about an existing rule.
- Thirteen conditions could be added to a permission to link, share, access, and use. When selected, conditions are annotated as constraints on a permission rule or as a duty that relies on the permission rule. Table 4 lists how each condition was translated, organized by dataset action.

Table 4: Translation of Conditions to Data Governance Metadata Schema

Action	Condition	Description of Translation
Link	The dataset may only be linked with the approval of the IRB of record	Permission to Link with a Duty to ObtainApproval assigned to the DataRequester by the IRB
Link	The dataset may only be linked with the approval of data contributing sites	Permission to Link with a Duty to ObtainApproval assigned to the DataRequester by the DataContributor
Link	The dataset may only be linked using a specific linkage method*	Permission to Link with a Constraint of LinkageMethod=ApprovedProtocol
Link	The dataset may only be linked for specific types of research or use*	Permission to Link with a Constraint of Purpose=ApprovedPurpose
Share	The dataset may only be shared as a de-identified dataset	Permission to Share with a Constraint of Output=DeidentifiedDataset
Share	The dataset may only be shared following the Safe Harbor de-identification method	Permission to Share with a Constraint of deidentificationMethod=SafeHarborMethod
Share	The dataset may only be shared if approved by a review body*	Permission to Share with a Duty to ObtainApproval assigned to a DataRequester by a ReviewCommittee
Share	The dataset may only be shared as a limited dataset	Permission to Share with a Constraint of Output=LimitedDataset
Share	The dataset may only be shared within a defined data release process*	**Imperfect match.** Could be imperfectly translated to a Permission to share with a Duty to ObtainApproval from a DataRepository.
Access	The dataset may only be accessed in a data enclave	Permission to Access with a Constraint of VirtualLocation=DataEnclave
Access	The dataset may only be accessed in a controlled environment	Permission to Access with a Constraint of AccessType=ControlledAccess
Use	This dataset may only be used for an approved purpose*	Permission to Use with a Constraint of Purpose=ApprovedPurpose

Action	Condition	Description of Translation
Use	This dataset may only be used for research on a specific topic	**Imperfect match.** Could be imperfectly translated to a Permission to Use with a Constraint of Purpose=ApprovedPurpose

* When selected, these five conditions trigger a follow-up question.

- Responses to condition follow up questions that are collected in a free text box are not able to be mapped to the schema classes. These include:
    - Select or enter the linkage method
    - Enter the specific type of research or use
    - Describe the data release process
    - Enter the name of the review body needed for approval for sharing
    - Enter the approved purpose
- Responses to "*Select or enter a prohibition*" can be mapped to attribute classes in the schema with imperfect representation.
    - User selections trigger the creation of a rule of type=prohibitions with the action terms reidentify, link, share, access, use, secondaryUse, CommercialUse, and sell.
    - A response of ***dataset may not be used beyond explicit permissions*** is not accurately represented with a prohibition. This concept could be represented as a prohibition against secondary use or represented as a permission to use with a constraint of purpose=approved purpose.

Section 7: Data Use Agreement

Translation exercise findings about questions that occur in a repeating pattern, in both sections 5 and 7, such as *Does the Data Use Agreement permit dataset linkage, …sharing, …secondary dataset access, …secondary dataset use*, etc. are presented in the previous section.

Other findings include:

- The name of the DUA is mapped to policy::uid and can be mapped to policy classes in the schema with accurate representation.
- The data collection tool does not collect the sources for policies. While the person entering governance information in the data collection tool could be conceptualized as the source for policies, the tool could be modified to store links to where the consent form, DUA, or policy document can be found for reference.
- Parties on the DUA (e.g., "*Select or enter the organization providing the dataset"* and "*Select or enter the organization receiving the dataset"*) were not able to be mapped to a corresponding schema class.
    - The schema accepts parties for rules but not policies.
    - The translation could assume that the parties for a DUA are the parties for the rules that the DUA holds, but this assumption requires further exploration.

The translation findings are the results of an experimental mapping to demonstrate how some responses to the governance information data collection tool questions could be fit into the schema. Even in its limited scope, translation highlighted multiple concepts where the schema vocabulary and structure were inadequate to completely and accurately capture real-world governance concepts.

4 Discussion

The Health FFRDC project team developed a proof-of-concept Governance Metadata Collection Tool for governance information, based on the governance metadata schema and using the LHC implementation of the FHIR SDC standard. NICHD ODSS, TEP members, and co-designers provided comprehensive review and feedback on questions, responses, and a question sequence that resulted in extensive iterations and 20 releases. The tool includes 165 questions presented in 11 sections, organized by governance policies relevant to research (e.g., consent, IRB, DUA, laws). Developing the right questions, responses, and business logic was the most challenging and resource intensive aspect of development, and these processes were critical to identify how the schema does or does not map to real-world data governance information.

The discussion section addresses the two aims of this proof-of-concept implementation project. The first aim is to explore how governance information may be collected from researchers and ascertain which governance information is the easiest and most challenging to collect. The second is to test how the data governance metadata schema structure and design perform in a real-world data collection setting. Discussion is organized by three lines of inquiry:

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

The technical development of the data collection tool using the LHC implementation of the FHIR SDC standard and the limitations of this effort as a whole are also discussed.

4.1 What are the questions to solicit governance metadata?

The project team's approach to soliciting governance information migrated from an ODRL-based approach to a research practice-based approach. The data governance metadata schema was the inspiration for the early prototype and in keeping with the schema, the tool initially asked users to enter policies and then enter the rules within those policies. However, feedback on the early prototype was poor. Co-designers thought that researchers would be unable to respond to open-ended questions about the policies and rules and suggested that governance information could most effectively be elicited through dedicated sections (e.g., IRB, consent, DUA, etc.) that mirror the practical hierarchy of their experience with research governance. The project team restructured and revised the data collection tool, iterating until co-designers felt the question content, order, and business logic was optimized. The final data collection tool contained seven sections to enter governance policies for each of the policy types: IRB, consent, privacy board, DUA, data submission agreement, other governance, and laws. Each policy section features a repeating five-question pattern:

- Does the <blank policy> permit dataset linkage?
- Does the <blank policy> permit dataset sharing?
- Does the <blank policy> permit secondary dataset access?
- Does the <blank policy> permit secondary dataset use?
- Select or enter the prohibitions [from <blank policy>].

Usability evaluation revealed that users were able to enter governance rules when asked about specific policy types that combine the who's (parties like the IRB and privacy board) and the what's (policy types such as consent and DUA) of research. However, creating distinct sections for important who's and what's of research is an approach that could be challenging to scale because many other parties and policy types exist in the broader research ecosystem.

User testing also demonstrated the "other governance" section for other policies, processes, agreements, certifications, and determinations was too broad. Users did not know how to enter governance in a section that could include a diversity of different types of governance, suggesting that even more dedicated sections for specific governance parties and policies may be required.

Creating multiple sections with a repeating question pattern also resulted in a longer questionnaire, adding over 50 questions to the tool for a total of 165 possible questions. However, each section starts with a ***Yes/No*** question about whether that governance policy applies (e.g., is this dataset governed by a Privacy Board?), and business logic and navigation show researchers only questions that apply based on their responses. If a ***Yes*** response is entered, that response triggers the option to enter a policy because ***Yes*** reports the existence of a policy. For example, a **Yes** to Question 5.13 "*Is a DUA required for dataset linkage, sharing, and secondary access and use?"* If the user enters ***No***, the section remains collapsed, and the user moves on to the next section. A downside of this approach is that ***No*** responses that report the absence of a policy do not record meaningful governance information and would need to be parsed out of the other governance information during metadata transformation. Even though the multi-section approach results in more total questions, the tool only presents the user with questions for policy topics that apply to their dataset. Though 165 questions are possible, it is unlikely any user would ever answer them all. User testers affirmed that the design and business logic made the questionnaire feel shorter and more manageable.

4.2 Can a researcher answer questions about governance metadata?

Yes, researchers can answer questions about governance metadata in this data collection tool. All four user testers successfully entered governance information for one dataset of their choice into the data collection tool but struggled with different sections and questions. Testers' confidence to answer governance questions was influenced by the challenge of clarity in governance terminology and varied levels of experience in linkage implementations.

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. User testers, TEP members, and co-designers affirmed that these terms are used widely across the research community, and yet terms such as linkage implementation mean different things to different groups. One user tester

interpreted data access and data sharing as the same while another was unclear if data sharing meant the same as data transfer. Despite the definitions provided in help text and the User Guide, unclear terms made entry into the data collection tool take longer and shook users' confidence in their own understanding of governance. While none of the User Testers reviewed the User Guide in advance of their usability evaluation session, it was developed for research teams to support governance information finding in advance of data entry including review of terms and definitions.

Collecting governance information at a dataset level raised questions about the meaning and concept of a dataset. Both the data governance metadata schema and the data collection tool capture governance information at a dataset level. The scope and boundaries to define a dataset are a challenge in practice, given one dataset can be a source for many other datasets. In research, a study is often the primary unit of analysis rather than a dataset. A single study may lead to many different datasets or versions of similar datasets. User testers, co-designers, and TEP members repeatedly questioned the definition and meaning of a dataset. As datasets are the unit at which linkage occurs, dataset level is the appropriate unit for governance information collection, but it is a challenging unit for researchers to conceptualize.

The data collection tool relies heavily on the policy type for its structure and question flow, and user understanding of policy type was mixed. Though some policy types are germane to specific types of biomedical research, some user testers were thrown when they encountered questions about determinations, institutional certifications, and data submission agreements. Unfamiliar policy types (e.g., foreign laws) shook multiple testers' confidence. A reliance on policy types for structuring the questionnaire could be additionally challenging as the name of a policy is not always an accurate indication of what policy type it is. For example, DUAs may be called data sharing agreements, data sharing and use agreements, network agreements, or master data agreements, and a researcher may not understand that an agreement functioning as a DUA with another name should be identified as a DUA in the tool.

Researchers struggled with how to represent the governance rules that differ based on whether a limited or deidentified dataset was being shared. In practice, many researchers have a fully identified dataset acting as a data source to generate a limited or deidentified dataset. This means the study's primary dataset can yield multiple different types of datasets for subsequent linkage and use. While some policies are the same, many rules governing a deidentified dataset differ from rules governing a limited dataset. Since the schema and tool treat the dataset as the unit of governance, researchers would potentially have to enter governance information more than once—separately for identifiable, limited, and deidentified datasets. Because multiple user testers grappled with this complexity, how the schema should represent rule variation for the same dataset is worth further exploration.

The order and sequence of the governance information questions in the data collection tool was essential to support successful data entry. The early version of the questionnaire began with questions about laws, a topic that co-designers and user testers shared they had the lowest confidence and highest level of concern about. Co-designers thought the law section should appear last and the tool should start with the IRB and consent as the primary authority about dataset rules. By starting with IRB and ending with law, the tool starts with the topics that researchers are the most knowledgeable and confident about and ends with the topic researchers are the least knowledgeable and confident about. User testers confirmed that researchers have the most familiarity with the rules from the IRB and

consent, the rules from IRB and consent carry the most weight in user testers' views, and consent and IRB were the sections where researchers were the most confident with their responses.

The schema is centered on two concepts: policies and the rules that those policies codify. User testers showed a strong command of the rules for a dataset and a weak orientation for the policies that those rules originate from. This gap could be problematic in the context of the repeated question pattern. For example, testers would often enter the same answer for "*Does the <blank policy> permit secondary dataset access?"* in the consent, IRB, privacy board, and DUA sections. However, permissions for secondary dataset access are a topic unlikely to be covered in a consent form. In that case, the tester is reporting the rules for the dataset but has forgotten which policy they are answering questions about or is not considering whether the consent form (as opposed to other policies) addresses secondary dataset access. This has the potential to create erroneous rules and inflate the number of rules when the governance information is queried or visualized. Further on this point, user testers struggled with the interrelatedness between governance policies. For example, testers questioned whether a section about laws was necessary, because it is the IRB's responsibility to ensure that research complies with relevant laws such as the Common Rule. Testers also identified other forms of governance that overlap with the IRB, such as consent forms approved by the IRB or DUAs informed by IRB policy. Future schema effort could consider the tradeoffs surrounding restricting rules to exist only within policies and whether the schema could or should allow rules to exist independent of policies or generate an unduplicated set of rules to be mapped to multiple policies.

Perhaps the greatest challenge for researchers to answer questions about governance metadata stems from researchers having to report governance for a hypothetical future research study that would involve linking their dataset. Testers struggled to step into a future study perspective. For datasets that have never been linked, researchers had to imagine what the rules would be. Some testers noted that their institution did not have a template DUA that they could use to answer questions about what rules a future DUA might contain. Because policies and rules can be formed on a case-by-case basis, a researcher's capacity to anticipate the governance policies and rules that could apply to a future linkage could result in an inaccurate metadata record that reports the wrong rules or an incomplete metadata record that is missing policies or rules.

4.3 Do the question responses generate metadata that fits within the schema? Do the value sets in the tool support governance metadata collection goals?

Some question responses generate metadata that can be accurately represented using the metadata schema. The policies and rules that originate from those policies fit in the schema attributes and terms.

The collected party information does not fit within the schema, as it was collected at the policy level rather than the rule level. Revising the data collection tool to collect parties at the rule level is a simple tool modification but could create more difficulty for the user as datasets have many more rules than policies. Before this change is made, additional user testing is recommended.

Most of the conditions on rules can be accurately represented as constraints or duties in the metadata schema. However, translating all conditions to constraints highlighted some conditions that can be better represented through the addition of terms to the schema vocabulary.

Testers had the option to enter free-text policies, rules, conditions, and parties when the governance they were describing could not be represented by the available value sets. The schema accepts only structured metadata so free-text responses cannot be directly mapped into schema classes. However, these custom values are an important part of the tool and schema's future evolution. Because the tool's value sets, policy types, and rules were populated from a sample of biomedical datasets, more widespread adoption of this tool will surface concepts from the broader research governance community. Free-text responses could help identify terms to add to the schema vocabulary and highlight other policy types and rules that the data collection tool does not currently represent.

The translation exercise highlighted several potential changes to the schema and key governance information that the data collection tool does not collect but should. The data collection tool should collect source information (e.g., links) for policies and policy names for consent and IRB. The data collection tool should add questions or features to encourage the collection of duties—actions a party is required to take. The data collection tool should collect raw policy language, when available. The data collection tool should consider whether and how to represent relationships between parties.

4.4 Technical development and the LHC implementation of the FHIR SDC standard

Building and evaluating a data collection tool highlighted the limitations of LHC Implementation of FHIR SDC standard that impacted user experience and the quality of responses. The project team built this tool in keeping with the FHIR SDC standard and considered how the experience would have been different if this pilot had selected REDCap as the tool for extension.

The LHC FHIR SDC Form Builder was a superb application for developing and revising the data collection tool. Updates and releases throughout the implementation were driven almost exclusively by changes to the question wording, order, response values, and conditional logic. Implementing these changes was easily facilitated by the LHC FHIR SDC Form Builder.

However, user testers desired more supportive navigation features than FHIR SDC could provide. Testers requested "HELP" text to define response options; because FHIR SDC only allows help text at the question level rather than the response/value level, explanations of response values could not be provided. Testers overlooked key navigation features that were visually subtle like the help text, looping function to create more entries, and wording about response formats (e.g., select one or more or type an entry). The project team was unable to alter the format of these tool elements to provide more obvious visual navigation cues. Co-designers and TEP members suggested that the tool should allow users to "SAVE" a questionnaire response and return to it later and "TRANSFER" a partially completed questionnaire to another party. These two features do not align with the FHIR SDC standard but were implemented.

4.5 Limitations

The project team did not test the exchange of governance metadata. FHIR SDC was selected in part with the knowledge that FHIR-based solutions are designed for exchange. While the ability to easily exchange metadata using FHIR SDC remains an important feature for potential future adoption of a metadata collection tool, during the project, the team ultimately focused on developing the data collection tool content and enabling metadata capture rather than implementing exchange.

The project team did not employ a comprehensive approach to searching for existing data collection tools to consider extension. There are thus some tools that could have been missed.

The project team engaged only four co-designers, which represents a small sample of biomedical research perspectives. The same limitation is true for user testers. These researchers were selected based on experience with NICHD and linkage. Thus, the usability evaluation did not capture the experience of a linkage-naïve researcher using this tool. It also did not capture the perspectives of institutional representatives, data repository stewards, policy and legal experts, and other community members who also play an important role in making data linkage possible. Engaging them as co-designers and user testers could have yielded different results.

The project team did not translate all sections in the data collection tool to the schema, rather only four sections were selected for the translation exercise to be largely representative of the governance information generated from the tool. However, the selected sections are broadly representative of the data collection tool response content. Future schema work could benefit from a thorough translation of multiple questionnaire responses.

5 Recommendations

Evaluating a governance information data collection tool generated recommendations for future data collection efforts and the data governance metadata schema.

5.1 For Governance Metadata Collection Tools

This project produced a prototype data collection tool not intended for production use. While the tool was successfully used by a limited number of user testers, this evaluation process highlighted ways that governance data collection tools can be refined. Usability evaluation demonstrated that governance information can be collected from researchers but requires an understanding of both governance and the mechanics of the data lifecycle to enter the information accurately. If a data collection tool were adopted in practice, policy or legal experts should play a role in entering and verifying governance information, rather than the responsibility falling to researchers alone. The project team recommends the following updates which could assist tool users across many roles.

Add help text for the response values. LHC FHIR SDC Form Builder only allowed for help text to be attached to a question, making it difficult to define terms or explain response values. Adding more help text and a separate glossary within the tool that covers all terms could alleviate this limitation.

Expand introductory text for orientation and provide support for new users. This content could emphasize the context of use, when the questionnaire should be filled out in the research process, and what will happen to the information once it is entered.

Develop additional policy-specific modules to prompt novices on potentially relevant rules and elicit more comprehensive governance information. Since user testers tended not to provide additional information when faced with open-ended questions, additional policy-specific modules could prompt governance novices to think of potentially relevant policies and provide more comprehensive governance information. Examples include policy sections for tribal laws, requirements from funders, or international laws. However, testers recognized that covering all forms of governance may not be possible and adding more policy-specific modules tunes the tool a specific audience and may make the tool more difficult to scale. Thus, as an alternative, examples could be provided in the introductory text to Section 11: Other Governance Information to prompt users.

Consider collecting policy names and parties systematically. Collect policy names and parties universally and refrain from hardcoding these elements based on policy types.

5.2 For the Data Governance Metadata Schema

Developing and testing the data collection tool highlighted ways that the data governance metadata schema can be improved.

Expand governance vocabulary to add new terms for specific actions, policy types, parties, and constraints that were highlighted during the translation exercise and user testing. Specific vocabulary suggestions include adding honest broker and network to parties; disclose, cede, and reconsent to actions; and PPRL to constraint.

Allow parties to be assigned to policies (in addition to rules). This change will allow users to accurately record all parties on consents, agreements, contracts, and IRB determinations as well as note relationships between parties, such as an agreement between three parties or multiple IRBs ceding to a lead IRB. This change could also allow for assigning more than two parties to policies or rules. This would be useful given that the act of data sharing often encompasses multiple entities (e.g., three parties on an agreement).

Consider refining how the schema can show relationships between policies. User testers articulated how many policies are related and flow into each other and those relationships are not well documented by the schema. However, it may not be feasible to codify these relationships as its often difficult to confirm whether a rule exists because it comes from another policy, or whether the rule stands alone. Also, when the raw text of a rule is not written in exactly the same way, it could be difficult to determine whether the rules are the same or merely similar. Regardless, despite the potential for duplication, it is important to understand all the policies that rules derive from so users can go back and more deeply evaluate their meaning.

Add a schema class to record the order of rules that involve duties. Some rules require actions to be taken (duties), for example obtaining IRB approval and signing a Data Use Agreement. Defining an order for these procedures would help researchers acting on governance information sequence actions that are required across the data lifecycle.

Ensure that all schema terms from the profile and imported from ODRL have accurate governance-relevant definitions. The schema endeavors to use existing ODRL terms whenever applicable rather than creating similar terms. However, many definitions for ODRL terms are not governance relevant. For example, ODRL defines the action term *execute* as "to run the computer program Asset" and in data governance, execute is often referencing making an agreement legally valid and binding. The schema should consider a systematic approach to addressing ODRL definitions that are not accurate for a data governance application.

Consider if specificity of organization, IRB, and committee names would yield more useful governance metadata. And if so, the schema can add structures to store these specific values. Users did not find selecting parties from a dropdown menu to be intuitive, especially compared to party-related questions that accepted free text responses where users could enter an organization-specific name. Users verbalized multiple instances where multiple party values were applicable to a single institution (e.g., a government organization that is also the data repository) and that made it difficult to select one value that was the most accurate. A single institution often fulfills multiple different party roles within a linkage implementation and at different levels (e.g., one organization may encompass many boards, committees, IRBs, and PIs). If the purpose of parties is to articulate who a rule policy or rule originates from and applies to, a schema class could be potentially added to capture more specific party names (e.g., Johns Hopkins Institutional Review Board) and then map names to one- or multiple-party roles (e.g. Johns Hopkins Institutional Review Board is an IRB and a Privacy Board).

Add specifications to schema documentation for handling multiple versions of datasets. Users expressed that different versions or subsets of datasets may have different rules. The primary example was differences in rules between a deidentified versus a limited dataset such as linkage being permitted for one and not the other, or how laws like HIPAA or the Common Rule apply. In general, the schema treats one dataset as having one set of governance information, but one fully identified dataset can generate a limited and/or a deidentified dataset. There could be other ODRL-based strategies for handling these caveats that avoid having a user enter two sets of governance information for a hypothetical deidentified and limited dataset. For example, one approach would be employing constraints to communicate rules that only apply in certain situations (e.g., if de-identified, or if a certain consent form was signed).

Examine tool questions/responses with no corresponding schema class and consider if/how the schema can be updated to accommodate them. The translation exercise revealed that responses to these 10 questions do not align with the schema:

- Are identifiers accessible outside of the dataset to generate a pseudo-identifier (e.g., hash or token)?
- Enter the name of the organization(s) that funded the collection of this dataset
- Does this dataset contain Personally Identifiable Information (PII) for use in individual-level dataset linkage?
- Yes/No navigation questions for consent and DUA (e.g., Were participants consented for the collection of this dataset? Is a DUA required for dataset linkage, sharing, and secondary access and use?)
- Will minor participants be re-consented when they become adults?

- Select or enter the linkage method
- Enter the specific type of research or use
- Describe the data release process
- Enter the name of the review body needed for approval for sharing
- Enter the approved purpose

The Yes/No questions listed above were initially intended as navigation features in the Data Collection Tool; however, through the translation exercise, the team discovered that some of them could convey governance information that might be worth capturing the schema. For example, responses of ***No human participants in the dataset***, ***No,*** or ***I don't know*** to "Were participants consented for the collection of this dataset?" could not be mapped to a schema class. However, a waiver of consent requirement or no human participants being included in the dataset are meaningful pieces of governance information that could be considered for inclusion in future updates to the schema.

The free text responses listed above also emerged as areas where the schema could be augmented to capture additional governance information that may be important for making decisions about future use, sharing, and linkage of the data.

6 Conclusion

The Health FFRDC project team developed a proof-of concept Governance Metadata Collection Tool designed around how biomedical researchers and research teams conduct research and manage data governance, rather than focusing strictly on guidelines for governance metadata organization based on the new schema alone. The team collaborated with researchers as co-designers, to support an exploration of how governance information can be collected, conducted usability testing, developed open-source documentation to support others to innovate further on this proof-of-concept effort, and conducted a translation exercise to examine alignment with the data governance metadata schema. The tool includes 165 questions presented in 11 sections, organized by governance policies relevant to research (e.g., consent, Institutional Review Board [IRB], DUAs, laws).

Four user testers successfully entered the governance for a dataset into the data collection tool validating that governance information can be collected from researchers in a structured format. The project team converted questionnaire responses into governance metadata by parsing and loading metadata elements into a relational database, architected based on the schema's specifications. The usability evaluation and translation exercise generated recommendations for future schema improvements and data collection tool evolution. All those who interacted with the tool were unanimous about its value for collecting structured governance metadata and its potential for exchanging governance metadata to advance linkage implementations for research.

The preferred data collection tool design does not mirror guidelines for governance metadata organization; instead, it is designed around how research is conducted. While best practice suggests that a data collection tool should prioritize having the least number of questions, researchers as co-designers suggested a data collection tool that asks repeated questions based on how they organize and understand their governance work. Integrating researchers into a human centered design development process improved the data collection tool.

This effort is an important step toward a governance metadata standard and research norms that include collecting and transmitting governance metadata with every research dataset. Developing and testing the data collection tool demonstrated significant progress in data governance for research and identified many remaining challenges. Clearly communicating about governance information is difficult without a common understanding of key governance terms. The research ecosystem is vast and diverse. Each research topic faces unique governance challenges and has corresponding specialized governance concepts that a schema would have to represent. Governance knowledge is fragmented across many individuals and policy documents for a single dataset, and governance for similar datasets may look different across institutions. Despite these challenges, every individual who engaged with the data collection tool understood its value and role in the future of research.

Researchers are concerned about reporting incorrect governance information and raised the importance of engaging appropriate authorities to confirm and communicate governance information. This highlights a potential role for the schema and associated data collection tools in facilitating conversations with policy and legal experts and supporting experts in performing thorough policy analyses. As described in the NICHD Record Linkage Implementation Checklist, collecting and interpreting dataset-level governance information is just one of several considerations in designing a new data linkage strategy that protects research participant privacy, addresses ethics, manages risks, facilitates compliance, and respects participant trust while accelerating research. Future work to evolve governance metadata collection tools will require collaboration among and input from data providers, institutional representatives, IRBs, and policy/legal experts for deciding and communicating rules about a dataset.

If widely adopted, this work would contribute to making data more findable, accessible, interoperable, and reusable for patient-centered outcomes research and promote trust and appropriate oversight in linking individual-level participant data when collected and combined from different resources. A refined metadata governance schema and production-level data collection tools could be leveraged throughout the HHS and NIH research ecosystem, supporting innovative and responsible research to improve health outcomes for all Americans.

7 Terms and Definitions in the User Guide

Condition

Condition refers to a constraint that is applicable to a rule. Each rule can have zero or more constraints.

Consent

A consent is an IRB-approved written record that complies with the Common Rule (45 CFR 46) and, as applicable, Protection of Human Subjects rules (21 CFR 50) and is used to demonstrate a participant's or guardian's consent to participate in research.

Data Collection

Data collection means obtaining data from participants for research, clinical, or administrative purposes.

Database/Data Repository

A database or data repository is virtual data storage that stores, organizes, and validates data, and makes the data accessible for use by others.

Dataset Access

Dataset access means acquiring data from a data repository or other data sharing system for secondary research purposes.

Dataset Linkage

Dataset linkage or record linkage means combining information from a variety of data sources for the same individual.

Dataset Sharing

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.

Dataset Use

Dataset use means working with data for secondary research or other analytical purposes.

Data Use Agreement

A Data Use Agreement (DUA) is a document that establishes who is permitted to use and receive data, and the permitted uses and disclosures of such information by the recipient.

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.

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.

Entity Resolution

Entity resolution is the process of joining or matching records from one data source with another that describes the same entity. In privacy preserving record linkage (PPRL), hash codes/tokens are used to match individual records without using PII/PHI.

Governance or Data Governance

Governance or data governance 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.

Honest Broker

An honest broker is a party that holds de-identified tokens ("hashes") and operates a service that matches tokens generated across disparate datasets to formulate a single Match ID for a specific use case.

Institutional Review Board

An Institutional Review Board (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.

Laws

Local, state, or federal laws that apply to a dataset. Specific laws that apply to the dataset in this tool include:

- **Confidential Information Protection and Statistical Efficiency Act (CIPSEA) rules**

  More information about CIPSEA: https://www.cio.gov/handbook/it-laws/cipsea/
- **Family Educational Rights and Privacy Act (FERPA) rules**

  More information about FERPA: https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html
- **Health Insurance Portability and Accountability Act (HIPAA) rules**

  More information about HIPAA: https://www.hhs.gov/hipaa/index.html
- **The Common Rule**

More information about the Common Rule: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html#:~:text=For%20all%20participating%20departments%20and,regulations%20of%20that%20department%2Fagency

- **The Privacy Act of 1974**

  More information about the Privacy Act: https://www.justice.gov/opcl/privacy-act-1974

Letter of Determination

A letter of determination documents an IRB decision on the status of research.

Metadata Schema

Metadata schema, as defined in this guide, 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.

Personally Identifiable Information

Personally identifiable information (PII) is 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.

Policies

Policies are the foundation of governance information and the source of rules (permissions and prohibitions) about how a dataset is handled across the data lifecycle. Laws, DUAs, and consent forms are examples of policies.

Privacy Board

A privacy board is 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.

Privacy Preserving Record Linkage

Privacy preserving record linkage (PPRL) is 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.

Protected Health Information

Protected health information (PHI) is 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.

Rules

Rules represent a permission, prohibition, or duty associated with a policy. For each type of rule, the tool collects information about permissions and prohibition for dataset sharing, dataset linkage, and secondary dataset access and use.

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, 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
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
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)
Data provider (Also 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)

Term	Definition
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[f]	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 de-identification (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)
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)

[f] 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
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, 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
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)

Term	Definition
Honest broker	A party that holds de-identified 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)
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.))

Term	Definition
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
AHRQ	Agency for Healthcare Research and Quality
CDAC	Controlled Data Access Coordination
CMS	Centers for Medicare & Medicaid Services
COVID	Coronavirus Disease
DUA	Data Use Agreement
EHR	Electronic Health Record
FAIR	Findable, Accessible, Interoperable, and Reusable
FDA	Food and Drug Administration
FFRDC	Federally Funded Research and Development Center
FHIR	Fast Health Information Resource
HHS	Department of Health and Human Services
HIPAA	Health Insurance Portability and Accountability Act
IRB	Institutional Review Board
JSON	JavaScript Object Notation
LHC	Lister Hill Center
NCI	National Cancer Institute
NICHD	National Institute of Child Health and Human Development
NIH	National Institutes of Health
NLM	National Library of Medicine
ODK	Open Data Kit
ODRL	Open Digital Rights Language
ODSS	Office of Data Science and Sharing
OS-PCORTF	Office of the Secretary Patient-Centered Outcomes Research Trust Fund
PHI	Protected Health Information
PI	Principal Investigator
PII	Personally Identifiable Information

Acronym	Definition
RE-AIM	Reach Effectiveness Adoption Implementation Maintenance Reach
SDC	Structured Data Capture
TEP	Technical Experts Panel
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: Tools Readiness Analysis Findings

The project team applied NICHD ODSS selection criteria to screen the candidate tools for extension to support the Governance Data Collection Tool, and surfaced insights to inform candidate tool selection by NICHD ODSS and the TEP.

The project team used a qualitative approach to examine each candidate tool against defined selection criteria. NICHD ODSS predetermined criteria include:

- Number of active clinical sites or other entities using the tool
- Tool's existing eConsent or related capabilities
- Open-source community engagement for the tool
- Ability to use the FHIR-based information exchange standards

The project team reviewed literature from the research data collection tool community to identify potential concepts applicable to use and adoption to further develop these criteria, and defined five selection criteria:

1. **Form Functionality:** How well does the tool allow for the design and management of forms and forms entry, collections, retrieval, and export? Are there any restrictions on form use?
2. **Electronic Consent (eConsent) Capabilities:** Does this tool enable eConsent or related capabilities and how well do those capabilities function?
3. **FHIR-Based Exchange**: Does this tool allow exchange of form-entered data, preferably with FHIR?
4. **License and Open-Source Status:** What are the restrictions on use of the candidate tool by end users? Does the tool require a license for us? Is the tool open source?
5. **Active Use and Community Adoption:** Has the tool been adopted and used within the research community? How many clinical sites are currently using the tool? Does the tool have an active user community? Does the tool provide adequate and complete documentation to support users? Is documentation easy to locate and access, and available for public reference, allowing users to obtain necessary information without difficulty? Does the tool offer support for users?

NICHD ODSS validated the selection criteria. The project team applied the selection criteria to identified tools and documented findings within a tools inventory.

Functional Capabilities

Table 6 summarizes how each tool enables form business logic, provides a user interface for data collection, stores data, and supports data exchange, interoperability, and standards.

Table 6: Functional Capabilities by Candidate Data Collection and Exchange Tool

Functional Capabilities	REDCap	FHIR SDC	ODK	Kobo Toolbox
Form Business Logic	Form design using the Online Designer tool or by constructing a data dictionary template in Microsoft Excel with form metadata. Supports skip-logic, condition-based navigation, and pagination on V10.6 or higher.	FHIR SDC STU 3 supports skip-logic, condition-based navigation, pagination, advanced form rendering, and behavior controls. Open-source NLM Form Builder can be used to build and edit FHIR Questionnaires.	Forms are created and deployed to an ODK Central server instance and administered using the ODK Collect mobile application or web forms. Supports skip-logic, condition-based navigation, and pagination on mobile devices.	kpi to create and manage forms, reusing forms in library. Supports skip-logic, condition-based navigation, and pagination on mobile devices.
User Interface	Web-based form rendering supported by REDCap backend.	Web-based and application-based form rendering supported by SDC Questionnaire App, a SMART on FHIR® open-source application that establishes a connection with a FHIR Server and provides an interface for selecting Questionnaires, filling them out, and saving Questionnaires and Observation data.	Web-based and mobile device form rendering by ODK platform.	Web-based and mobile device form rendering by Kobo Toolbox platform based on ODK. The forms are compatible with Enketo web forms, and KoboCollect Android application for collecting data on mobile.

Functional Capabilities	REDCap	FHIR SDC	ODK	Kobo Toolbox
Data-Collection Backend	REDCap is a PHP application that can be deployed using a web server (e.g. Microsoft IIS or Apache) with PHP 7.2.5 and higher and database server with MySQL 5.5.5+ or MariaDB 5.5.5+. REDCap can also be deployed using cloud providers such as Amazon's AWS CloudFormation service and Microsoft's Azure Cloud Platform, which have collaborated with REDCap consortium to enable setup of a REDCap server environment.	FHIR capable back-end systems such as the open-source, object-relational database, PostgreSQL with the FHIRbase extension for storing and retrieving FHIR resources and LinuxForHealth. FHIR Server such as HAPI, Spark, Node on FHIR, and LinuxForHealth. For cloud implementations, Microsoft offers open-source FHIR Server for Azure.	ODK Central is the ODK back-end service and can be deployed using a Postgres database. ODK Central manages accounts and permissions, and stores form definitions. Additionally, ODK Central allows clients to connect for form download and upload.	kobocat and kobocat-templates for deploying surveys, collecting, and analyzing data; enketo-express web application for collecting data, previewing forms, editing data submission. KoboCAT is used for receiving form submissions and can be deployed on a Linux server via docker images.
Data Exchange, Interoperability, and Standards	API supports the import, export, and modification of data in the data migration process. Data can be exported in common formats (SPSS, SAS, Stata, R, CSV, Excel, and CDISC ODM). Form designs can be standardized, e.g., CDISC CDASH for interoperable reuse. REDCap code standard is PHP.	Exchange (import and export) via FHIR API with any endpoint supporting the FHIR API standard. Bulk FHIR enables exchanging large datasets. FHIR SDC uses Questionnaire and QuestionnaireResponse as standards.	ODK Central allows download of submitted form data as CSV files or as JSON via an OData endpoint and XML via the ODK Briefcase component. Uses W3C XForm XML mark for form standard and Enketo web forms.	Kobo Toolbox KPI provides API to export form data into other applications and allows for easy integration with other software. Uses XForm standard for form development.

Non-functional Capabilities

Table 7 summarizes how each tool enables non-functional requirements including performance and load time response, privacy and security considerations, integration capabilities, and ability to scale in operational environment.

Table 7: Non-functional Capabilities by Candidate Data Collection and Exchange Tool

Non functional Capabilities	REDCap	FHIR SDC	ODK	Kobo Toolbox
Performance and Load Time Response	Lightweight application with minimal load demand for form design and data entry. No hard requirements for server processing speed, memory, or hard drive space. Record Status Dashboard may be slow if the project contains a large number of records.	SDC Questionnaire App is a lightweight and performant JavaScript application. FHIR services are successfully deployed in high demand operational environments.	ODK back-end services were designed to be performant on modest hardware. Tool performance notes indicate the services have been benchmarked on a small VM server to support 50 concurrent 5 MB submissions.	KPI and KoboCAT services are built using Python and Django on top of performant Postgres and MongoDB databases and offer sufficient performance and load time.
Privacy and Security Considerations	Supports access control and permissioning; recommended infrastructure configuration is secure, recommend implementing the Web Server in a DMZ,[g] keeping the database and file servers behind a firewall, and using WebDAV protocol (SSL supported) for communicating with the Web Server.	All FHIR Resources have privacy and security related standards and metadata based on FHIR Data Segmentation for Privacy Implementation Guide. FHIR Security Module and Implementer's Safety Checklist provide guidance and standards for all FHIR artifacts. SMART on FHIR App for authentication and authorization of FHIR RESTful API actions.	Supports access control and permissioning, supports form encryption, and recommends secure database configuration, encrypting server-client connections and data. ODK Authentication API authorizes HTTP transactions.	Kobo Security supports disk-level encryption at rest and optional data-level encryption; access control; and permissioning for access, creation, use, and deletion at different levels for different users. Free public KoboToolbox servers are not HIPAA Privacy and Security compliant.

[g] DMZ or "demilitarized zone" is a specially designed network that segregates internal systems and network communications from servers communicating with the internet. Servers and external services such as email, VOIP, and file transfer within the DMZ are separated from internal servers by a firewall or other security gateways that filter communications from the DMZ.

Non functional Capabilities	REDCap	FHIR SDC	ODK	Kobo Toolbox
Integration Capabilities	REDCap Clinical Data Interoperability Service can be integrated with EHR and other source systems using a FHIR API to extract REDCap project data such as the USCDI data required for EHR certification. REDCap projects can extract medical device and payer data as well. REDCap projects can upload data captured by MyCap Android and iOS mobile device applications. REDCap has also been integrated with ODK and Kobo Toolkit.	By converting non-standard data to the FHIR data model, searches and exchange can be integrated among disparately formatted systems. Doing so facilitates compilation of large datasets for analytics and machine learning.	REDCap can be integrated with ODK. ODK is also integrated with Kobo Toolbox platform.	Integration using an API (application programming interface) and REST services to download forms and retrieve data from other systems. REDCap can be used with Kobo Toolbox.
Ability to Scale in Operational Environment	Currently operating at scale in diverse operational environments. Operational REDCap infrastructure requirements are modest according to REDCap Technical Overview.	An operational FHIR Server and database are sufficient for governance metadata form entry, collection, management, and exchange at a single repository. These FHIR infrastructure capabilities are also available from cloud providers such as FHIR Server for Azure	ODK is primarily used as a hosted service ODK Cloud, which could scale for many users to share a governance metadata form. Alternatively, ODK Central database can be implemented on private servers or cloud service and scaled by the implementing entity. Alternatively, ODK can be implemented on private servers and scaled by the implementing entity.	Kobo Toolkit is primarily used as a hosted service, which could scale for many users to share a governance metadata form.

Table 8: Estimation of Level of Effort for Tool Extension

Development Activity	Development Activity Description	Estimated Hours
1. Establish REDCap instance	Deploy REDCap capable web server that supports current release of PHP, database server with MySQL 5.5.5+ or MariaDB 5.5.5+ and MySQL client, SMTP email server, and File server if the web server is internet accessible.	8
2. Add questions to REDCap and create form	Add template questions in the order laid out the template spreadsheet into REDCap form using the Online Designer tool. Add form controls that aid data entry, navigation and form management, and pop-ups to display associated information or instructions.	10
3. Create business logic	Develop any logic (e.g., to skip questions) and necessary value sets and reference capabilities needed to populate variables into REDCap database.	20
Total hours	-	**38**

The expert estimated 38 hours to establish a REDCap instance and create the form(s) for data collection. No level of effort is estimated for back-end data collection as a back-end database is part of REDCap and populated through form creation. Additionally, REDCap's exchange capabilities are limited but development effort is not estimated for setting up FHIR-based data exchange functions.

A MITRE FHIR expert was consulted to define the development activities associated with applying FHIR SDC for governance metadata collection. The estimation assumes that the FHIR Questionnaire standard is applied through SDC. Table 9 lists FHIR development activities and the level of effort associated with each activity.

Table 9: FHIR Level of Effort Estimation by Development Activity

Development Activity	Development Activity Description	Estimated Hours
1. Establish FHIR Server and database	Establish FHIR server and database for data storage. Includes setup server, API, Postgres database, and data access layer.	16
2. Build FHIR Questionnaire	Assumes there are multiple Questionnaires for different use cases. Each Questionnaire is static and saved on server.	16
3. Build front-end web client	Build front-end web client using LHC component.	16
4. Integrate back-end data collection	Read/save QuestionnaireResponse: including generate QuestionnaireResponse from web client, send QuestionnaireResponse to server, save QuestionnaireResponse to database, and later retrieve saved QuestionnaireResponse from server.	16

Development Activity	Development Activity Description	Estimated Hours
5. Implement security controls	Authentication and Authorization for integration into a repository. This does not include implementation of SMART App Launch (an optional feature).	12
Total hours	_	**76**

Appendix C: Research Co-designers

Table 10: Researcher Co-designers

Researcher	Affiliation
Ananth Annapragada, PhD, FAIMBE, FNAI	**PreVAIL Kids Investigator** Director Translational Imaging Group: TIGr Vice-Chair for Research, Texas Children's Hospital Department of Radiology Professor of Radiology and Professor of Obstetrics and Gynecology, Baylor College of Medicine
Cedric Manlhiot, PhD	**PreVAIL Kids Investigator** Director, Cardiovascular Analytic Intelligence Initiative (CV-Ai^2) Assistant Professor, Department of Pediatrics, Division of Cardiology Blalock-Taussig-Thomas Pediatric and Congenital Heart Center, Johns Hopkins School of Medicine
Adam C. Resnick, PhD	**Kids First Investigator** Director, Center for Data Driven Discovery in Biomedicine Alexander B. Wheeler Endowed Chair in Neurosurgical Research Research Professor of Neurosurgery at the Perelman School of Medicine Children’s Hospital of Philadelphia
Elizabeth E. Umberfield, PhD, RN, NI-BC	**Nurse Scientist** Division of Nursing Research and Department of Artificial Intelligence and Informatics Mayo Clinic Member of the Technical Experts Panel

Appendix D: 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 collection tool 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 tool, 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 met at our orientation meeting if you were able to attend, who will be observing the session and taking notes.

Purpose of the Usability Evaluation

As a quick refresher, we have worked with the 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 collection tool to collect governance information about a dataset through a questionnaire and then transform questionnaire responses into structured metadata. Structured governance metadata can facilitate the determination of whether a dataset can be linked (combined with data from other sources that relate to the same person) and if so, what rules flow down to the linked dataset.

The tool, developed as a Fast Healthcare Information Resource Structured Data Capture questionnaire, enables research study teams to convey dataset governance information in a consistent, machine-readable format.

In this session, we will be testing the usability of the data collection tool to gather governance information about a research dataset 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 who wants to share a dataset from one of your studies with other researchers. This tool assumes that you are entering governance information for a single dataset retrospectively, after the data collection has occurred. You will document the policies and regulations that other researchers would have to follow in order to link, share, access, and use your dataset.

Before this session we requested that you think of a dataset that you have created through your research that you would like to use as an example as you fill out the questions in the data collection tool. Please use that dataset to answer the questions in the tool.

The goal of this session is not to test your knowledge of data governance or to gather information about your chosen dataset. The goal is to see if the tool can be easily used by researchers, and if it's helpful for gathering governance information. You will be entering fake names for your research institution and chosen dataset to protect your privacy. I will give you more instructions before you begin interacting with the tool.

Overall Session Flow

This session will start with some basic questions about your research and past experience with dataset linkage for individual-level data. Then we will ask you to interact with the tool and answer questions related to the governance information of your chosen dataset, and we will observe those interactions. At the end, we will have questions about your impression of the tool's usability and usefulness, challenges you faced using the tool, 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 been conducting 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.
        - Both—cases where datasets were all owned by my institution, and cases where they were owned by multiple different institutions.

Governance Metadata Collection Tool Interaction

Data Entry Instructions

Thank you for answering those initial questions! Now we will ask you to enter data governance information into the tool. I want you to navigate to the data collection tool using the link that I have placed in the chat: https://warp.mitre.org/data-linkage-governance/collection-tool/. 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 tool. I have a few instructions for you before you begin.

To protect your privacy, you will not be entering your name, the name of your dataset, or the name of your institution into the tool. For today, your name will be Researcher #, your dataset will be Dataset #, your research study will be Study #, and your research institution will be Institution #.

Part of our evaluation considers the time required to complete each section. This questionnaire is organized into 11 different sections. For each section I will give you a verbal cue requesting that you start filling out information in that section. When you have finished filling it out, please verbally let me know that you are finished. If you get to a point in a section where you are unsure of what to do or where to go, and feel that you can't continue with the section, please tell us and we will move on to the next section.

As you are navigating this tool, I'd like you to “think aloud” or verbalize your thought process as you move through the questionnaire, consider questions, and provide answers.

Anything that you are thinking related to the tool, or any questions that you may have as you work through the tool, please talk through them out loud, including anything that is unclear or confusing.

I may not answer questions that you ask during the session, but please speak to them out loud so that we can collect them for our discussion at the end of the session.

If you have questions about any of the wording or terms that are used in the questionnaire, please let us know, and we can put definitions in the chat window.

As you navigate the tool, you are not expected to reference any documentation about the governance of your chosen dataset—documentation such as IRB protocols or consent forms. Please answer the questions to the best of your knowledge.

To manage time for this session, at points I may interrupt as you are entering information and request that you move on to the next section.

At the bottom of the questionnaire there will be a button titled “Save.” At the end of this session, we would like you to save your responses so that we can review them after the session.

Do you have any questions before we begin?

Appendix E: Usability Evaluation Analysis Codebook

Table 11: Usability Evaluation Analysis 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.

Code	Source	Definition
Price Value	UTAUT	Cognitive tradeoff between the perceived benefits of an application and the monetary cost for using it.
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.
Interrelated Policies	Evaluation Data	Situation where one form of governance sources its rules from another form of governance.
Governance Authorities	Evaluation Data	The involvement of an authority who has the vested power to establish or interpret dataset governance policy.
Sharing Context	Evaluation Data	Situation where the rules governing the linkage of a dataset depend on the context of sharing, such as what subset of the data are shared and the entity the data are being shared with.
Missing Governance	Evaluation Data	Sources of dataset governance not addressed directly by the metadata collection tool.
Knowledge Required	Evaluation Data	Indication that a user would need training on dataset governance or the tool to use the tool effectively.

Appendix F: Data Collection Tool Questions and Response Options

The Data Collection Tool version 2.8.6 included the 165 questions organized into 11 sections. Each question allows users to either type a value, select one from defined response values, select one from defined response values and/or type a value, or select one or more from defined response values and/or type a value.

The data collection tool includes conditional logic that helps the user skip question when one response negates the relevance of subsequence questions. Conditional logic is represented below as (*skip to question X)*.

Section 1: User Information

1.1. Question: Enter your first name [type a value]
   a. Response: [type a value]

1.2. Question: Enter your last name [type a value]
   a. Response: [type a value]

1.3. Question: Enter the organization(s) that you represent [type a value]
   a. Response: [type a value]

1.4. Question: Select or enter the value that best represents your role related to this dataset [select one or more or type a value]
   a. Response Value: Associate investigator
   b. Response Value: Co-principal investigator
   c. Response Value: Principal investigator
   d. Response Value: Research study coordinator
   e. Response Value: IRB representative
   f. Response Value: Signing official
   g. Response Value: Legal representative
   h. Response: [type a value]

Section 2: Dataset Information

2.1. Question: Enter the name of the dataset [type a value]
   a. Response: [type a value]

2.2. Question: Enter the name of the study that generated this dataset [type a value]
   a. Response: [type a value]

2.3. Question: Select or enter the type of data in the dataset [select one or more or type a value]
   a. Response Value: Administrative
   b. Response Value: Claims

c. Response Value: Clinical
d. Response Value: Electronic Health Record (EHR)
e. Response Value: Environmental
f. Response Value: Data generated from biospecimens
g. Response Value: Genomic data
h. Response Value: Patient generated health data (e.g., PROs, RPM, wearables, devices, sensors)
i. Response Value: Survey
j. Response: [type a value]

2.4. Question: Enter the name of the organization that collected the dataset [type a value]
   a. Response: [type a value]

2.5. Question: Enter the name of the organization(s) that funded the collection of this dataset [type a value]
   a. Response: [type a value]

2.6. Question: Enter the grant number for the collection of this dataset [type a value]
   a. Response: [type a value]

2.7. Question: Does this dataset contain Personally Identifiable Information (PII) for use in individual-level dataset linkage? [select one]
   a. Response Value: Yes
   b. Response Value: No *(skip to question 2.10)*
   c. Response Value: I don't know *(skip to question 2.10)*

2.8. Question: Select which PII elements the dataset contains [select one or more or type a value]
   a. Response Value: Name
   b. Response Value: Geographic subdivision smaller than state (e.g., address or zip code)
   c. Response Value: Elements of dates (except for year) related to birth, death, or medical encounters
   d. Response Value: Telephone number
   e. Response Value: Fax number
   f. Response Value: Email address
   g. Response Value: Social security number
   h. Response Value: Medical record number
   i. Response Value: Health plan beneficiary number
   j. Response Value: Account number
   k. Response Value: Certificate or license number
   l. Response Value: Vehicle identifier and serial numbers, including license plate numbers
   m. Response Value: Device identifier
   n. Response Value: Web URL
   o. Response Value: Internet protocol address

p. Response Value: Biometric identifiers (e.g., fingerprint or voice)
q. Response Value: Photographic image
r. Response: [type a value]

2.9. Question: Enter the organization that holds these PII elements. [type a value]
   a. Response: [type a value] (*skip to question 3.1)*

2.10. Question: Select or enter the method used to de-identify the dataset. [select one or type a value]
   a. Response Value: HIPAA - Safe Harbor
   b. Response Value: HIPAA - Expert Determination
   c. Response Value: Dataset contains no individual-level data
   d. Response: [type a value]

2.11. Question: Are identifiers accessible outside of the dataset to generate a pseudo-identifier (e.g., hash or token)? [select one]
   a. Response Value: Yes
   b. Response Value: No
   c. Response Value: I don't know

Section 3: History of Dataset Linkage

3.1. Question: Has this dataset previously been linked? [select one]
   a. Response Value: Yes
   b. Response Value: No *(skip to question 4.1)*
   c. Response Value: Information not available/found *(skip to question 4.1)*
   d. Response Value: I don't know *(skip to question 4.1)*

3.2. Question: Select or enter the linkage method [select one or more or type a value]
   a. Response Value: Privacy Preserving Record Linkage (PPRL)
   b. Response Value: Clear text
   c. Response: [type a value]

3.3. Question: Enter the name of the project [type a value]
   a. Response: [type a value]

3.4. Enter the name of the dataset or describe the group of datasets [type a value]
   a. Response: [type a value]

3.5. Question: Enter who collected the dataset or group of datasets [type a value]
   a. Response: [type a value]

Section 4: Institutional Review Board

4.1. Question: Is this dataset governed by an IRB Protocol? [select one]
   a. Response Value: Yes

b. Response Value: Yes, but determined to be non-human subject's research *(skip to question 5.1)*

c. Response Value: No *(skip to question 5.1)*

4.2. Question: How many IRBs are involved with this dataset [select one or type a value]

a. Response Value: A single IRB

b. Response Value: Multiple IRBs

c. Response: [type a value]

4.3. Question: Enter the name of the IRB of record [type a value]

a. Response: [type a value]

4.4. Question: For a future linkage study, would a requester of this dataset need to receive study-specific approval from an IRB? [select one]

a. Response Value: Yes

b. Response Value: No

c. Response Value: I don't know

4.5. Question: Does the IRB permit dataset linkage? [select one]

a. Response Value: Yes *(skip to question 4.9)*

b. Response Value: Yes, with conditions

c. Response Value: No *(skip to question 4.9)*

d. Response Value: I don't know *(skip to question 4.9)*

e. Response Value: It doesn't say *(skip to question 4.9)*

4.6. Question: Select or enter the dataset linkage conditions that the IRB applies [select one or more or type a value]

a. Response Value: The dataset may only be linked with the approval of the IRB of record *(skip to question 4.9)*

b. Response Value: The dataset may only be linked with the approval of data contributing sites *(skip to question 4.9)*

c. Response Value: The dataset may only be linked using a specific linkage method

d. Response Value: The dataset may only be linked for specific types of research or use

e. Response: [type a value] *(skip to question 4.9)*

4.7. Question: Select or enter the linkage method [select one or type a value]

a. Response Value: PPRL

b. Response Value: Clear text

c. Response: [type a value]

4.8. Question: Enter the specific types of research or use [type a value]

a. Response: [type a value]

4.9. Question: Does the IRB permit dataset sharing? [select one]

a. Response Value: Yes *(skip to question 4.13)*

b. Response Value: Yes, with conditions
c. Response Value: No *(skip to question 4.13)*
d. Response Value: I don't know *(skip to question 4.13)*
e. Response Value: It doesn't say *(skip to question 4.13)*

4.10. Question: Select or enter the dataset sharing conditions that the IRB applies [select one or more]
a. Response Value: The dataset may only be shared as a de-identified dataset *(skip to question 4.13)*
b. Response Value: The dataset may only be shared following the Safe Harbor de-identification method *(skip to question 4.13)*
c. Response Value: The dataset may only be shared if approved by a review body
d. Response Value: The dataset may only be shared as a limited dataset *(skip to question 4.13)*
e. Response Value: The dataset may only be shared within a defined data release process
f. Response: [type a value] *(skip to question 4.13)*

4.11. Question: Enter the review body needed for approval for sharing
a. Response: [type a value]

4.12. Question: Describe the data release process
a. Response: [type a value]

4.13. Question: Does the IRB permit secondary dataset access? [select one]
a. Response Value: Yes *(skip to question 4.15)*
a. Response Value: Yes, with conditions
b. Response Value: No *(skip to question 4.15)*
c. Response Value: I don't know *(skip to question 4.15)*
d. Response Value: It doesn't say *(skip to question 4.15)*

4.14. Question: Select or enter the secondary dataset access conditions that the IRB applies [select one or more or type a value]
a. Response Value: The dataset may only be accessed in a data enclave
b. Response Value: The dataset may only be accessed in a controlled environment
c. Response: [type a value]

4.15. Question: Does the IRB permit secondary dataset use? [select one]
a. Response Value: Yes *(skip to question 4.18)*
b. Response Value: Yes, with conditions
c. Response Value: No *(skip to question 4.18)*
d. Response Value: I don't know *(skip to question 4.18)*
e. Response Value: It doesn't say *(skip to question 4.18)*

4.16. Question: Select or enter the secondary dataset use conditions that the IRB applies [select one or more]
a. Response Value: This dataset may only be used for an approved purpose

b. Response Value: This dataset may only be used for research on a specific topic *(skip to question 4.18)*
c. Response: [type a value] *(skip to question 4.18)*

4.17. Question: Enter the approved purpose
a. Response: [type a value]

4.18. Question: Select or enter the prohibitions [select one or more or type a value]
a. Response Value: Individuals in the dataset may not be re-identified
b. Response Value: The dataset may not be linked
c. Response Value: The dataset may not be shared
d. Response Value: The dataset may not be accessed
e. Response Value: The dataset may not be used for secondary purposes
f. Response Value: The dataset may not be used for commercial purposes
g. Response Value: The dataset may not be used for any purposes beyond explicit permissions
h. Response Value: The dataset may not be sold
i. Response: [type a value]

Section 5: Consent

5.1. Question: Were participants consented for the collection of this dataset? [select one]
a. Response Value: Yes
b. Response Value: No, consent was waived by the IRB *(skip to question 6.1)*
c. Response Value: I don't know *(skip to question 6.1)*
d. Response Value: No human participants in this dataset *(skip to question 6.1)*

5.2. Question: Will minor participants be re-consented when they become adults? [select one]
a. Response Value: Yes
b. Response Value: No
c. Response Value: I don't know
d. Response Value: There are no minors in this dataset
e. Response: [type a value]

5.3. Question: Does the consent permit dataset linkage? [select one]
a. Response Value: Yes *(skip to question 5.7)*
b. Response Value: Yes, with conditions
c. Response Value: No *(skip to question 5.7)*
d. Response Value: I don't know *(skip to question 5.7)*
e. Response Value: It doesn't say *(skip to question 5.7)*

5.4. Question: Select the dataset linkage conditions that the consent applies [select one or more or type a value]
a. Response Value: The dataset may only be linked with the approval of the IRB of record

b. Response Value: The dataset may only be linked with the approval of data contributing sites *(skip to question 5.7)*
c. Response Value: The dataset may only be linked using a specific linkage method
d. Response Value: The dataset may only be linked for specific types of research or use
e. Response: [type a value] *(skip to question 5.7)*

5.5. Question: Select or enter the linkage method [select one or type a value]
a. Response Value: PPRL
b. Response Value: Clear text
c. Response: [type a value]

5.6. Question: Enter the specific type of research or use [type a value]
a. Response: [type a value]

5.7. Question: Does the consent permit dataset sharing? [select one]
a. Response Value: Yes *(skip to question 5.11)*
b. Response Value: Yes, with conditions
c. Response Value: No *(skip to question 5.11)*
d. Response Value: I don't know *(skip to question 5.11)*
e. Response Value: It doesn't say *(skip to question 5.11)*

5.8. Question: Select the dataset sharing conditions that the consent applies [select one or more]
a. Response Value: The dataset may only be shared as a de-identified dataset *(skip to question 48)*
b. Response Value: The dataset may only be shared following the Safe Harbor de-identification method *(skip to question 48)*
c. Response Value: The dataset may only be shared if approved by a review body
d. Response Value: The dataset may only be shared as a limited dataset *(skip to question 48)*
e. Response Value: The dataset may only be shared within a defined data release process
f. Response: [type a value]

5.9. Question: Describe the data release process
a. Response: [type a value]

5.10. Question: Enter the name of the review body needed for approval for sharing
a. Response: [type a value]

5.11. Question: Does the consent permit secondary dataset access? [select one]
a. Response Value: Yes *(skip to question 5.13)*
b. Response Value: Yes, with conditions
c. Response Value: No *(skip to question 5.13)*
d. Response Value: I don't know *(skip to question 5.13)*
e. Response Value: It doesn't say *(skip to question 5.13)*

5.12. Question: Select the secondary dataset access conditions that the consent applies [select one or more]
    a. Response Value: The dataset may only be accessed in a data enclave
    b. Response Value: The dataset may only be accessed in a controlled environment
    c. Response: [type a value]
5.13. Question: Does the consent permit secondary dataset use? [select one]
    a. Response Value: Yes *(skip to question 5.16)*
    b. Response Value: Yes, with conditions
    c. Response Value: No *(skip to question 5.16)*
    d. Response Value: I don't know *(skip to question 5.16)*
    e. Response Value: It doesn't say *(skip to question 5.16)*
5.14. Question: Select the secondary dataset use conditions that the consent applies [select one or more]
    a. Response Value: This dataset may only be used for an approved purpose
    b. Response Value: This dataset may only be used for research on a specific topic *(skip to question 5.16)*
    c. Response: [type a value]
5.15. Question: Enter the approved purpose
    a. Response: [type a value]
5.16. Question: Select or enter the prohibitions [select one or more]
    a. Response Value: Individuals in the dataset may not be re-identified
    b. Response Value: The dataset may not be linked
    c. Response Value: The dataset may not be shared
    d. Response Value: The dataset may not be accessed for secondary use
    e. Response Value: The dataset may not be used for secondary purposes
    f. Response Value: The dataset may not be used for commercial purposes
    g. Response Value: The dataset may not be used for any purposes beyond explicit permissions
    h. Response Value: The dataset may not be sold
    i. Response: [type a value]

Section 6: Privacy Board

6.1. Question: Is this dataset governed by a Privacy Board? [select one]
    a. Response Value: Yes
    b. Response Value: No *(skip to question 7.1)*
    c. Response Value: I don't know *(skip to question 7.1)*
6.2. Question: Does the Privacy Board policy permit dataset linkage? [select one]
    a. Response Value: Yes *(skip to question 6.6)*

b. Response Value: Yes, with conditions
c. Response Value: No *(skip to question 6.6)*
d. Response Value: I don't know *(skip to question 6.6)*
e. Response Value: It doesn't say *(skip to question 6.6)*

6.3. Question: Select or enter the dataset linkage conditions that the Privacy Board applies [select one or more]
a. Response Value: The dataset may only be linked with the approval of the IRB of record *(skip to question 6.6)*
b. Response Value: The dataset may only be linked with the approval of data contributing sites *(skip to question 6.6)*
c. Response Value: The dataset may only be linked using a specific linkage method
d. Response Value: The dataset may only be linked for specific types of research or use
e. Response: [type a value] *(skip to question 6.6)*

6.4. Question: Select or enter the linkage method [select one]
a. Response Value: PPRL
b. Response Value: Clear text
c. Response: [type a value]

6.5. Question: Enter the type of research or use
a. Response: [type a value]

6.6. Question: Does the Privacy Board policy permit dataset sharing? [select one]
a. Response Value: Yes *(skip to question 6.10)*
b. Response Value: Yes, with conditions
c. Response Value: No *(skip to question 6.10)*
d. Response Value: I don't know *(skip to question 6.10)*
e. Response Value: It doesn't say *(skip to question 6.10)*

6.7. Question: Select or enter the dataset sharing conditions that the Privacy Board applies [select one or more]
a. Response Value: The dataset may only be shared as a de-identified dataset *(skip to question 6.10)*
b. Response Value: The dataset may only be shared following the Safe Harbor de-identification method *(skip to question 6.10)*
c. Response Value: The dataset may only be shared if approved by a review body
d. Response Value: The dataset may only be shared as a limited dataset *(skip to question 6.10)*
e. Response Value: The dataset may only be shared within a defined data release process
f. Response: [type a value] *(skip to question 6.10)*

6.8. Question: Describe the review body needed for approval for sharing
a. Response: [type a value]

6.9. Question: Describe the data release process
    a. Response: [type a value]
6.10. Question: Does the Privacy Board policy permit secondary dataset access? [select one]
    a. Response Value: Yes *(skip to question 6.12)*
    b. Response Value: Yes, with conditions
    c. Response Value: No *(skip to question 6.12)*
    d. Response Value: I don't know *(skip to question 6.12)*
    e. Response Value: It doesn't say *(skip to question 6.12)*
6.11. Question: Select or enter the secondary dataset access conditions that the Privacy Board applies [select one or more]
    a. Response Value: The dataset may only be accessed in a data enclave
    b. Response Value: The dataset may only be accessed in a controlled environment
    c. Response: [type a value]
6.12. Question: Does the Privacy Board policy permit secondary dataset use? [select one]
    a. Response Value: Yes *(skip to question 6.15)*
    b. Response Value: Yes, with conditions
    c. Response Value: No *(skip to question 6.15)*
    d. Response Value: I don't know *(skip to question 6.15)*
    e. Response Value: It doesn't say *(skip to question 6.15)*
6.13. Question: Select or enter the secondary dataset use conditions that the Privacy Board applies [select one or more]
    a. Response Value: This dataset may only be used for the approved purpose
    b. Response Value: This dataset may only be used for research on a specific topic *(skip to question 6.15)*
    c. Response: [type a value] *(skip to question 6.15)*
6.14. Question: Enter the approved purpose
    a. Response: [type a value]
6.15. Question: Select or enter the prohibitions [select one or more]
    a. Response Value: Individuals in the dataset may not be re-identified
    b. Response Value: The dataset may not be linked
    c. Response Value: The dataset may not be shared
    d. Response Value: The dataset may not be accessed for secondary purposes
    e. Response Value: The dataset may not be used for secondary purposes
    f. Response Value: The dataset may not be used for commercial purposes
    g. Response Value: The dataset may not be used for any purposes beyond explicit permissions
    h. Response Value: The dataset may not be sold
    i. Response: [type a value]

Section 7: Data Use Agreement

7.1. Question: Is a DUA required for dataset linkage, sharing, and secondary access and use? [select one]

a. Response Value: Yes

b. Response Value: No *(skip to question 8.1)*

c. Response Value: I don't know *(skip to question 8.1)*

7.2. Question: Enter the name of the DUA

a. Response: [type a value]

7.3. Question: Select or enter the organization providing the dataset [select one]

a. Response Value: Data coordinating center

b. Response Value: Data provider

c. Response Value: Data repository

d. Response Value: Government organization

e. Response: [type a value]

7.4. Question: Select or enter the data receiving organization [select one]

a. Response Value: Data coordinating center

b. Response Value: Data provider

c. Response Value: Data requester

d. Response Value: Government organization

e. Response Value: Data repository

f. Response: [type a value]

7.5. Question: Does the DUA permit dataset linkage? [select one]

a. Response Value: Yes *(skip to question 7.9)*

b. Response Value: Yes, with conditions

c. Response Value: No *(skip to question 7.9)*

d. Response Value: I don't know *(skip to question 7.9)*

e. Response Value: It doesn't say *(skip to question 7.9)*

7.6. Question: Select or enter the dataset linkage conditions that the DUA applies [select one or more]

a. Response Value: The dataset may only be linked with the approval of the IRB of record *(skip to question 7.9)*

b. Response Value: The dataset may only be linked with the approval of data contributing sites *(skip to question 7.9)*

c. Response Value: The dataset may only be linked using a specific linkage method

d. Response Value: The dataset may only be linked for specific types of research or use

e. Response: [type a value] *(skip to question 7.9)*

7.7. Question: Enter the specific types of research or use

a. Response: [type a value]

7.8. Question: Select or enter the linkage method [select one]
   a. Response Value: PPRL
   b. Response Value: Clear text
   c. Response: [type a value]

7.9. Question: Does the DUA permit dataset sharing? [select one]
   a. Response Value: Yes *(skip to question 7.13)*
   b. Response Value: Yes, with conditions
   c. Response Value: No *(skip to question 7.13)*
   d. Response Value: I don't know *(skip to question 7.13)*
   e. Response Value: It doesn't say *(skip to question 7.13)*

7.10. Question: Select or enter the dataset sharing conditions that the DUA applies [select one or more]
   a. Response Value: The dataset may only be shared as a de-identified dataset *(skip to question 7.13)*
   b. Response Value: The dataset may only be shared following the Safe Harbor de-identification method *(skip to question 7.13)*
   c. Response Value: The dataset may only be shared if approved by a review body
   d. Response Value: The dataset may only be shared as a limited dataset *(skip to question 7.13)*
   e. Response Value: The dataset may only be shared within a defined data release process
   f. Response: [type a value] *(skip to question 7.13)*

7.11. Question: Enter the name of the review body needed for approval for sharing
   a. Response: [type a value]

7.12. Question: Describe the data release process
   a. Response: [type a value]

7.13. Question: Does the DUA permit secondary dataset access? [select one]
   a. Response Value: Yes *(skip to question 7.15)*
   b. Response Value: Yes, with conditions
   c. Response Value: No *(skip to question 7.15)*
   d. Response Value: I don't know *(skip to question 7.15)*
   e. Response Value: It doesn't say *(skip to question 7.15)*

7.14. Question: Select or enter the secondary dataset access conditions that the DUA applies [select one or more]
   a. Response Value: The dataset may only be accessed in a data enclave
   b. Response Value: The dataset may only be accessed in a controlled environment
   c. Response: [type a value]

7.15. Question: Does the DUA permit secondary dataset use? [select one]
   a. Response Value: Yes *(skip to question 7.18)*

b. Response Value: Yes, with conditions
c. Response Value: No *(skip to question 7.18)*
d. Response Value: I don't know *(skip to question 7.18)*
e. Response Value: It doesn't say *(skip to question 7.18)*

7.16. Question: Select or enter the secondary dataset use conditions that the DUA applies [select one or more]
a. Response Value: This dataset may only be used for the approved purpose
b. Response Value: This dataset may only be used for research on a specific topic *(skip to question 86)*
c. Response: [type a value]

7.17. Question: Enter the approved purpose
a. Response: [type a value]

7.18. Question: Select or enter the prohibitions [select one or more]
a. Response Value: Individuals in the dataset may not be re-identified
b. Response Value: The dataset may not be linked
c. Response Value: The dataset may not be shared
d. Response Value: The dataset may not be accessed for secondary purposes
e. Response Value: The dataset may not be used for secondary purposes
f. Response Value: The dataset may not be used for commercial purposes
g. Response Value: The dataset may not be used for any purpose beyond explicit permissions
h. Response Value: The dataset may not be sold
i. Response: [type a value]

Section 8: Data Submission Agreement

8.1. Question: Is the dataset governed by a data submission agreement or institutional certification? [select one]
a. Response Value: Yes
b. Response Value: No *(skip to question 9.1)*
c. Response Value: I don't know *(skip to question 9.1)*
d. Response Value: Not yet, but will be in the future *(skip to question 9.1)*

8.2. Question: Enter the name of the data submission agreement or institutional certification
a. Response: [type a value]

8.3. Question: Select the most accurate description of this policy [select one]
a. Response Value: Data submission agreement *(skip to question 8.7)*
b. Response Value: NIH Institutional Certification
c. Response Value: Other *(skip to question 8.6)*
d. Response: [type a value] *(skip to question 8.7)*

8.4. Question: Select the research data use limitation [select one]
   a. Response Value: General Research Use (GRU)
   b. Response Value: Health/Medical/Biomedical (HMB)
   c. Response Value: Disease-specific (DS)

8.5. Question: Select any modifiers [select one or more]
   a. Response Value: IRB approval required *(skip to question 8.7)*
   b. Response Value: Publication required *(skip to question 8.7)*
   c. Response Value: Collaboration required *(skip to question 8.7)*
   d. Response Value: Not-for-profit use only *(skip to question 8.7)*
   e. Response Value: Methods *(skip to question 8.7)*
   f. Response Value: Genetic studies only *(skip to question 8.7)*

8.6. Question: Is this an agreement or certification? [select one]
   a. Response Value: Institutional Certification
   b. Response Value: Agreement

8.7. Question: Who is the assigning party? [select one]
   a. Response Value: Certification organization
   b. Response Value: Data coordinating center
   c. Response Value: Data provider
   d. Response Value: Data repository
   e. Response Value: Government organization
   f. Response Value: Principal investigator
   g. Response: [type a value]

8.8. Question: Who is the assigned party? [select one]
   a. Response Value: Data coordinating center
   b. Response Value: Data provider
   c. Response Value: Data repository
   d. Response Value: Government organization
   e. Response Value: Principal investigator

8.9. Question: Does the data submission agreement or institutional certification permit dataset linkage? [select one]
   a. Response Value: Yes *(skip to question 8.13)*
   b. Response Value: Yes, with conditions
   c. Response Value: No *(skip to question 8.13)*
   d. Response Value: I don't know *(skip to question 8.13)*
   e. Response Value: It doesn't say *(skip to question 8.13)*

8.10. Question: Select the data linkage conditions that the data submission agreement or institutional certification applies [select one]
    a. Response Value: The dataset may only be linked with the approval of the IRB of record *(skip to question 8.13)*
    b. Response Value: The dataset may only be linked with the approval of data contributing sites *(skip to question 8.13)*
    c. Response Value: The dataset may only be linked using a specific linkage method
    d. Response Value: The dataset may only be linked for specific types of research or use
    e. Response: [type a value] *(skip to question 8.13)*
8.11. Question: Enter the type of research or use
    a. Response: [type a value]
8.12. Question: Select or enter the linkage method [select one]
    a. Response Value: PPRL
    b. Response Value: Clear text
    c. Response: [type a value]
8.13. Question: Does the data submission agreement or institutional certification permit dataset sharing? [select one]
    a. Response Value: Yes *(skip to question 8.17)*
    b. Response Value: Yes, with conditions
    c. Response Value: No *(skip to question 8.17)*
    d. Response Value: I don't know *(skip to question 8.17)*
    e. Response Value: It doesn't say *(skip to question 8.17)*
8.14. Question: Select the dataset sharing conditions that the data submission agreement or institutional certification applies [select one]
    a. Response Value: The dataset may only be shared as a de-identified dataset *(skip to question 8.17)*
    b. Response Value: The dataset may only be shared following the Safe Harbor de-identification method *(skip to question 8.17)*
    c. Response Value: The dataset may only be shared if approved by a review body
    d. Response Value: The dataset may only be shared as a limited dataset *(skip to question 8.17)*
    e. Response Value: The dataset may only be shared within a defined data release process
    f. Response: [type a value] *(skip to question 8.17)*
8.15. Question: Enter the name of the review body
    a. Response: [type a value]
8.16. Question: Describe the data release process
    a. Response: [type a value]

8.17. Question: Does the data submission agreement or institutional certification permit secondary dataset access? [select one]
   a. Response Value: Yes *(skip to question 8.19)*
   b. Response Value: Yes, with conditions
   c. Response Value: No *(skip to question 8.19)*
   d. Response Value: I don't know *(skip to question 8.19)*
   e. Response Value: It doesn't say *(skip to question 8.19)*

8.18. Question: Select the secondary dataset access conditions that the data submission agreement or institutional certification applies [select one]
   a. Response Value: The dataset may only be accessed in a data enclave
   b. Response Value: The dataset may only be accessed in a controlled environment
   c. Response: [type a value]

8.19. Question: Does the data submission agreement or institutional certification permit secondary dataset use? [select one]
   a. Response Value: Yes *(skip to question 8.22)*
   b. Response Value: Yes, with conditions
   c. Response Value: No *(skip to question 8.22)*
   d. Response Value: I don't know *(skip to question 8.22)*
   e. Response Value: It doesn't say *(skip to question 8.22)*

8.20. Question: Select the secondary dataset use conditions that the data submission agreement or institutional certification applies [select one]
   a. Response Value: This dataset may only be used for an approved purpose
   b. Response Value: This dataset may only be used for research on a specific topic *(skip to question 8.22)*
   c. Response: [type a value] *(skip to question 8.22)*

8.21. Question: Enter the approved purpose
   a. Response: [type a value]

8.22. Question: Select or enter the prohibitions [select one or more]
   a. Response Value: Individuals in the dataset may not be re-identified
   b. Response Value: The dataset may not be linked
   c. Response Value: The dataset may not be shared
   d. Response Value: The dataset may not be accessed for secondary purposes
   e. Response Value: The dataset may not be used for secondary purposes
   f. Response Value: The dataset may not be used for commercial purposes
   g. Response Value: The dataset may not be used for any purposes beyond explicit permissions
   h. Response Value: The dataset may not be sold
   i. Response: [type a value]

Section 9: Other Governance Policies

9.1. Are there other agreements, certifications, contracts, determinations, policies, or processes that have rules about dataset linkage, sharing, or secondary access and use? [select one]

a. Response Value: Yes

b. Response Value: No *(skip to question 10.1)*

c. Response Value: I don't know *(skip to question 10.1)*

9.2. Question: Enter the name of the policy

a. Response: [type a value]

9.3. Question: Select the type of document or source for governance information [select one]

a. Response Value: Agreement

b. Response Value: Certification

c. Response Value: Contract (

d. Response Value: Determination

e. Response Value: Policy

f. Response Value: Process

g. Response: [type a value]

9.4. Question: Select or enter the assignee party [select one]

a. Response Value: Coordinating center

b. Response Value: Data provider

c. Response Value: Data requester

d. Response Value: Principal investigator

e. Response Value: Repository

f. Response: [type a value]

9.5. Question: Select or enter the assigning party [select one]

a. Response Value: Coordinating center

b. Response Value: Data provider

c. Response Value: Government organization

d. Response Value: Principal investigator

e. Response Value: Repository

f. Response: [type a value]

9.6. Question: Select or enter the assigner [select one]

a. Response Value: Certification organization

b. Response Value: Government organization

c. Response Value: Data provider

d. Response Value: Data coordinating center

e. Response Value: Principal investigator

f. Response: [type a value]

9.7. Question: Select or enter the assignee [select one]

a. Response Value: Data coordinating center
b. Response Value: Data provider
c. Response Value: Data repository
d. Response Value: Government organization
e. Response Value: Principal investigator
f. Response: [type a value]

9.8. Question: Select or enter the contracted party [select one]

a. Response Value: Data coordinating center
b. Response Value: Data provider
c. Response Value: Data requester
d. Response Value: Data repository
e. Response Value: Principal investigator
f. Response: [type a value]

9.9. Question: Select or enter the contracting party [select one]

a. Response Value: Data coordinating center
b. Response Value: Data provider
c. Response Value: Data repository
d. Response Value: Government organization
e. Response Value: Principal investigator
f. Response: [type a value]

9.10. Question: Select or enter the party that is receiving this determination [select one]

a. Response Value: Data coordinating center
b. Response Value: Data provider
c. Response Value: Data requester
d. Response Value: Data repository
e. Response Value: Principal investigator
f. Response: [type a value]

9.11. Question: Select or enter the party that is issuing this determination [select one]

a. Response Value: Data provider
b. Response Value: Government organization
c. Response Value: IRB
d. Response Value: Privacy board
e. Response Value: Data repository
f. Response Value: Review committee

g. Response Value: Data coordinating center
h. Response: [type a value]

9.12. Question: Select or enter the party that is subject to this policy [select one]
a. Response Value: Coordinating center
b. Response Value: Data provider
c. Response Value: Data requester
d. Response Value: Government organization
e. Response Value: Principal investigator
f. Response Value: Repository
g. Response Value: Not applicable
h. Response: [type a value]

9.13. Question: Select or enter the party that is defining or enforcing the policy [select one]
a. Response Value: Coordinating center
b. Response Value: Data provider
c. Response Value: Government organization
d. Response Value: Repository
e. Response Value: Review committee
f. Response: [type a value]

9.14. Question: Select or enter the party that carries out this process [select one]
a. Response Value: Data coordinating center
b. Response Value: Data provider
c. Response Value: Data repository
d. Response Value: Government organization
e. Response Value: Principal investigator
f. Response: [type a value]

9.15. Question: Select or enter the party that carries out this process [select one]
a. Response Value: Data coordinating center
b. Response Value: Data provider

9.16. Response Value: Data requester
a. Response Value: Data repository
b. Response Value: Principal investigator
c. Response: [type a value]

9.17. Question: Does this policy permit dataset linkage? [select one]
a. Response Value: Yes
b. Response Value: Yes, with conditions
c. Response Value: No

d. Response Value: I don't know
e. Response Value: It doesn't say

9.18. Question: Select the data linkage conditions that this policy applies [select one or more]
a. Response Value: The dataset may only be linked with the approval of the IRB of record
b. Response Value: The dataset may only be linked with the approval of data contributing sites
c. Response Value: The dataset may only be linked using a specific linkage method
d. Response Value: The dataset may only be linked for specific types of research or use
e. Response: [type a value]

9.19. Question: Select or enter the linkage method [select one]
a. Response Value: PPRL
b. Response Value: Clear text
c. Response: [type a value]

9.20. Question: Enter the specific types of research or use
a. Response: [type a value]

9.21. Question: Does this policy permit dataset sharing? [select one]
a. Response Value: Yes
b. Response Value: Yes, with conditions
c. Response Value: No
d. Response Value: I don't know
e. Response Value: It doesn't say

9.22. Question: Select the dataset sharing conditions that this policy applies [select one or more]
a. Response Value: The dataset may only be shared as a de-identified dataset
b. Response Value: The dataset may only be shared following the Safe Harbor de-identification method
c. Response Value: The dataset may only be shared if approved by a review body
d. Response Value: The dataset may only be shared as a limited dataset
e. Response Value: The dataset may only be shared within a defined data release process
f. Response: [type a value]

9.23. Question: Describe the review body needed for approval for sharing
a. Response: [type a value]

9.24. Question: Describe the data release process
a. Response: [type a value]

9.25. Question: Does this policy permit secondary dataset access? [select one]
a. Response Value: Yes
b. Response Value: Yes, with conditions
c. Response Value: No

d. Response Value: I don't know
e. Response Value: It doesn't say

9.26. Question: Select the secondary dataset access conditions that this policy applies [select one or more]
a. Response Value: The dataset may only be accessed in a data enclave
b. Response Value: The dataset may only be accessed in a controlled environment
c. Response: [type a value]

9.27. Question: Does this policy permit secondary dataset use? [select one]
a. Response Value: Yes
b. Response Value: Yes, with conditions
c. Response Value: No
d. Response Value: I don't know
e. Response Value: It doesn't say

9.28. Question: Select the secondary dataset use conditions that this policy applies [select one or more]
a. Response Value: This dataset may only be used for the approved purpose
b. Response Value: This dataset may only be used for research on a specific topic
c. Response: [type a value]

9.29. Question: Enter the approved use
a. Response: [type a value]

9.30. Question: Select or enter the prohibitions [select one or more]
a. Response Value: Individuals in the dataset may not be re-identified
b. Response Value: The dataset may not be linked
c. Response Value: The dataset may not be shared
d. Response Value: The dataset may not be accessed for secondary purposes
e. Response Value: The dataset may not be used for secondary purposes
f. Response Value: The dataset may not be used for commercial purposes
g. Response Value: The dataset may not be used for any purposes beyond explicit permissions
h. Response Value: The dataset may not be sold
i. Response: [type a value]

Section 10: Laws

10.1. Question: Do local, state, or federal laws apply to this dataset? [select one]
a. Response Value: Yes
b. Response Value: No
c. Response Value: I don't know

10.2. Question: Select the applicable federal laws [select one]
a. Response Value: CIPSEA: Confidential Information Protection and Statistical Efficiency

b. Response Value: FERPA: Family Educational Rights and Privacy Act
c. Response Value: HIPAA Privacy Rule
d. Response Value: The Common Rule: 45 CFR 46 Part A
e. Response Value: The Federal Privacy Act (of 1974)
f. Response Value: The Public Health Services Act

10.3. Question: Select or enter the CIPSEA rules that apply to the dataset [select one or more]
a. Response Value: Prohibits participant re-identification
b. Response Value: Permits data sharing with approved researchers who are designated agents
c. Response Value: Guarantees confidentiality of participants and contributing organizations
d. Response Value: Permits secondary data use for research
e. Response Value: Penalizes confidentiality violations
f. Response Value: Supersedes the Freedom of Information Act (FOIA) such that data collected under CIPSEA are not subject to FOIA
g. Response: [type a value]

10.4. Question: Select or enter the FERPA rules that apply to the dataset [select one or more]
a. Response Value: Permits sharing identified data for approved purposes [excludes research]
b. Response Value: Permits sharing of de-identified data
c. Response: [type a value]

10.5. Question: Select or enter the HIPAA rules that apply to the dataset [select one or more]
a. Response Value: Permits sharing of a deidentified dataset for research purposes
b. Response Value: Permits sharing of limited dataset for research without participant consent by entering into a data use agreement with a recipient organization
c. Response Value: Defines a limited dataset as a dataset where 16 categories of direct identifiers have been removed
d. Response Value: Defines Safe Harbor and Expert Determination as the allowed methods for deidentification
e. Response Value: Permits use of limited dataset for research without participant consent by entering into a data use agreement with a recipient
f. Response Value: Permits data use for research with either 1) individual authorization (consent) or 2) IRB or a Privacy Board approval of waiver of consent requirement
g. Response Value: Data de-identified according to HIPAA standards are no longer subject to HIPAA and can be used for research without participant consent
h. Response: [type a value]

10.6. Question: Select the HIPAA designated type of dataset [select one]
a. Response Value: Deidentified dataset
b. Response Value: Limited dataset
c. Response Value: Fully identified dataset

d. Response Value: I don't know

10.7. Question: Select or enter the Common Rule rules that apply to the dataset [select one or more]

a. Response Value: Permits sharing of de-identified data

b. Response Value: Permits human subjects research use of data with PII with either 1) participant consent or 2) IRB or a Privacy Board approval of consent waiver

c. Response: [type a value]

10.8. Question: Select or enter the Privacy Act of 1974 rules that apply to the dataset [select one or more]

a. Response Value: Permits collection of data that includes PII by federal agencies that have published a system of records notice (or "SORN") in the Federal Register

b. Response Value: Permits sharing of data with PII if federal agencies take the data into their SORN

c. Response Value: Permits federal agency use of the data with PII if the federal agencies take the data into their SORN

d. Response: [type a value]

10.9. Question: Select or enter the Public Health Services Act rules that apply to the dataset [select one or more]

a. Response Value: Permits data collection

b. Response Value: Permits data sharing with approved researchers who are designated agents

c. Response Value: Permits data use for purposes described in participant consent

d. Response: [type a value]

10.10. Question: Do other laws apply to this dataset? [select one]

a. Response Value: Yes

b. Response Value: No

c. Response Value: I don't know

10.11. Question: What is the name of the law?

a. Response: [type a value]

10.12. Question: Does the law permit dataset linkage? [select one]

a. Response Value: Yes

b. Response Value: Yes, with conditions

c. Response Value: No

d. Response Value: I don't know

e. Response Value: It doesn't say

10.13. Question: Select the dataset linkage conditions that this law applies [select one or more]

a. Response Value: The dataset may only be linked with the approval of the IRB of record

b. Response Value: The dataset may only be linked with the approval of data contributing sites

c. Response Value: The dataset may only be linked using a specific linkage method

d. Response Value: The dataset may only be linked for specific types of research or use
e. Response: [type a value]

10.14. Question: Select or enter the linkage method [select one]
a. Response Value: PPRL
b. Response Value: Clear text
c. Response: [type a value]

10.15. Question: Describe the specific types of research or use
a. Response: [type a value]

10.16. Question: Does the law permit dataset sharing? [select one]
a. Response Value: Yes
b. Response Value: Yes, with conditions
c. Response Value: No
d. Response Value: I don't know
e. Response Value: It doesn't say

10.17. Question: Select the dataset sharing conditions that this law applies [select one or more]
a. Response Value: The dataset may only be shared as a de-identified dataset
b. Response Value: The dataset may only be shared following the Safe Harbor de-identification method
c. Response Value: The dataset may only be shared if approved by a review body
d. Response Value: The dataset may only be shared as a limited dataset
e. Response Value: The dataset may only be shared within a defined data release process
f. Response: [type a value]

10.18. Question: Describe the data release process
a. Response: [type a value]

10.19. Question: Enter the name of the review body
a. Response: [type a value]

10.20. Question: Does the law permit secondary dataset access? [select one]
a. Response Value: Yes
b. Response Value: Yes, with conditions
c. Response Value: No
d. Response Value: I don't know
e. Response Value: It doesn't say

10.21. Question: Select the secondary dataset access conditions that this law applies [select one or more]
a. Response Value: The dataset may only be accessed in a data enclave
b. Response Value: The dataset may only be accessed in a controlled environment

c. Response: [type a value]

10.22. Question: Does the law permit secondary dataset use? [select one]

a. Response Value: Yes
b. Response Value: Yes, with conditions
c. Response Value: No
d. Response Value: I don't know
e. Response Value: It doesn't say

10.23. Question: Select the secondary dataset use conditions that this law applies [select one or more]

a. Response Value: This dataset may only be used for the approved purpose
b. Response Value: This dataset may only be used for research on a specific topic
c. Response: [type a value]

10.24. Question: Enter the approved purpose

a. Response: [type a value]

10.25. Question: Select or enter the prohibitions [select one or more]

a. Response Value: Individuals in the dataset may not be re-identified
b. Response Value: The dataset may not be linked
c. Response Value: The dataset may not be shared
d. Response Value: The dataset may not be accessed for secondary purposes
e. Response Value: The dataset may not be used for secondary purposes
f. Response Value: The dataset may not be used for commercial purposes
g. Response Value: The dataset may not be used for any purposes beyond explicit permissions
h. Response Value: The dataset may not be sold
i. Response: [type a value]

Section 11: Other Governance Information

11.1. Question: Is there any other governance information that applies to this dataset and could not be entered above? [select one]

a. Response Value: Yes
b. Response Value: No

11.2. Question: Enter the governance information related to dataset linkage, sharing, and secondary access and use

a. Response: [type a value]

References

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

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

[3] ASPE Office of the Secretary Patient-Centered Outcomes Research Trust Fund, OS-PCORTF Strategic Plan for 2020-2029: https://aspe.hhs.gov/collaborations-committees-advisory-groups/os-pcortf/os-pcortf-strategic-plan-2020-2029

[4] 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

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

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

[7] 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

[8] 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

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

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

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

[12] 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.

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

[14] Venkatesh, Viswanath and Thong, James Y.L. and Xu, Xin, 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

[15] REDCap website: https://www.project-redcap.org/

[16] Fast Healthcare Interoperability Resource Structured Data Capture website: https://build.fhir.org/ig/HL7/sdc/

[17] Open Data Kit website: https://getodk.org/

[18] Kobo Toolbox website: https://www.kobotoolbox.org/

[19] HTML Javascript website: https://html.spec.whatwg.org/multipage/

[20] S2O website: https://sourceforge.net/projects/s2o/

[21] APCDR EQ website: https://github.com/apcdr/questionnaire

[22] 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

[23] NLM Form Builder website: https://lhcformbuilder.nlm.nih.gov/

[24] LHC-Forms Widget Website: https://lhncbc.github.io/lforms/

[25] React Website: (https://react.dev/)

[26] Kushniruk, A. W., & Patel, V. L. (2004). Cognitive and usability engineering methods for the evaluation of clinical information systems. *Journal of biomedical informatics*, 37(1), 56–76. https://doi.org/10.1016/j.jbi.2004.01.003

[27] Venkatesh, Viswanath and Thong, James Y.L. and Xu, Xin, 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

[28] Open Digital Rights Language Model: https://www.w3.org/TR/odrl-model/