Governance Metadata Standards: Landscape and Gap Analysis

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

Prepared for:

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

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

Prepared by:

MITRE | CMS Alliance to Modernize Healthcare

CMS Alliance to Modernize Healthcare (The Health FFRDC)

A Federally Funded Research and Development Center

February 2024

Department of Health and Human Services, National Institutes of Health, *Eunice Kennedy Shriver* National Institute for 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 wellbeing, and reduce societal costs associated with illness and disability. NICHD's mission is to lead research and training to understand human development, improve reproductive health, enhance the lives of children and adolescents, and optimize abilities for all.

NICHD Office of Data Science and Sharing (ODSS)

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

- Develop a diverse, secure, and interoperable research data ecosystem
- Advise on best practices for data collection, standards, management, sharing, and use across the research and funding lifecycles
- Advance scientific discovery in support of NICHD's mission to understand human development, improve reproductive health, enhance the lives of children and adolescents, and optimize abilities for all

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

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

NIH NICHD Office of Data Science and Sharing 31 Center Drive, Bldg. 31, Rm. 2A03, Bethesda, MD, 20892

Rebecca Rosen PhD, Director rebecca.rosen@nih.gov

Citation

CMS Alliance to Modernize Healthcare (The Health FFRDC). Data Governance Metadata Standards: Landscape and Gap Analysis. Prepared under Contract No. 75N94023F00171. February 2024.

Authors

Emily Kraus, PhD, MPH
Gregory Shemancik, MHA
Tomasz Adamusiak, MD, PhD
Callie St. Claire Phillips, MS
Katherine K. Kim, PhD, MPH, MBA, FAMIA
Susan C. Hull, MSN, RN, NI-BC, NEA-BC, FAMIA
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, to whom we extend our sincere appreciation.

This report was prepared by 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

Contents

Figures

Tables

Executive Summary

Individual-level linkages[a] between biomedical study datasets and the U.S. Department of Health and Human Services (HHS) administrative and survey datasets provide opportunities to maximize the value of existing data. Linkage enables researchers to deduplicate participants across datasets, introduce new variables into analysis plans, reduce costly redundancies in data generation, perform longitudinal analysis, and ask new scientific questions of the enriched dataset. To appropriately link datasets, researchers and data stewards must understand the consent, policy, regulatory, and/or other legal frameworks that apply to each of the original datasets and how the resulting linked dataset inherits rules and controls from these frameworks. They must also understand if and how new limitations arise that impact the sharing and use of the resulting linked dataset; for example, a need to implement new rules and controls to mitigate increased risk of participant identifiability. The collective set of these rules and controls is referred to as data governance, and it defines and enforces appropriate collection, sharing, access, linking, and use of the data, across the data lifecycle. The standardization of data governance information about datasets will help researchers and data stewards determine whether multiple datasets can be linked, and, if so, what data governance applies to the linked dataset.

Metadata related to data governance to support decision making around dataset linkage is widely acknowledged as necessary to enable knowledge discovery from and responsible use of existing data. These metadata should address information related to appropriate data linkage, sharing, and use; provenance; roles and responsibilities of stakeholders; and decision making.

The National Institutes of Health (NIH) has identified the need to improve efficiency and harmonization among controlled-access repositories to make NIH data more findable, accessible, interoperable, and reusable (FAIR) and to ensure appropriate oversight when data from different resources are combined. The *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) Office of Data Science and Sharing (ODSS) is leading an effort to assess the usage of privacy preserving record linkage for pediatric patient-centered outcomes research, with a focus on pediatric Coronavirus disease 2019 (COVID-19) research with support from HHS Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF) and the NIH Office of Data Science Strategy (NIH ODSS). During foundational governance work that informs the current project, NICHD ODSS uncovered a rich and complex governance information ecosystem, for which no data governance metadata schema or user tools exists.

NICHD ODSS has engaged the Health federally funded research and development center (Health FFRDC), operated by The MITRE Corporation, to develop a robust metadata schema for data governance information relevant to linking individual-level participant data and sharing and using linked datasets. Both structured and unstructured text, often stored in a narrative format within policy documents, data use agreements, and consent forms, must be annotated through the application of a data governance

[a] Data linkage is defined by the Agency for Healthcare Research and Quality as “combining information from a variety of data sources for the same individual”; in the context of this report, it is synonymous with individual-level dataset linkage.

metadata schema.[b] Standards, including ontologies, terminologies, vocabularies, schemas, and common data models, are the tools and methods to organize, codify, value, and annotate unstructured governance information with structured governance metadata that may be extensible and machine-readable.

The purpose of the landscape analysis is to identify existing standards that could be used in a data governance metadata schema. The analysis consisted of the development of an inventory of existing standards, an assessment of utility of those standards, and a gap analysis based on governance information domains derived from the NICHD ODSS-developed Governance Information Framework.[1]

A multi-pronged and iterative search yielded 47 standards of which 33 met inclusion criteria of being a standard in use or in development in the United States or abroad that could be applied to any governance information domain or attribute derived from the Governance Information Framework, or data linkage or use concepts discussed in the preceding NICHD ODSS reports.[2,3] The project team did not recommend 20 of these based on assessment of the utility which included criteria related to completeness and community intent, logical consistency and coherence, accessibility, active use and community adoption, and maturity. The project team recommended the remaining 13 standards for use in the data governance metadata schema and included them in the gap analysis. The standards that proceeded to the gap analysis phase are: Data Catalog Vocabulary, Data Documentation Initiative, Data Tags Suite, Data Use Ontology, Dublin Core, Fast Health Information Resources Consent Resource, Fast Health Information Resources Data Segmentation for Privacy and Security Labeling, Informed Consent Ontology, National Cancer Institute Thesaurus, Organization for the Advancement of Structured Information Standards (OASIS) LegalRuleML, Observational Medical Outcomes Partnership Common Data Model, Open Digital Rights Language, and Operational Data Model.

The project team identified gaps by mapping standards and utility assessment findings to the governance information domains. They identified gaps in nine of the 13 governance information domains: Dataset Information, Linkage, Consent, Institutional Review Board, Policy, Rules, Controls, Party, and Data Lifecycle. They determined there are adequate standards to address only four domains: Governing Body, Law (includes Regulations and Statutes), Agreement, and Authorization.

Summary recommendations for the NICHD ODSS data governance metadata schema development focus on the Open Digital Rights Language standard and Fast Health Information Resources (FHIR) Consent information models, with value sets drawn from FHIR terminology and Data Use Ontology. The findings from this landscape analysis, utility assessment, and gap analysis inform this approach. No single standard fully addresses the schema requirements across all governance information domains, necessitating the use of multiple standards and combining elements from various sources, as well as potentially developing new value sets (e.g., to capture linkage metadata). The maturity and licensing of existing standards are also significant factors influencing their utility and adoption.

[b] 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. "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.

The findings from this landscape and gap analysis will support the advancement of a data governance metadata schema. The schema should balance the need for consistency and interoperability with the need for flexibility and adaptability to accommodate evolving research needs and regulatory requirements. Minimizing the number of standards referenced and maximizing coverage of domains allows for a more practical and flexible solution, better management of variability over time, and a better optimized schema. Furthermore, schema development should be guided by a strong commitment to collaboration and engagement with researchers, data providers, and policy makers, to ensure that the schema is both practical and effective in addressing the diverse needs of the research community. The schema will subsequently contribute to NIH-wide strategic goals and activities on Controlled Data Access Coordination.[4]

By adopting a thoughtful and strategic approach to governance metadata schema development, informed by the findings and recommendations presented in this report, the NICHD ODSS can pave the way for a more standardized, efficient, and transparent system of metadata data governance that supports the advancement of research, data sharing and reuse, and innovation in the field. We also hope this report will be useful to researchers generating datasets, data stewards, stakeholders interested in research using linked datasets across HHS agencies and NIH as well as more broadly, and the patient-centered outcomes research community.

The OS-PCORTF funded this project as part of a portfolio of data capacity projects related to patient-centered outcomes research and data linkage, aligned with Goal 2, Data Standards and Linkages for Longitudinal Research, in the next decade's strategic plan.[5]

1 Introduction

1.1 Background

The National Institutes of Health (NIH), a part of the U.S. Department of Health and Human Services (HHS), is the nation's primary medical research agency, making important discoveries that improve health and save lives. NIH is now one of the world's foremost biomedical research agencies and serves as the focal point for biomedical research within the Federal Government. NIH began in 1887, and today comprises 27 separate Institutes and Centers, most of which are located in Bethesda, Maryland. NIH works toward its mission to seek fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability by 1) conducting research in its own laboratories; 2) supporting non-federal scientists at universities, teaching hospitals, and other academic institutions around the world; 3) sponsoring training programs for research investigators; and 4) fostering the communication of research-based health information.

The *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) Office of Data Science and Sharing (ODSS) is leading an effort to assess the usage of privacy preserving record linkage (PPRL) for pediatric patient-centered outcomes research, with a focus on pediatric Coronavirus disease 2019 (COVID-19) research with support from NIH ODSS and HHS Office of the Secretary Patient-Centered Outcomes Research Trust Fund (OS-PCORTF). PPRL holds significant promise for enhancing the value of de novo clinical research data collection, through linkages across different studies and linkages with HHS administrative and survey datasets.

Individual-level dataset linkages[c] could enable researchers to deduplicate subjects across studies, introduce new variables into analysis plans, and reduce costly redundancies in the generation of genomic sequencing data. In order for individual-level datasets to be linked using PPRL or any other linkage method, however, researchers and data stewards must ensure that the linkages are appropriate, based on factors such as if or how the data were consented for use by the research participant, whether the scope of linkage encompasses other data sources, and if there are regulatory and/or legal frameworks that apply to the use of the data. It is important to understand how the resulting linked dataset inherits rules and controls that are associated with the original datasets that contribute to the linkage and if new limitations arise; for example, to address increased identifiability of linked data.

NICHD ODSS is developing a robust metadata schema for data governance information relevant to linking individual-level participant data and sharing and using linked datasets. This effort aligns with NICHD ODSS's larger goal of developing a governance and technology strategy for implementing individual-level record linkage for pediatric research, driven by pediatric COVID-19 research use cases.

[c] Data linkage is defined by the Agency for Healthcare Research and Quality as "combining information from a variety of data sources for the same individual"; in the context of this report, it is synonymous with individual-level dataset linkage.

The NICHD ODSS-developed data governance metadata schema[d] will contribute to NIH-wide strategic goals and activities on Controlled Data Access Coordination (CDAC).[6] To serve NIH-wide priorities, NICHD ODSS will develop a prototype data governance metadata search and visualization tool and underlying database that will inform researchers how datasets of interest can be linked and used. The overall goal is to provide researchers and other stakeholders 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.

To develop a robust data governance metadata infrastructure, unstructured text, often stored in a narrative format within policy documents, data use agreements, and consent forms, must be annotated with structured data through the application of a data governance metadata schema. Standards, including ontologies, terminologies, vocabularies, schemas, common data models and taxonomies, are the tools and methods to organize, codify, value, and annotate unstructured governance information with structured governance data that may be extensible and machine-readable. Future efforts may enrich the data governance metadata schema with a rules engine and automation tools that call on the structured values within the metadata schema.

NICHD engaged the Health federally funded research and development center (Health FFRDC), operated by The MITRE Corporation, to support NICHD ODSS to conduct a landscape assessment and gap analysis of existing metadata standards. In preparation for this work, the Health FFRDC, under the oversight of the NICHD ODSS project leadership team, engaged community experts in the form of a Technical Experts Panel (TEP) to draw on their expertise to guide this landscape analysis and subsequent efforts to develop the data governance metadata schema. See Appendix A for TEP Membership.

1.2 Purpose

The purpose of this Governance Metadata Standards Landscape and Gap Analysis is to identify existing standards that could be used in a data governance metadata schema. For this analysis, standards are defined as ontologies, terminologies, vocabularies, schemas, common data models, and taxonomies. The project team conducted the analysis through the development of governance information domains and attributes, compilation of an inventory of existing standards, an assessment of utility of those standards, and an analysis of gaps related to the governance information domains. The findings of this landscape analysis will inform the development of an extensible and machine-readable schema for the standardized collection and exchange of data linkage and use governance metadata that has been derived from participant assent or consent, regulatory or other policy requirements, or any other agency, study, system, or participant determinations.

[d] 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. “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.

1.3 Audience

The intended audience of this public report includes: 1) researchers generating datasets from a study or program that are or could be linked; 2) stewards of data repositories who accept and expose metadata for datasets they host; 3) stakeholders including policy makers considering research that involves record linkage and those interested in the ontologies, terminologies, or standards that may be useful in the collection and use of governance metadata; 4) the broader metadata and standards community; 5) the patient-centered outcomes research community; and 6) researchers and data scientists across HHS and NIH agencies.

1.4 Foundational Governance Work

This landscape analysis builds on deep data governance discovery work that occurred over 2022 and 2023, led by NICHD ODSS and culminating in two reports described below and a Governance Information Framework.[7]

Privacy Preserving Record Linkage (PPRL) for Pediatric COVID-19 Studies Report[8]

Published in September 2022, this report (hereafter referred to as "the 2022 Report") aimed to inform an NIH-wide strategy on the use of PPRL for pediatric COVID-19 studies. The project assessed 13 existing record linkage implementations and developed technical and governance considerations for appropriately linking data. The 2022 Report summarizes the current state of pediatric COVID-19 studies that could benefit from use of PPRL, 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.

PCORTF Pediatric Record Linkage Governance Assessment[9]

Following the exploration of what governance information was necessary to make a determination about the ability to conduct linkage and the subsequent limitations and controls that would apply to such a linkage, NICHD ODSS sought to collect and examine the governance information from 11 HHS and other federally funded datasets that represent three theoretical pediatric COVID-19 research use cases. This 2023 report (hereafter referred to as "the 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 of a standardized and machine-readable data governance metadata schema. The collected governance information was documented in an NICHD ODSS-developed Governance Information Framework. Based on this dataset governance assessment, the 2023 Report provides considerations for the development and implementation of a data governance metadata schema, including:

- Publicly sharing the data governance information specified by the schema in a predictable and easy-to-find location will facilitate the ability to create linked datasets
- Publicly shared data governance information, and the associated schema, should:
    - Explicitly describe whether linkage is permissible for a given dataset and, if so, include general guidance for what types of linkages are allowed or prohibited, and what rules and controls the linked data would inherit from the individual dataset

    - Incorporate the provenance of data governance origins including authorizations for data collection, linking, sharing, access, and use as well as applicable laws, regulations, and policies
    - Capture the roles and responsibilities of the multiple stakeholders involved in implementing data governance across the data lifecycle
    - Incorporate information regarding decisions made for previous and new linkages involving a given dataset to communicate appropriate linkage of the data and to inform future linkage involving the same dataset; this information may streamline decision making when linkage governance is not explicitly specified by any dataset governance source
- The schema should describe data governance in a standard way to facilitate human interpretation and machine-readability, which in turn promotes adherence
- A concerted effort is required to encourage adoption of the schema across federal and other health agencies that generate datasets that could be linked and used by researchers

In the process of collecting governance information, NICHD ODSS uncovered a rich and complex governance information ecosystem, for which no data governance metadata schema or user tools exist. 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

This report takes these points into further consideration.

1.5 State of the Science of Data Governance Metadata

Metadata related to data governance is widely acknowledged as necessary to enable knowledge discovery from existing datasets. Both federal and private sector calls for data sharing and open science identify the need for governance metadata.[10,11] These metadata should address information related to appropriate rules for data linkage, sharing, access, and use; provenance; roles and responsibilities of stakeholders; and decision-making processes. The FAIRsharing inventory[12] reveals numerous metadata standards developed or in use across thousands of primarily community-driven data repositories. These standards include checklists, terminologies, ontologies, and formats or syntax. Yet relatively few governance-specific metadata standards or models have proved to be interoperable in practice. A recent systematic review focused on metadata related to the health/clinical domain, which is the context of this landscape analysis, found only seven papers that described data linkage and reuse among the 80 papers reviewed.[13] The review found remaining critical gaps in representation of administrative metadata that provides information about provenance, reasons for the study, and other context, in comparison to more prevalent technical metadata about the measures or data elements themselves. A recent report compared data governance of NIH-supported platforms that share genetic/genomic data.[14] The paper identified governance functionality in these platforms including data submission, data ingestion, user authentication and authorization, data security, data access, auditing, and sanctions, offering another perspective of gaps in standardized terminology for data governance processes and procedures.[15] To progress the field, governance metadata must become easier to share by data providers, annotate by data curators, and use by researchers.

Alignment with CDAC and NIH Data Access Goals

NIH has identified the need to improve efficiency and harmonization among controlled-access repositories to make NIH data more findable, accessible, interoperable, and reusable (FAIR)[16] and to ensure appropriate oversight when data from different resources are combined. Toward addressing this need, NIH established an internal CDAC working group in 2021 that delivered a series of recommendations for implementation in 2022 and onward. Implementation of these recommendations would require a harmonized approach to collecting, exchanging, and visualizing information about controlled-access data governance.

CDAC aims 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.

The ultimate goal of this NICHD ODSS governance work is to streamline the appropriate access to and use of federal patient-centered outcomes datasets by developing a metadata schema to facilitate decision making relevant to individual-level record linkage, sharing, and research use.

2 Approach

A landscape analysis is a set of general approaches used to scan the field and provide early and high-level findings to identify, organize, and evaluate the state of affairs related to a particular issue. The approach can include both specified methods and ad hoc steps for a complete analysis. The project team that conducted this governance metadata landscape analysis included the Health FFRDC and NICHD ODSS, with guidance through consultation with the TEP.

The project team's approach to this landscape analysis of metadata governance standards included:

- Review of recent work conducted by NICHD ODSS, including the NICHD ODSS-developed Governance Information Framework, to define the scope of the analysis
- Identification of potential candidate standards
- Development of utility assessment and gap analysis methodology
- Development of a data collection instrument to capture the standards inventory and results of the utility assessment
- Organization of standards by relevant characteristics
- Assessment of standards relative to the purpose of the project

2.1 Methods

The NICHD ODSS-developed Governance Information Framework[17] served as the foundation for the landscape analysis, providing the structure to understand the scope of standards and analyze the gaps. The project team applied categories defined by the Governance Information Framework to organize data governance metadata standards into applicable domains and describe the attributes of each standard associated with those domains. This logical model breaks down a complex topic into tangible segments enabling both the researcher and reader to understand a standard's applicability to each domain and any resulting gaps in standards for that domain.

Defining Data Governance Information Domains and Related Attributes

The project team defined an analytic structure of data governance information domains and child attributes, derived from the NICHD ODSS-developed Governance Information Framework. Through an interactive collaboration process, the project team further refined definitions of the domains and attributes. This structure helped inform search activities, characterization of standards, utility assessment, and gap analysis.

Search for Relevant Standards and Determining Relevance

The project team searched for data governance-relevant candidate standards and documented an initial list in a shared workspace, in this case an Excel standard inventory specifically designed to list, define, and categorize each standard. The TEP confirmed that the project team's inventory of standards was comprehensive for the purpose of supporting the use cases described in the 2023 Report and the data governance metadata schema development. The TEP also provided additional standards, literature, and relevant project recommendations.

Description of Standards

The project team assigned reviewers to each standard. Reviewers collected website links from open resources from the internet that were readily available, placing these within the standards inventory collection instrument and adding descriptions of each standard.

Development of Inclusion and Exclusion Criteria

The project team defined inclusion and exclusion criteria that were affirmed by the TEP. Reviewers used descriptions to determine which inclusion and exclusion criteria either 1) likely applied, 2) possibly applied with further review necessary, or 3) likely did not apply, and color coded each standard accordingly. The project team discussed each standard as a team, to ensure that there was consensus on the color coding and that resulting inclusions and exclusions were accurate.

Utility Assessment and Recommended Standards

The project team conducted the utility assessment. The team scrutinized the included standards in greater detail using utility criteria to determine whether or not the standard was to be recommended for use in the data governance metadata schema. Standards the team did not recommend were removed from the subsequent gap analysis.

Gap Analysis

The project team mapped the recommended standards to associated governance information domains. This allowed the team to identify which domains (and attributes as applicable) are sufficiently addressed by standards available and which are not (i.e., a gap). The team documented results from this exercise by governance information domain.

2.2 Defining Governance Information Domains and Attributes

The NICHD ODSS-developed Governance Information Framework served as the foundation for defining a structure of governance information domains and attributes that the project team used to categorize each standard based on what type of governance information the standard could annotate. The governance information domains and attributes are specified in Table 1.

Table 1. Governance Information Framework Domains and Attributes

Domain	Attribute	Governance Attribute Description
Dataset Information	Dataset Name	Dataset source name
Dataset Information	Dataset Source	Dataset source agency
Dataset Information	Data Type	Data type (clinical, survey, genomics, etc.)
Dataset Information	Dataset Point of Contact	Dataset, repository, or other point of contact
Dataset Information	Dataset Granularity	Dataset level of aggregation (individual level or aggregate)
Dataset Information	Dataset Special Population	Dataset includes special populations such as tribal populations, minors, or pregnant women
Dataset Information	Dataset Common Data Model	Use of common data model, if any, for data collection
Linkage	Personally Identifiable Information Present	Personally identifiable information elements collected
Linkage	Personally Identifiable Information Holder	Personally identifiable information elements holder (i.e., party that holds the identifiers)
Linkage	Past Linkage	Has this dataset been linked with other datasets?
Linkage	Linked Dataset	Name of other linked dataset
Linkage	Linked Dataset Type	Other dataset type (e.g., clinical, survey, claims)
Linkage	Linked Dataset Source	Other dataset source(s)

Domain	Attribute	Governance Attribute Description
Linkage	Linkage Method	Linking methodology and technology
Linkage	Linkage Personally Identifiable Information	Personally identifiable information elements used for the linkage
Linkage	Entity Resolver	Entity resolver (data originator or data linker or third party)
Linkage	Linkage Entity	Party linking the data
Linkage	Linkage Quality	Linkage quality assessment
Linkage	Linkage Sharing Method	Linked data sharing method (linkage maps or pre-linked dataset)
Consent	Consent Waived	Consent waived
Consent	Assent	Assent used
Consent	Assent Contents	Assent contents including permissions for dataset linkage and use
Consent	Consent	Consent used
Consent	Consent Contents	Consent contents including permissions for dataset linkage and use as well as complex consent issues such as tiered consent
Consent	Consent Subgroups	Dataset includes consent subgroups that have data use or linkage implications
Institutional Review Board (IRB)	IRB Entity	IRB of record or non-IRB privacy board
Institutional Review Board (IRB)	IRB Entity Type	IRB or non-IRB privacy board
Institutional Review Board (IRB)	IRB Determination	IRB determination (e.g., full approval of an IRB protocol, IRB waiver of consent, IRB determination that a study is not human subjects research)
Governing Body	Governing Body	Name of organization or group of individuals that has decision making authority about a dataset’s linkage or use
Governing Body	Governing Body Determination	Governance body determination (e.g., full approval of linkage or data use)

Domain	Attribute	Governance Attribute Description
Law (includes Regulations and Statutes)	Law Type	Type of law that applies (e.g., local, state, federal, international, other)
Law (includes Regulations and Statutes)	Law Identification	Name of law and/or statutory number
Law (includes Regulations and Statutes)	Law Content	Contents of applicable laws (sections and meaning)
Agreement	Agreement Type	Agreement type (e.g., data use agreement, network agreement, contract, NIH institutional certification)
Agreement	Agreement Name	Agreement name (e.g., COVID-19 Registry participation agreement)
Agreement	Agreement Content	Agreement content (e.g., data use agreement specifies a disclosure review process)
Policy	Policy Document	Name of policy document that may be the origin or source of authorizations, controls, and rules
Policy	Policy Level	Local, state, tribal, international
Policy	Policy Name	Name of policy within policy document
Policy	Policy Content	Content of applicable policy
Rule	Rule	Define what must occur or not occur, including limitations or constraints on how data are handled
Authorization	Authorization Type	Is there an authorization for dataset linkage or use?
Authorization	Authorization Determination	Authorization determination (e.g., this dataset may be linked to other datasets)
Authorization	Authorization Specification	When authorization is present, additional specifications within that authorization
Authorization	Authorization Source	What is the source of this authorization (e.g., consent form, data use agreement, policy document)?

Domain	Attribute	Governance Attribute Description
Controls	Control	Description of control (e.g., dataset may only be linked by a third-party entity resolver, de-identification status, disclosure review, no sharing permitted, access requirements, committee approvals, signed agreements)
Controls	Control Type	Technical or administrative control
Party	Entity	The name of the organization that is a party
Party	Role	Description of the entity's role in governance (data owner, data steward, entity resolver, secondary user, etc.); could also represent how a control or rule applies to an entity
Data Lifecycle	Collection	Obtaining data from participants for research, clinical, or administrative purposes
Data Lifecycle	Linking	Combining information from a variety of data sources for the same individual
Data Lifecycle	Sharing	Making data available to the broader data user community, for example, by submitting the data to a data repository for dissemination
Data Lifecycle	Access	Acquiring data from a data repository or other data sharing system
Data Lifecycle	Use	Working with data for secondary research or other analytical purposes

The project team formed domains as categories, and attributes as subcategories to serve as examples of the types of metadata that would exist within respective information domains. The project team refined the domains and attributes iteratively through discussion to arrive at the set for this analysis. The team also began to identify relationships between domains and attributes (e.g., agreements and laws are the source of authorizations and controls).

Future work to define and test the data governance metadata schema may lead to further refinements.

2.3 Search Strategy

The project team bounded the scope of search efforts according to the team's development of domains and attributes as a standard set of requirements. These requirements were aligned with and according to the three pediatric COVID-19 use cases previously defined in the 2023 Report. The TEP provided feedback on which standards to consider based on their industry and working knowledge.

The search strategy consisted of a multi-pronged and iterative approach to identify candidate standards for consideration. Steps to execution included:

1. The project team conducted a targeted internet search using terms such as “governance,” “metadata,” “consent,” “research standards,” “legal data standards,” etc.
2. The project team identified standards based on knowledge and experience of team members and guidance from the TEP
3. The project team reviewed research projects, consortiums, and broad data sharing initiatives that the TEP identified as additional sources of potential standards

After assembling an initial inventory, the project team compiled basic descriptive information including a narrative description, affiliations with projects or larger standards repositories, an initial assessment of which governance information domains the standard could apply to, and links to relevant resources.

The TEP suggested that the project team explore related projects, consortiums, and initiatives where standards and schemas may exist or be indirectly found. This strategy added to the comprehensive nature of the project team’s review, and the team extracted and reviewed any standards recommended or used by relevant resources and, when applicable to this work, added them to the standards inventory. To ensure a complete review, the project team where applicable also examined communities and projects from which standards originated.

2.4 Inclusion and Exclusion Criteria

The landscape analysis included standards in use or in development in the United States or abroad that could be applied to any governance information domain or attribute (Table 1), or data linkage or use concepts discussed in the preceding NICHD ODSS reports.

Because no well-established methodology exists for determining the scope of governance metadata concepts or data linkage and use concepts, the project team consulted with the TEP about the inclusion and exclusion criteria and came to consensus on the agreed upon criteria.

The project team excluded standards from the landscape analysis if they met any one of the following criteria:

- No recent activity (no posts, activity, or releases in 5 years)
- Inadequate publicly available documentation for use
- No recent evidence of an active user community or support
- No relevance to data governance metadata concepts
- Documentation of an inability to be applied to health or research
- No examples of real-world application
- Formally deprecated or incorporated into other standards
- Discovered to be not a standard

2.5 Utility Assessment

The objective of the utility assessment was to identify the standards from among those included in the standards inventory that might be appropriate for use in the governance metadata schema. The project team developed a qualitative utility assessment approach to evaluate each standard against defined criteria.

To develop the criteria, the project team reviewed literature from the biomedical research and standards community to identify potential concepts applicable to standard evaluation, including accuracy, completeness, coherence, consistency, accessibility, maturity, active use, community adoption, and conformance.[18] The team then drafted a definition for each concept and validated it on two standards. Validation revealed that 1) accuracy was duplicative with completeness, consistency, and coherence, and the team removed this criterion; and 2) conformance to expectations was not able to be applied in a systematic way because the expectations were not well defined for each standard, and thus the team also removed this criterion.

Utility assessment criteria included:

1. **Application:** Does the standard offer terms, attributes, structures, or vocabularies that can be applied to governance information domains and attributes?
2. **Completeness and Community Intent**: Does the standard meet community need with required resources? Do the resources cover all relevant aspects of the domain in question, such as necessary information, examples, and guidelines?
3. **Logical Consistency and Coherence**: Is the standard well-structured, comprehensive, easy to understand, and free from contradictions or ambiguities?
4. **Accessibility**: Is necessary information easy to locate and access, and available for public reference, allowing users to obtain necessary information without difficulty?
5. **Active Use and Community Adoption**: Are there notable instances of active use, such as published references, case studies, or endorsements? Is the standard listed in European Bioinformatics Institute Ontology Lookup Service or National Center for Biomedical Ontology BioPortal repositories?
6. **Maturity**: At what maturity level is the standard currently operating? This approach to examining maturity combines several established methodologies:[19,20,21,22]
   - Maturity Level 1: Informal and Initial
   - Maturity Level 2: Developing
   - Maturity Level 3: Defined and Implemented
   - Maturity Level 4: Managed and Repeatable
   - Maturity Level 5: Integrated and Optimized

The TEP validated and confirmed the utility criteria as being appropriate and acceptable.

Four Health FFRDC team reviewers participated in the utility assessment. The utility assessment followed these steps:

1. Two reviewers were assigned to each standard for trustworthiness and accuracy in the qualitative analysis
2. Reviewers relied on information and resources already collected, along with new resources (links to GitHub or other descriptors) including publicly available information that was accessible by the project team
3. Reviewers considered each of the utility assessment criteria for each standard
4. Reviewers described to what extent criteria were met or not in the data collection instrument and made a recommendation for use in the metadata schema

At least two Health FFRDC reviewers evaluated the utility assessment for each standard, depending on complexity and technical content needed for the review. Reviewers discussed findings and reached consensus to support concurrence across each standard and associated domain.

2.6 Gap Analysis

The project team mapped the recommended standards against the full set of governance information domains (Table 1). Since a standard can apply to multiple domains, reviewers assessed each standard for application to all 13 governance information domains (e.g., dataset information, linkage, consent, IRB). The project team defined a gap as a governance information domain for which they could not identify a standard that met the utility criteria (Section 2.5).

To perform the gap analysis, four Health FFRDC team reviewers participated in the following steps:

1. Each standard was assigned to the same reviewers who conducted the utility assessment with at least two reviewers for each standard.
2. Reviewers considered each standard for application within each governance information domain.
3. When a standard had relevant terms or concepts for any attribute within the governance information domain, reviewers mapped the standard to the domain and noted examples at the attribute level.
4. Reviewers described how the standard could encode domain-specific information in a summary table (Table 4) and examples where the attribute level with references were denoted.
5. The project team validated mappings and examples, and summarized the degree to which all of the domain attributes were covered by one or more standards.

Notably, the work to apply and test standards across all attributes in all 13 governance information domains is a key activity in the development of the data governance metadata schema, the subsequent effort that this report will inform. The findings below present examples of how standards can apply to selected governance attributes but are not comprehensive.

2.7 Limitations

A landscape analysis by definition is not systematic and is most appropriate as an approach to explore a topic that may not have extensive evidence or study resources available. Thus, this analysis has several limitations.

The project team did not employ a systematic approach to literature search. The team may have missed standards that could be applicable to the governance information domains and topics of interest using the search strategy identified above.

Well-validated or tested utility criteria for governance metadata standards and metadata standards in general do not exist. The project team developed utility criteria and then confirmed the criteria with the TEP.

In addition, available documentation for some standards was limited, and the project team made decisions on only readily available public information. Other publications, resources, or websites may exist that explain how an identified standard meets the utility criteria or aligns with a governance information domain or topic. The project team may have found standards to be less applicable or appropriate for use in the metadata schema than they might have upon review of complete information.

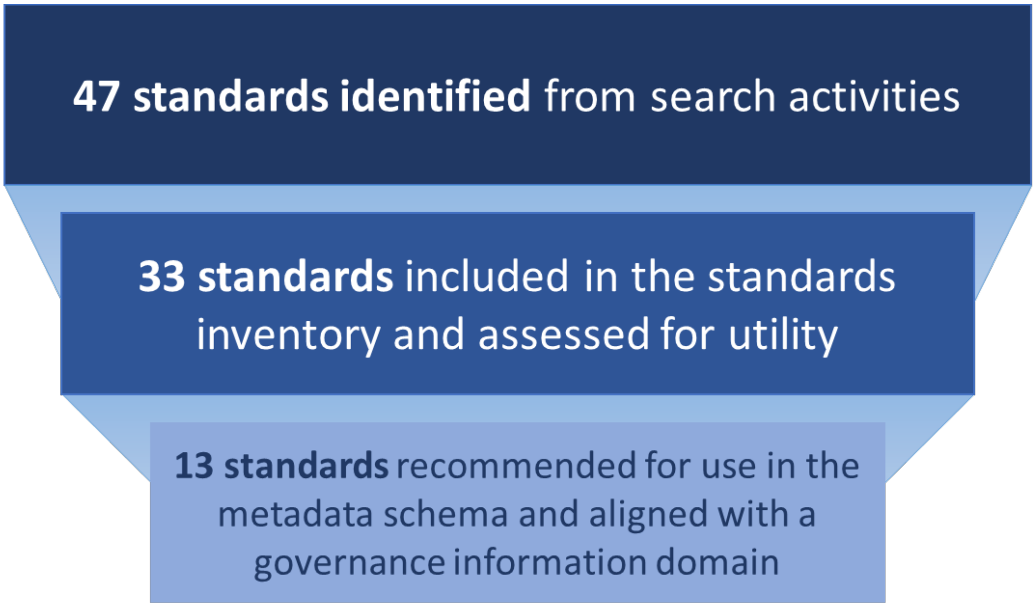
The landscape analysis methodology did not allow for testing the application and extension of identified standards to governance metadata attributes. Thus, recommendations of standards to be used in a metadata schema may be determined to be infeasible or inappropriate once the metadata schema is defined. However, testing of standards will occur during subsequent metadata schema development activities.

The decision to examine standards as the unit of analysis and the varied interpretations of the term standard may have affected the findings. For example, Fast Health Information Resources (FHIR) is a standard and a specific FHIR resource is also a standard. The project team included multiple FHIR resources in the analysis as distinct standards. Their findings may have differed based on an assessment of FHIR as one collective standard and its inherent capabilities.

3 Findings

The landscape assessment identified 47 candidate standards for consideration. The project team excluded 12 standards at the outset based on inclusion and exclusion criteria, resulting in 35 standards included in the initial standards inventory and put through the utility assessment. The project team subsequently discovered another two candidates that were not a standard in practice and removed these from the standards inventory (n=33). Based on the utility assessment, the project team did not recommend 20 standards for use in the data governance metadata schema. From this set, 13 standards were recommended for use and considered in the gap analysis (Figure 1).

Figure 1: Summary of Standards by Landscape Analysis Activity



47 standards identified from search activities

33 standards included in the standards inventory and assessed for utility

13 standards recommended for use in the metadata schema and aligned with a governance information domain

3.1 Inventory of Standards

After excluding 14 standards based on inclusion and exclusion criteria (Table 6), the project team included 33 standards in the standards inventory (Table 2).

Table 2. Standards Inventory

	Standard Name	Governance Information Domain
1	Clinical Data Acquisition Standards Harmonization (CDASH)	Dataset Information
2	Control Objectives for Information and Related Technologies (COBIT)	Policy
3	Datacite 4.3	Dataset Information
4	Datasheets for Datasets	Dataset Information
5	Data Catalog Vocabulary (DCAT)	Dataset Information
6	Data Documentation Initiative (DDI)	Dataset Information
7	Data Tags Suite (DATS)	Dataset Information
8	Data Use Ontology (DUO)	Authorizations, IRB
9	Dublin Core (DCMI)	Dataset Information, Governing Body, Party
10	Extensible Access Control Markup Language (XACML)	Controls, Policy
11	FHIR Consent Resource	Consent
12	FHIR Data Segmentation for Privacy (DS4P) and Security Labeling	Controls, Policy

	Standard Name	Governance Information Domain
13	FHIR Provenance Resource	Data Lifecycle
14	FHIR US CORE	Dataset Information
15	Information Artifact Ontology (IAO)	Dataset Information
16	Informed Consent Ontology (ICO)	Consent, IRB
17	ISO/IEC 38500:2015 – Governance of IT for organization (ISO)	Policy
18	NCI Thesaurus (NCIt)	Dataset Information, Linkage
19	OASIS LegalRuleML TC (LegalRuleML)	Agreement, Policy, Law, Rules, Authorizations
20	Observational Medical Outcomes Partnership Common Data Model (OMOP CDM)	Dataset Information
21	Ontology for Biomedical Investigations (OBI)	Dataset Information
22	Ontology of Information Security (OIS)	Controls
23	Open Digital Rights Language (ODRL)	Dataset Information, Rules, Consent, Governing Body, Law, Policy, and Party
24	OpenAIRE	Dataset Information
25	Operational Data Model (ODM)	Dataset Information
26	Provenance, Authoring, and Versioning (PAV)	Data Lifecycle
27	Provenance Ontology (PROV-O)	Dataset Information
28	Science On Schema.Org (SOSO)	Dataset Information
29	Study Data Tabulation Model (SDTM)	Dataset Information, Data Lifecycle
30	terms4FAIRskills (T4FS)	Dataset Information
31	Unified Medical Language System (UMLS)	Dataset Information
32	US Core Data for Interoperability	Dataset Information
33	Web Access Controls	Controls

3.2 Additional Resources

The project team reviewed 23 projects, consortiums, initiatives, frameworks, and principles to identify additional candidate standards or relevant guidance for the formation of a governance metadata schema. The 17 projects, consortiums, and initiatives and 6 frameworks and principles are:

Projects, Consortiums, and Initiatives

1. Biomedical and Healthcare Data Discovery Index Ecosystem[23]
2. Bioschemas[24]
3. cancer Biomedical Informatics Grid[25]
4. Clinical Data Interchange Standard Consortium[26]
5. Creative Commons Licenses[27]
6. The database of Genotypes and Phenotypes[28]
7. Global Alliance for Genomics and Health Consortium[29]
8. Globus Toolkit[30]
9. InCommon Identity Federation[31]
10. Integrating the Healthcare Enterprise[32]
11. Kidney Precision Medicine Project[33]
12. Observational Health Data Sciences and Informatics[34]
13. Open Biological and Biomedical Ontologies Foundry[35]
14. Research Data Alliance[36]
15. Schema.org[37]
16. The Social Data Foundation[38]
17. Vulcan HL7 FHIR Accelerator[39]

Frameworks and Principles

1. Anonymization Decision Making Framework[40]
2. Data Management Body of Knowledge[41]
3. National Institute of Standards and Technology (NIST) Cybersecurity Framework[42]
4. Principles of Least Privilege[43]
5. Trusted Exchange Framework and Common Agreement[44]
6. Zero Trust Architecture[45]

The project team reviewed these additional resources in parallel with the review of standards.

Some resources, such as the Social Data Foundation, yielded no recommended standards whereas others, like the Open Biological and Biomedical Ontologies Foundry, offered many standards for consideration. When reviews surfaced standards that were not already included in the standards inventory, the project team noted those standards in the data collection instrument, examined them for relevance, and, when relevant, added them to the standards inventory. This process resulted in very few additions to the standards inventory. For example, the Clinical Data Interchange Standard Consortium (CDISC) referenced SDTM, AdaM, and Biomedical Research Integrated Domain Group Model (BRIDG) for consideration. SDTM and AdaM were already present on the standards inventory and the project team noted, examined, and discarded BRIDG based on relevance. Some projects like Research Data Alliance highlighted other resources for review, like Schema.org. When relevant, the project team added those other resources and reviewed them.

The review of frameworks and principles offered guidance mostly applicable to potential future data linkage, sharing, and use that could be informed by a governance metadata schema and associated information system. Notably three frameworks and principles the project team reviewed as distinct resources were in fact interrelated—NIST cybersecurity framework, Principle of Least Privilege, and the Zero Trust Architecture—and had similar recommendations for implementers of data governance. For example, the NIST cybersecurity framework and Principle of Least Privilege in general recommend employing role-based access control, need-to-know basis, regular audits, segregation of duties, monitoring and logging, engagement of alerts and notifications, creation of an incident response plan, and implementation of backup and recovery strategies. Multiple principles and frameworks recommended the use of standards to develop a metadata schema and to implement data governance and related security requirements. However, the security standards used within the reviewed frameworks were not applicable to the creation of a governance metadata schema.

A review of the Anonymization Decision Making Framework recommended that implementers of a governance metadata schema apply appropriate privacy-enhancing techniques such as data masking or pseudonymization while balancing privacy protection with data utility. Implementers could apply data masking and pseudonymization if the metadata schema captures specific individuals as points of contact or investigators.

Table 8 includes detailed findings for each resource.

3.3 Utility Assessment

Of the 35 standards evaluated in the utility assessment, the project team retroactively excluded two, did not recommend 20, and recommended 13 for considered use in the data governance metadata schema. Table 3 provides a summary of utility assessment findings for the 13 recommended standards. Section 5.2 includes a profile for each recommended standard.

Table 3. Summary of Utility Assessment Findings for Recommended Standards

	Standard	Governance Information Domain	Summary of Relevance and Recommendations from Utility Assessment
1	Data Catalog Vocabulary	Dataset Information	Data Catalog Vocabulary (DCAT) is a Resource Description Framework vocabulary designed to facilitate interoperability between data catalogs published on the Web. DCAT demonstrates sufficient completeness, logical consistency, coherence, accessibility, active use, community adoption, and maturity. DCAT is highly relevant for organizations that publish or consume datasets, as it provides a standardized way to describe and discover data catalogs. DCAT is recommended for organizations looking to improve the interoperability and discoverability of their datasets. However, DCAT may have limited application to governance metadata as it does not have a healthcare focus and lacks concepts such as consent or more fine-grained access rights.
2	Data Documentation Initiative	Dataset Information	Data Documentation Initiative (DDI) is a free international standard for describing the data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences. DDI demonstrates sufficient completeness, logical consistency, coherence, accessibility, active use, community adoption, and maturity. DDI may be useful in representing dataset information. DDI has a variety of tools and resources, yet the utility and currency are unclear. DDI is a relatively mature standard with extensive adoption. The latest version of the standard (DDI Lifecycle 3.3) was published in 2020.

	Standard	Governance Information Domain	Summary of Relevance and Recommendations from Utility Assessment
3	Data Tags Suite	Dataset Information	Data Tags Suite (DATS) is the core metadata specification of the Biomedical Research Computing System, which is used in a number of NIH data repositories. DATS is primarily focused on metadata and data discovery. DATS demonstrates sufficient completeness, logical consistency, coherence, accessibility, and active use, and has limited community adoption. DATS covers various aspects of data governance, such as licensing, storage location, access, and adherence to data standards. However, it does not address policy or more detailed rules like those found in Open Digital Rights Language (ODRL), such as prohibitions or duties/obligations. Organizations may need to consider additional standards or custom solutions to address those aspects of data governance. DATS also has limitations in the consent domain, as it only models the participant and consent dates, lacking other common consent elements such as status and scope of the consent.
4	Data Use Ontology	Authorizations	Data Use Ontology (DUO) is a comprehensive and community-driven effort to standardize data use conditions, specifically for research data in the biomedical domain. DUO was developed to address the challenges associated with unique language used in informed consent forms and the lack of a standard universal system for categorizing data use conditions. DUO demonstrates completeness, logical consistency, coherence, accessibility, active use, and community adoption but exhibits moderate maturity. While DUO was initially designed for managing consent-based restrictions, it has the potential to be adopted in other types of agreements, such as authorizations, rules, and controls. However, it will be essential to continually monitor its development and community adoption to ensure it remains suitable and beneficial for broader applications.

	Standard	Governance Information Domain	Summary of Relevance and Recommendations from Utility Assessment
5	Dublin Core	Dataset Information	Dublin Core™ Metadata Element Set (also known as “the Dublin Core” or DCMI) includes 15 core metadata terms plus several dozen properties, classes, datatypes, and vocabulary encoding schemes. DCMI demonstrates robust completeness, logical consistency, coherence, accessibility, active use, and community adoption and is a mature and stable standard. DCMI is a useful standard set of data elements for consideration and support, industry standard for dataset information organization, and reference. DCMI’s ability to represent individuals and organizations could potentially be applicable to governance information domains beyond dataset information.

	Standard	Governance Information Domain	Summary of Relevance and Recommendations from Utility Assessment
6	Fast Health Information Resources Consent Resource	Consent	FHIR Consent standard is defined as a record of a choice by a healthcare consumer (grantor) or their personal representative, which permits or denies an authorized entity (grantee) to perform one or more actions within a given policy context, for specific purposes and periods of time. Anticipated uses for FHIR Consent are written or verbal agreements by a healthcare consumer (grantor) or a personal representative, made to an authorized entity (grantee). FHIR Consent is highly relevant for data governance in healthcare and its application to research is limited but growing. The FHIR Consent resource does not inherently support the more complex needs of research and has not been fully modeled for all potential use cases, but documentation does notate a Research Consent Directive: Consent to participate in research protocol and information sharing required.[46] The Office of the National Coordinator (ONC) Leading Edge Acceleration Projects (LEAP) Consent Management Services grant and ONC Patient Choice projects have supported efforts to begin applying the FHIR Consent resource in research; however, the project team could not identify any strong examples of using FHIR Consent resource in research. The FHIR Consent resource originates from earlier formats in the Clinical Document Architecture (CDA) by both Integrating the Healthcare Enterprise and HL7 (Health Level Seven International), which were primarily developed to address the use case of sharing electronic health records via Health Information Exchanges; thus, the standard is geared toward healthcare treatment. FHIR Consent demonstrates sufficient completeness, logical consistency, coherence, accessibility, active use, and community adoption but it is a newer standard with limited maturity. It provides a comprehensive framework for managing patient consent, ensuring that sensitive healthcare data is shared and used according to the patient's preferences and in compliance with regulatory requirements and organizational policies. However, it is important to note that FHIR Consent has not been fully evaluated in a clinical research setting.

	Standard	Governance Information Domain	Summary of Relevance and Recommendations from Utility Assessment
7	Fast Health Information Resources Data Segmentation for Privacy and Security Labeling	Controls	Fast Health Information Resources Data Segmentation for Privacy and Security Labeling (FHIR DS4P) is a standard for applying security labels with coded tags for use in access control systems governing the collection, access, use, and disclosure of the target FHIR Resource(s) as required by applicable organizational, jurisdictional, or personal “sharing with protection” policies and is highly relevant for data governance in healthcare information systems. FHIR DS4P demonstrates sufficient completeness, logical consistency, coherence, accessibility, active use, and community adoption but it is a newer standard with limited maturity. FHIR DS4P is a comprehensive framework for implementing data privacy and security policies, ensuring that sensitive healthcare data is protected and managed according to regulatory requirements and organizational policies.
8	Informed Consent Ontology	Consent	Informed Consent Ontology (ICO) is an ontology that represents processes and information pertaining to obtaining informed consent in medical investigations that could be applied to governance schema to represent various aspects of consent. ICO demonstrates sufficient completeness, logical consistency, coherence, and accessibility, but active use, community adoption, and maturity are limited. ICO directly reuses more than 40 DUO terms, including *data use modifier* and *consent code* classes, along with their descendants. However, without employing an explicit update mechanism, such as the *owl:imports* annotation, the DUO terms within the ICO ontology might become outdated and inconsistent with their original definitions in the DUO ontology. The lack of recent updates and limited community engagement raise concerns about ICO ongoing maintenance and maturity.

	Standard	Governance Information Domain	Summary of Relevance and Recommendations from Utility Assessment
9	National Cancer Institute Thesaurus	Dataset Information	National Cancer Institute Thesaurus (NCIt) is a reference terminology and biomedical ontology that covers vocabulary for cancer-related clinical care, translational and basic research, and public information and administrative activities, and is a relevant and recommended standard. NCIt demonstrates robust completeness, logical consistency, coherence, accessibility, active use, and community adoption and is a mature and stable standard. NCIt's extensive structure and coverage make it suitable for a wide range of applications in oncology and biomedical research. NCIt has undergone multiple revisions and updates since its initial release and has been tested and proven in various real-world scenarios and applications in cancer research and clinical care.
10	OASIS LegalRuleML	Agreement, Policy, Law	Oasis LegalRuleML offers rule interchange language for the legal domain to enable modeling and reasoning that allows implementers to structure, evaluate, and compare legal arguments constructed using the rule representation tools provided. LegalRuleML is relevant to the governance domains of laws and agreements and demonstrates completeness, logical consistency, coherence, accessibility, active use, and community adoption. LegalRuleML is a mature and stable standard. It enables markup language and tagging of specific legal concepts listed in node elements and vocabulary, features that will benefit the classification and interpretation of legal agreements pertinent to pediatric data and supported use cases.
11	Observational Medical Outcomes Partnership Common Data Model	Dataset Information	The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) is an open community data standard, designed to standardize the structure and content of observational data and to enable efficient analyses that can produce reliable evidence. OMOP CDM contains two metadata tables that can capture metadata concepts and dataset information. OMOP is a mature and robust standard that demonstrates completeness, logical consistency, coherence, accessibility, active use, and community adoption but may offer limited utility across the breadth of governance metadata domains. Considering the engagement of the user community, OMOP may be an opportunity to make recommendations about additions to their existing metadata capture.

	Standard	Governance Information Domain	Summary of Relevance and Recommendations from Utility Assessment
12	Open Digital Rights Language	Rules	ODRL is a language for the Digital Rights Management community for the standardization of expressing rights information over content. ODRL is intended to provide flexible and interoperable mechanisms to support transparent and innovative use of digital resources in publishing, distributing, and consuming of electronic publications, digital images, audio and movies, learning objects, computer software, and other creations in digital form. ODRL demonstrates completeness, logical consistency, coherence, accessibility, and widespread active use. ODRL is fully mature and usable. The ODRL Information Model provides a standard description model and format to express permission, prohibition, and obligation statements that are directly applicable to governance metadata.
13	Operational Data Model	Dataset Information	The Operational Data Model (ODM), created by the Clinical Data Interchange Standards Consortium (CDISC), facilitates the archive and interchange of the metadata and data for clinical research that is vendor neutral and platform independent. ODM is a mature and relevant standard to encode dataset information that demonstrates completeness, logical consistency, coherence, accessibility, active use, and community adoption. However, license information from CDISC may have further restrictions for use similar to other CDISC maintained standards.

The 20 standards the project team did not recommend based on the utility assessment include:

1. Clinical Data Acquisition Standards Harmonization
2. Control Objectives for Information and Related Technologies
3. Datacite 4.3
4. Datasheets for Datasets
5. eXtensible Access Control Markup Language
6. FHIR Provenance Resource
7. FHIR US Core
8. Information Artifact Ontology
9. International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) 38500:2015— Governance of IT for organization

10. Ontology for Biomedical Investigations
11. Ontology of Information Security
12. OpenAIRE
13. Provenance, Authoring and Versioning
14. Provenance Ontology
15. Science On Schema.Org
16. Study Data Tabulation Model
17. terms4FAIRskills
18. Unified Medical Language System
19. US Core Data for Interoperability
20. Web Access Control

The primary reasons the project team did not recommend candidate standards from the utility assessment were applicability and restrictive licenses. Section 5.3 summarizes the rationale for 20 standards not recommended for use in the metadata schema based on utility criteria.

Despite high-level alignment with one or multiple data use, data linkage, or data governance concepts, a detailed assessment of relevance revealed that many standards do not cover the desired governance information domains within their scope. For example, though PROV-O and PAV standards represent important aspects of provenance, these standards are not relevant because they do not effectively represent the origin of authorizations, rules, or controls. WAC and XACML are highly relevant as technical standards for administering access controls, but not for defining the governance around data access. US Core and USCDI lack standardization on common research governance metadata. COBIT and ISO Governance of IT for the organization are frameworks rather than standards for encoding governance metadata, and though highly relevant, both require purchase for use and a detailed review.

Multiple standards that could be relevant to this work have restrictive licenses, representing a significant functional limitation. Two examples are SDTM and CDASH for which the CDISC license prohibits derivative work. UMLS is relevant and recommended for those seeking a comprehensive and standardized system for integrating and harmonizing various biomedical and health-related terminologies; however, all UMLS users and clients of users need a UMLS license agreement.[47] UMLS license has many restrictions that can be challenging to navigate, (e.g., some materials in the UMLS Metathesaurus are from copyrighted sources, and the licensee is responsible for complying with copyright, patent, and trademark restrictions).[48]

3.4 Gap Analysis

The project team evaluated 13 standards to identify which governance information domain(s) (Table 1) the standards could be applied to. After domain-specific review, some standards entirely addressed the attribute requirements of a given domain, whereas other standards incompletely addressed attributes and requirements in their respective domain. If incompletely addressed, such gaps were identified and documented. Table 4 provides these results from the gap analysis according to governance information domain.

Table 4. Gaps by Governance Information Domain

	Domain	Alignment of Standards with Examples	Gaps
1	Dataset Information	Many recommended standards offer terms and concepts to represent basic dataset information such as a dataset name. DCAT, Dublin Core, NCI Thesaurus, ODM, ODRL, and OMOP are standards that could encode dataset information. OMOP includes a METADATA table that contains attributes for a dataset name, metadata unique key, and the date and time of the metadata entry.[49] A CDM_SOURCE table can store detail about the source database such as the name of the common data model used, and holder of the common data model instance and version.	Though the project team identified several standards to encode dataset information, they did not identify any standard to capture a dataset's inclusion of special populations, data type, or dataset granularity (e.g., individual level or aggregate). Different standards can encode the same property differently, e.g., *dc:title* (Dublin Core[50]) vs. *foaf:name* (FOAF – Friend of a Friend[51]).
2	Linkage	The linkage domain holds a wide variety of information from the presence of personally identifiable information (PII) in a dataset to the entity resolver in past linkages. NCIt includes a limited number of concepts applicable to data linkage, such as PII elements, the concepts of *Honest Broker* and *Record Linkage Study*. However, it does not offer a comprehensive method for modeling privacy-preserving record linkage.	There are no existing standards that comprehensively address all attributes in the linkage domain.

3	Consent	FHIR Consent resource, ICO, ODRL, and DATS are standards that could encode consent information. The FHIR Consent resource offers data elements such as field for *consent.provision* that allows a user to represent the information regarding the constraints of the consent policy. The resource also includes a data element for consent *decision* that includes the values deny or permit. ICO supports conformed consent data integration and reasoning in the clinical research space targeting researchers. The ICO top level entities define and standardize informed consent forms, policies governing informed consent, actors involved in the consent process, and the process of consent itself. ODRL contains the concepts describing consenting and consented party that could be used to encode the properties related to the subject giving the consent. DATS contains a *ConsentInformation* schema that captures any conditions limiting the use of a dataset. The schema includes a property for modification of consents over time. It also contains fields such as an identifier, description, and additional properties.	The FHIR Consent resource and ICO adequately cover the consent domain through a series of data elements. Though the project team identified several standards to encode attributes of the consent process, few standards adequately captured assent.
4	IRB	DUO and ICO are standards that could annotate IRB information. The ICO ontology contains two terms related to IRB: *institutional review board* (a subclass of the *organization* class) and *institutional review board approval* (a subclass of the *document* class), which is defined as a document detailing the IRB's approval for a clinical study involving human subjects at a specific site.[52] Incidentally, the latter term is a direct import from NCI Thesaurus. This contrasts with the DUO class *ethics approval required*, which is a data use modifier indicating that the requestor must provide evidence of IRB approval.[53]	Both ICO and DUO only encompass a limited number of terms associated with the IRB process and may need to be expanded depending on the specific use cases.

	Domain	Alignment of Standards with Examples	Gaps
5	Governing Body	Dublin Core and ODRL may be applicable to encoding governance body information. ODRL is a policy expression language that provides a flexible and interoperable information model, enabling the support of many types of rights and obligations, particularly in the realm of digital content, Web content, data, and services.[54] Dublin Core offers a property for a *rightsHolder* that represents a relationship between the resource and a person or an organization owning or managing rights over the resource. In this case, Dublin Core could potentially be used to represent decision-making rights over a dataset resource. It is possible that this property could be applied to represent the role of a governing body, but this application requires more investigation.[55] As for ODRL, an example of its application could be its *Party* class, which includes entities or collections of entities that undertake roles in a rule. This could potentially be used to represent a governance body or its members. However, the applicability of this class in this context also requires further exploration.	Dublin Core and ODRL offer adequate coverage of this domain. The project team notes no gaps.

	Domain	Alignment of Standards with Examples	Gaps
6	Law	LegalRuleML and ODRL adequately cover the Law domain through a series of concepts or "node elements" within its standardized vocabulary. LegalRuleML addresses the following attributes of interest for the Law domain.[56] Law Type is accounted for with LegalRuleML's concepts of Authority, recognizing a person or organization to create, endorse, and enforce Legal Norms, and Jurisdiction as the designated geographic area or subject matter for which the given Authority applies legal power—therefore accounting for local, state, federal, international, and any other given level or legal restriction. Law identification is accounted for with LegalRuleML's concepts of *Reference*, which provides a standardized internal identifier to reference a given statute. The concepts *LegalSource* and *Source* captured in the LegalRuleML standard also permit easy reference of source information and formulated legal norms. Law Content is organized according to Deontic Specification, which is a logical model assigning obligations, permissions, and related concepts according to attributes organized within the LegalRuleML standard, including concepts for Obligation, Agents, Prohibition, Permission, Right, Bearer, Auxiliary Party, SuborderList, and Time. ODRL, while not ideal in addressing law, includes concepts of Policy (parent class to the Set, Offer, and Agreement subclasses) as well as Rule (an abstract concept that represents the common characteristics of Permissions, Prohibitions, and Duties), which can interpret constraints for a given law or legal decision.[57]	The LegalRuleML standard adequately addresses the Law domain and desired attributes of Law Type, Law ID, and Law Content, and so the project team notes no gaps.

	Domain	Alignment of Standards with Examples	Gaps
7	Agreement	LegalRuleML adequately covers the Agreement domain through a series of concepts or “node elements” within its standards vocabulary, in particular.[58] Agreement Type is accounted for by the ability to Reference through internal identifiers the type of agreement found and through the FactualStatement concept where the vocabulary can express the type of agreement referenced. Agreement Name is accounted for through the expression of LegalSource identifying the legal norm format and Authority asserting an agreement. Agreement Content is accounted for according to Deontic Specification, or node elements including Obligation, Agents, Prohibition, Permission, Right, Bearer, Auxiliary Party, SuborderList, and Time that categorize what is permitted or excluded from an agreement, and then context in which it is presented and applicable.	The project team notes no gaps. The LegalRuleML vocabulary or “node elements” address all necessary attributes of the Agreement domain.

	Domain	Alignment of Standards with Examples	Gaps
8	Policy	LegalRuleML, ODRL, and FHIR DS4P can be used to annotate policy. LegalRuleML can formally capture policy constraints and concepts around permissions/prohibitions; however, it may be challenging with less structured or informally documented policies referenced and so presents a gap.[59] ODRL addresses Policy as part of core vocabulary and may better interpret less structured/formal policies referenced where LegalRuleML falls short. ODRL allows the grouping of one or more Rules into the concept of Policy, which can also be validated by the ODRL Validator checking the conformance of the ODRL Policy expressions including cardinality of values and proper expression.[60] FHIR DS4P provides a series of concepts that can be applied as a security label to resource bundles, providing specific security metadata about the information it is fixed to.[61] The access control decision engine uses security label combined with provenance resources to approve, read, change, and determine what resources may be returned and how to handle caveats. Policy is a DS4P tag under security Category, allowing security metadata that segments an IT resource to convey a mandate, obligation, requirement, rule, or expectation relating to its privacy.[62] As such, DS4P covers policy in a security and privacy context, meaning policy for data access and use with consideration to security and privacy policies established, with interpretation of associated constraints needed. This may more so pertain to enacting IRB or written policy determinations where privacy labels on data for restricted access are necessary.	Gaps exist. With a formally structured policy reference where legal language can be clearly interpreted, LegalRuleML through permissions and prohibitions is able to connote structured policy in a given context. This includes interpreting policy documents, levels, and possibly content. Where policy requires multiple rules to be constructed and considered, and written legal language is not available, ODRL is able to associate and compound given Rules into a Policy interpretation and application. Therefore, whether LegalRuleML or ODRL are useful depends on application. DS4P is more applicable to privacy controls as a result of policy or IRB, rather than interpreting legislative policy.

	Domain	Alignment of Standards with Examples	Gaps
9	Rules	The Rules domain aims to define what must occur or not occur, including limitations or constraints on how data are handled. LegalRuleML and ODRL can encode rules information. LegalRuleML contains several "node elements" as part of its vocabulary including Prohibition and Permission to standardize associated constraints and legal concepts.[63] This standard may be too focused on legal agreements to be of benefit to this domain, and so ODRL is recommended instead as a more general and universally applicable standard for Rule representation. ODRL can define and reference permissions appropriately through the designated Rule component of its core vocabulary. ODRL defines Rule as an abstract concept that represents the common characteristics of Permissions, Prohibitions, and Duties. Rule represents a class, with Duty, Permission, Prohibition as sub-classes to Rule. Further enhancing classes are available in ODRL associated with Rule, including Relation, Function, and Failure to further define and constrain criteria for rulemaking.[64]	Gaps exist depending on intended use and context; for example, LegalRuleML may benefit rules established from formally documented agreements, where ODRL may benefit representing a series of Rules that combine into an intended Policy. Depending on use, rules logic may require testing to meet desired functionality, and so the standard adopted may vary.
10	Authorizations	DUO, ODRL, and LegalRuleML are applicable to this domain. DUO can annotate authorized use including Type, Determination, and Spec in precoordinated terms; however, it's unclear if ODRL can provide Auth Source for use.[65] With regard to authorizations for data use and linkage that originate from consent forms, ODRL may be appropriate with its Obtain Consent Action, which may be used as a Duty to ensure that the Assigner or a Party is authorized to approve such actions on a case-by-case basis. May link to a Party with the role "consentingParty" function, relating to the Party domain.[66] If authorization comes from a data use agreement or policy document, other standards such as LegalRuleML in related domains may be applicable.	DUO may be an appropriate standard for all attributes except Auth Source. ODRL may be an appropriate standard for consenting parties in particular, or as relates to previously established Policy and Rules using ODRL. LegalRuleML may be applicable if authorization source and constraints are according to a formal legal agreement.

	Domain	Alignment of Standards with Examples	Gaps
11	Controls	FHIR DS4P provides a series of concepts that can be applied as a security label to resource bundles, providing specific security metadata about the information it is fixed to.[67] The access control decision engine uses security label combined with provenance resources to approve, read, change, and determine what resources may be returned and how to handle caveats. DS4P offers a range of tags that align with the desired security category and manage data access and usage. For instance, it includes *Purpose of Use* tag that encompasses various research or public health activities.[68] As such, a series of technical controls can be annotated, along with other desired information bundles in FHIR.	DS4P is an appropriate standard for representing security controls. However, the landscape of potential data use and linkage controls is extensive—likely extending beyond the coverage of DS4P.
12	Party	Dublin Core, LegalRuleML, and ODRL may be appropriate when describing the entity and entity's role in the data governance process. ODRL defines the concept party as an entity or a collection of entities that undertake roles in a rule. Additionally, the defined concept assignee of identifies an ODRL policy for which the party undertakes. This could represent the role of the entity. Dublin Core defines the terms *agent* and *contributor*, which encode an entity responsible for a resource. These terms can represent party. Dublin Core offers properties for a *contributor*, *creator*, *publisher* that could be used to encode the individual and organization that created and/or submitted metadata as well as other parties that are dataset owners or decision makers.[69] The most relevant LegalRuleML classes include *Agent*, *Authority*, *AuxiliaryParty*, and *Bearer*. *Agent* represents entities that act or have the capability to act, denoting the parties involved in the data governance process. *Authority* signifies a person or organization with the power to create, endorse, or enforce legal norms. This could be the organization or regulatory body that sets the data governance policies. *AuxiliaryParty* and *Bearer* may represent the roles of different entities in the data governance process.	Dublin Core, LegalRuleML, and ODRL can represent organizations and roles. However, as the variety of roles in the governance process further develop, gaps may emerge.

	Domain	Alignment of Standards with Examples	Gaps
13	Data Lifecycle	DDI-Lifecycle 3.3 is designed to document and manage data across the entire lifecycle, from conceptualization to data publication, analysis, and repurposing.[70] Potential functionality of DDI to meet the Data Lifecycle domain requirements includes: • Descriptive documentation of the content, meaning, provenance, and access for a single dataset • Archival preservation of descriptive content • Input basis for more complex descriptions • Input content for discovery and exchange of data at the study, data file, variable, and question levels	DDI addresses majority of requirements for Data Lifecycle domain. The Linking attribute may be a gap; all other attributes appear addressed by DDI including Collection, Sharing, Access, Use.

4 Conclusions

The project team conducted a landscape analysis to identify and evaluate existing standards for use in a data governance metadata schema. A multi-pronged and iterative search yielded 47 standards, of which 33 met inclusion criteria applicable to data linkage or use concepts discussed in the 2022 Report and the 2023 Report. Of those, the project team did not recommend 20 standards on assessment of the utility, which included criteria related to application, completeness and community intent, logical consistency and coherence, accessibility, active use and community adoption, and maturity. The project team recommended the remaining 13 standards for considered use in the data governance metadata schema and included them in the gap analysis: Data Catalog Vocabulary, Data Documentation Initiative, Data Tags Suite, Data Use Ontology, Dublin Core, Fast Health Information Resources Consent Resource, Fast Health Information Resources Data Segmentation for Privacy and Security Labeling, Informed Consent Ontology, National Cancer Institute Thesaurus, Oasis LegalRuleML, Observational Medical Outcomes Partnership Common Data Model, Open Digital Rights Language, and Operational Data Model.

The project team excluded or did not recommend most of the candidate standards (>70%) for use in the data governance metadata schema. The team excluded 14 standards primarily on relevance and recent activity. The team considered some standards to be not relevant because they did not meet the project definition of a standard. They excluded several standards like iDASH (Integrating Data for Analysis, Anonymization and Sharing) and PPO (Privacy Preference Ontology) based on no recent activity or formal deprecation. Furthermore, 20 standards that were relevant to data governance did not demonstrate utility for the governance metadata schema, primarily due to limited application and license limitations. In some cases, standards were overly specialized to a given area. For example, WAC and XACML are highly relevant for implementing access control policies but are not designed to represent governance information about access at the metadata level. Licenses limited three standards, especially those standards maintained by CDISC. Similarly, while UMLS is a comprehensive meta-ontology of medical terminologies, restrictive licensing may limit its applicability and accessibility for certain users and purposes, e.g., outside of the US. As compiling, encoding, and sharing governance metadata is essential to enabling dataset linkage and reuse, the limited success of this effort to identify standards for use in the metadata schema reinforces the need to develop or extend existing standards to achieve greater coverage across governance metadata.

The project team then aligned the 13 standards recommended for use with governance information domains to identify potential gaps, where key governance information should be annotated but no recommended standard could be identified. The project team identified gaps by mapping standards and utility assessment findings to the governance domains. They identified gaps in which the standards are inadequate to address all of the attributes in nine of the 13 governance information domains: Dataset Information, Linkage, Consent, IRB, Policy, Rules, Controls, Party, and Data Lifecycle. There are adequate standards to address the entirety of only four governance information domains: Governing Body, Law (includes Regulations and Statutes), Agreement, and Authorization.

The project team recognized the importance of simplicity and straightforwardness in whichever series of standards it recommends and ultimately adopts in the data governance metadata schema. If too many standards are adopted piecemeal, the final schema runs the real-world risk of being an ineffective and overly tailored compilation that does not perform efficiently and is itself not easily standardizable. The project team optimized to a minimum the number of standards referenced where possible. For example, if ODRL is clearly preferable in a given governance information domain, and then in another domain where ODRL was equivalent to LegalRuleML, ODRL would be prioritized to ensure agreement across domains. Fewer standards and greater coverage of domains with the determined subset of standards allows for better management of variability and various datatypes, and fewer communities, resources, updates, documentation, and source material from each standard that must be monitored and managed over time.

Summary recommendations for the NICHD ODSS data governance metadata schema development focus on the ODRL standard and FHIR Consent information models, with value sets drawn from FHIR terminology and DUO. The findings from this landscape analysis, utility assessment, and gap analysis inform this approach.

- No single standard fully addresses the schema requirements across all domains, necessitating the use of multiple standards and combining elements from various sources, as well as potentially developing new value sets (e.g., to capture linkage metadata).
- The maturity and licensing of existing standards are also significant factors influencing their utility and adoption.
- As the findings from this landscape and gap analysis will inform the development of data governance metadata schema, the schema should balance the need for consistency and interoperability with the need for flexibility and adaptability to accommodate evolving research needs and regulatory requirements.
- The schema development process should be guided by a strong commitment to collaboration and engagement with relevant stakeholders, including researchers, data providers, and policy

makers, to ensure that the resulting schema is both practical and effective in addressing the diverse needs of the research community.

- The schema will subsequently contribute to NIH-wide strategic goals and activities on CDAC.[71]

The development of a robust and extensible governance metadata schema will require careful consideration of the specific requirements based on the 3 pediatric COVID-19 use cases developed by NICHD ODSS and resulting governance information collected by NICHD ODSS. By leveraging the strengths of the ODRL and FHIR consent information models, as well as incorporating relevant value sets and ontology terms from FHIR and DUO, the NICHD ODSS can create a comprehensive and adaptable governance metadata schema that addresses the diverse needs of the community.

It is important to acknowledge the need for continuous evolution of standards to meet community and research needs and to consider the implications of using multiple standards with varying levels of maturity and licensing restrictions. Metadata standards are not a one-size-fits-all solution, and the project team should approach the development of the data governance metadata schema with a clear understanding of the limitations and potential challenges associated with using multiple standards.

By adopting a thoughtful and strategic approach to governance metadata schema development, informed by the findings and recommendations presented in this report, the NICHD ODSS can pave the way for a more standardized, efficient, and transparent system of metadata data governance that supports the advancement of research, data sharing and reuse, and innovation in the field. The project team also hopes this report will be useful to researchers generating datasets, data stewards, stakeholders interested in research using linked datasets across HHS agencies and NIH as well as more broadly, and the patient-centered outcomes research community.

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 common data model (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 data-set linkage

Term	Definition
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 originator/ contributor/submitter	Institutions/organizations/researchers that collect data from patients or study participants or that collect administrative data; they may also be the party to submit the data to a repository for sharing
Data pseudonymization	De-identification technique that replaces an identifier (or identifiers) for a data principal with a pseudonym in order to hide the identity of that data principal (NIST pseudonymization)
Data science	Interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data
Dataset	Collection of related sets of information composed of separate elements that can be manipulated computationally as a unit
Data sharing[e]	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

[e] 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
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)

Term	Definition
Electronic health records (EHRs)	EHRs are electronic versions of the paper charts in your doctor's or other healthcare provider's office. An EHR may include your medical history, notes, and other information about your health including your 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 can 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

Term	Definition
HIPAA Privacy Rule	The Standards for Privacy of Individually Identifiable Health Information are codified in 45 C.F.R. Parts 160 and 164 promulgated by the U.S. Department of Health and Human Services under the Health Insurance Portability and Accountability Act (HIPAA) of 1996. The HIPAA Privacy Rule establishes national standards to protect individuals' medical records and other individually identifiable health information (collectively defined as "protected health information") and applies to health plans, healthcare clearinghouses, and those healthcare providers that conduct certain healthcare transactions electronically. The Rule requires appropriate safeguards to protect the privacy of protected health information and sets limits and conditions on the uses and disclosures that may be made of such information without an individual's authorization. The Rule also gives individuals rights over their protected health information, including rights to examine and obtain a copy of their health records, to direct a covered entity to transmit to a third party an electronic copy of their protected health information in an electronic health record, and to request corrections (HHS Health Information Privacy)
Honest broker	A party that holds 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)

Term	Definition
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 (Taylor, A. G. (2004). Introduction to cataloging and classification (10th ed.))
Ontology	A set of terms or concepts defining the properties or identities of subjects (e.g., genes, proteins, conditions) and relationships between them; similar to a standardized vocabulary
Open access	Data within this category presents minimal risk of participant identification. Access to these data does not require user certification, and researchers may explore data content without restriction (NCI open access) No access restrictions or registration required to access (NIH open access) [see also data access model]
Patient identifier	Unique data used to represent a person’s identity and associated attributes (NIST patient identifier)
Personally identifiable information (PII)	Any information that can be used to distinguish or trace an individual's identity, either alone or when combined with other information that is linked or linkable to a specific individual (NIST PII) and (CODI PII)

Term	Definition
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)
Software	Programs and other operating information used by a computer
Subject ID	A de-identified subject/participant identifier that can be generated by hashing or non-hashing methods. If hashing is used, it is different from a hash code/token (hashed ID) generated using a PPRL tool

Abbreviations and Acronyms

Acronym	Definition
ADaM	Analysis Data Model
AdaM	Automatable Discovery and Access Matrix
ADF	Anonymization Decision Making Framework
AHRQ	Agency for Healthcare Research and Quality
ALFA	Abbreviated Language for Authorization
API	Application Programing Interface
BFO	Basic Formal Ontology
BRIDG	Biomedical Research Integrated Domain Group Model
CC	Creative Commons
CDAC	Controlled Data Access Coordination
CDASH	Clinical Data Acquisition Standards Harmonization

Acronym	Definition
CDC	Centers for Disease Control and Prevention
CDISC	Clinical Data Interchange Standards Consortium
CDM	Common Data Model
CMM	Capability Maturity Model
CMS	Center for Medicare and Medicaid Services
COBIT	Control Objectives for Information and Related Technologies
CODI	Childhood Obesity Data Initiative
COVID	Coronavirus Disease
DATS	Data Tags Suite
dbGaP	Database of Genotypes and Phenotypes
DCAT	Data Catalog Vocabulary
DCMI	Dublin Core
DDI	Data Documentation Initiative
DMBOK	Data Management Body of Knowledge
DUO	Data Use Ontology
DUOS	Data Use Oversight System
EGA	European Genome-phenome Archive
EHR	Electronic Health Record
EOSC-EDMI	European Open Science Cloud Datasets Minimum Information
ESIP	Earth Science Information Partners
FAIR	Findable, Accessible, Interoperable, Reusable
FDA	Food and Drug Administration
FFRDC	Federally Funded Research and Development Corporation
FHIR	Fast Health Information Resource
FOAF	Friend of a Friend
GA4GH	Global Alliance for Genomics and Health
GDPR	General Data Protection Regulation

Acronym	Definition
HHS	Department of Health and Human Services
HIPAA	Health Insurance Portability and Accountability Act
HIT	Health Information Technology
HTML	HyperText Markup Language
IAO	Information Artifact Ontology
ICO	Informed Consent Ontology
iDASH	Integrating Data for Analysis, Anonymization and Sharing
IG	Implementation Guide
IHE	Integrating the Healthcare Enterprise
IPR	Intellectual Property Rights
IRB	Institutional Review Board
ISACA	Information Systems Audit and Control Association
ISO	International Organization for Standardization
ISO/IEC	International Organization for Standardization/International Electrotechnical Commission
JSON	JavaScript Object Notation
KPMP	Kidney Precision Medicine Project
NCI	National Cancer Institute
NCIt	National Cancer Institute Thesaurus
NCVHS	National Committee on Vital and Health Statistics
NICHD	National Institute for Child Health Development
NIH	National Institutes of Health
NIST	National Institute of Standards and Technology
NLM	National Library of Medicine
OBI	Ontology for Biomedical Investigations
OBO	Open Biological and Biomedical Ontologies
ODM	Operational Data Model

Acronym	Definition
ODRL	Open Digital Rights Language
ODSS	Office of Data Science and Sharing
OHDSI	Observational Health Data Sciences and Informatics
OMOP	Observational Medical Outcomes Partnership Common Data Model
ONC	Office of the National Coordinator
ORP	Other Research Products
OS-PCORTF	Office of the Secretary Patient Centered Outcomes Research Trust Fund
OWL	Web Ontology Language
PAV	Provenance Authoring and Versioning
PHI	Protected Health Information
PII	Personally Identifiable Information
PMCID	PubMed Central Identifier
POLP	Principle of Least Privilege
PPO	Privacy Preference Ontology
PPRL	Privacy Preserving Record Linkage
PROV-O	Provenance Ontology
RDA	Research Data Alliance
RDF	Resource Description Framework
REGO	Requirements for Establishing Ground Truth in Observational Data
SDDL	Security Descriptor Definition Language
SDF	Social Data Foundation for Health and Social Care
SDM-XML	Study/Trial Design Model in XML
SDTM	Study Data Tabulation Model
TEFCA	Trusted Exchange Framework and Common Agreement
TEP	Technical Experts Panel
UMLS	Unified Medical Language System
URI	Uniform Resource Identifier

Acronym	Definition
US	United States
USCDI	United States Core Data for Interoperability
W3C	World Wide Web Consortium
WAC	Web Access Controls
XACML	Extensible Access Control Markup Language
XML	Extensible Markup Language
ZTA	Zero Trust Architecture

5 Appendices

5.1 Appendix A Technical Expert Panel Membership

Table 5. Technical Expert 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	Nurse Scientist, Division of Nursing Research and Department of Artificial Intelligence & Informatics, Mayo Clinic

5.2 Appendix B Profiles for Recommended Standards

One profile was created for each standard recommended by the utility assessment. Profiles include a basic characterization of the standard and detailed responses related to utility assessment criteria.

Data Catalog Vocabulary (DCAT)

Description

Data Catalog Vocabulary is a Resource Description Framework (RDF) vocabulary designed to facilitate interoperability between data catalogs published on the Web. DCAT is highly relevant for organizations that publish or consume datasets, as it provides a standardized way to describe and discover data catalogs. It is recommended for organizations looking to improve the interoperability and discoverability of their datasets. However, it does not have a healthcare focus and lacks concepts such as consent or more fine-grained access rights.

Date of Last Update

March 7, 2023

Affiliation

World Wide Web Consortium (W3C)

Recency of Support

Support is current.

Intended User and Community

The intended community includes data publishers, data consumers, data catalog developers, and data management professionals.

License

W3C Software and Document license – 2015 version

DCAT Utility Assessment

Utility Criteria	Response
Application	DCAT may be applicable for encoding dataset information as it is a vocabulary for describing datasets and data catalogs.
Completeness and Community Intent	DCAT is a comprehensive vocabulary with a strong focus on interoperability and facilitating data discovery. The community intends to continuously improve and extend the standard based on feedback and new requirements.
Logical Consistency and Coherence	DCAT is logically consistent and coherent, as it is based on the RDF data model and follows the principles of Linked Data.
Accessibility	DCAT is accessible through the W3C website, and its documentation is available in multiple languages. The standard is also machine-readable, which makes it easy to process and integrate with other systems.

Utility Criteria	Response
Active Use and Community Adoption	DCAT is actively used by various organizations, including governments, research institutions, and businesses, to publish and discover datasets. The standard has been widely adopted by the data management community and is considered a best practice for describing data catalogs and notably used/extended by data.gov and data.gov.uk.
Maturity	DCAT is a mature standard, with its first version published in 2014, the second version (DCAT 2) published in 2019, and the latest version (DCAT 3) published in 2023. The standard has evolved based on community feedback and requirements, and it is expected to continue to develop in the future. It could be considered between the Managed (Level 4) and Optimizing (Level 5) stages of the Capability Maturity Model.
Recommendation	DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. DCAT is highly relevant for organizations that publish or consume datasets, as it provides a standardized way to describe and discover data catalogs. It is recommended for organizations looking to improve the interoperability and discoverability of their datasets. However, it does not have a healthcare focus and lacks concepts such as consent or more fine-grained access rights.
Reference Links	https://www.w3.org/TR/vocab-dcat-3/ https://www.w3.org/standards/history/vocab-dcat-3 Albertoni, R., Browning, D., Cox, S., Gonzalez-Beltran, A. N., Perego, A., & Winstanley, P. (2023). The W3C Data Catalog Vocabulary, version 2: Rationale, design principles, and uptake. arXiv preprint arXiv:2303.088 Gißner, A. (2023). Modeling institutional research data repositories using the DCAT3 Data Catalog Vocabulary (Doctoral dissertation, Humboldt Universitaet zu Berlin (Germany)).

Data Documentation Initiative (DDI)

Description

Data Documentation Initiative is an international standard for describing data from the social, behavioral, and economic sciences and may be relevant to encoding dataset information. Two versions of the standard are currently maintained in parallel. DDI has a variety of tools and resources, though many appear dated, with unclear utility. DDI is a relatively mature standard with extensive adoption, though it appears dated.

Date of Last Update

DDI Lifecycle 3.3 – released April 15, 2020

Affiliation

DDI Alliance – Executive and Scientific Board governing

Recency of Support

Support is current.

Intended User and Community

DDI-Lifecycle is designed to document and manage data across the entire lifecycle, from conceptualization to data publication, analysis, and beyond. Based on Extensible Markup Language (XML) Schemas, DDI-Lifecycle is modular and extensible. This version also supports improvements in classification management (based on Generic Statistical Information Model/Neuchatel), non-survey data collection (Measurements), sampling, weighting, questionnaire design, and support for DDI as a Property Graph.

License

DDI-Lifecycle 3.3 XML Schema is free software; DDI may be redistributed or modified under the terms of the Creative Commons Attribution 4.0 International license. Other DDI documents are similarly distributed under the same Creative Commons license. The development of DDI 3.3 draws on earlier DDI versions and work of the committees and individuals that developed them as well as the collective ideas, needs, and work of the Expert Committee of the DDI Alliance. Major contributions to DDI-Lifecycle 3.3 were made by many individuals and organizations.

DDI Utility Assessment

Utility Criteria	Response
Application	DDI may be applied to encoding dataset information as it is designed to describe data produced by surveys and other observational methods in the social, behavioral, economic, and health sciences.

Utility Criteria	Response
Completeness and Community Intent	DDI is an international and free standard that can document and manage different stages in the research data lifecycle, such as conceptualization, collection, processing, distribution, discovery, and archiving. Documenting data with DDI facilitates understanding, interpretation, and use—by people, software systems, and computer networks.
Logical Consistency and Coherence	DDI provides a codebook specification, guidance on lifecycle management, archives, and documented best practices for use on the main DDI specification site. Some resources are dated to older versions but appear applicable.
Accessibility	Resources are accessible namely on the DDI maintained website; however, a GitHub is also available though it appears dated. The DDI GitHub site includes this disclaimer: This directory has been superseded by the Research Data Alliance (RDA) Metadata Standards Catalog and is no longer maintained.
Active Use and Community Adoption	The full list of DDI adopters is extensive, though slightly dated. GitHub does not provide actively managed projects for DDI. The latest version of the standard (DDI Lifecycle 3.3) was published in 2020.
Maturity	DDI is a managed and repeatable standard. It is considered a Level 4 of the Capability Maturity Model. Despite extensive adoption, GitHub does not indicate new versions of the standard have been published recently. The latest version of the standard (DDI Lifecycle 3.3) was published in 2020.
Recommendation	DDI demonstrates sufficient completeness, logical consistency, coherence, accessibility, active use, community adoption, and maturity. DDI may be useful in representing dataset information. DDI has a variety of tools and resources, yet the utility and currency are unclear. DDI is a relatively mature standard with extensive adoption.

Utility Criteria	Response
Reference Links	https://ddialliance.org/ https://ddialliance.org/history.html https://ddialliance.org/Specification/ https://ddialliance.org/Specification/DDI-Lifecycle/3.3/ https://ddialliance.org/ddi-adopters https://rd-alliance.github.io/metadata-directory/standards/ddi-data-documentation-initiative.html

Data Tags Suite (DATS)

Description

Data Tags Suite is the core metadata specification of the Biomedical Research Computing System, which is used in several National Institutes of Health (NIH) data repositories. DATS is primarily focused on metadata and data discovery. DATS demonstrates sufficient completeness, logical consistency, coherence, accessibility, and active use, but has limited community adoption. DATS covers various aspects of data governance, such as licensing, storage location, access, and adherence to data standards. However, it does not address policy or more detailed rules like those found in Open Digital Rights Language (ODRL), such as prohibitions or duties/obligations. Organizations may need to consider additional standards or custom solutions to address those aspects of data governance. DATS also has limitations in the consent domain, as it only models the participant and consent dates, lacking other common consent elements such as status and scope of the consent.

Date of Last Update

September 5, 2022

Affiliation

The bioCADDIE (Biomedical and Healthcare Data Discovery Index Ecosystem) Project, funded by NIH through the Big Data to Knowledge (BD2K) program

Recency of Support

Support is current.

Intended User and Community

The intended users include researchers, data curators, and developers working with biomedical and healthcare datasets, particularly in the context of data discovery and indexing.

License

CC BY-SA 3.0

DATS Utility Assessment

Utility Criteria	Response
Application	Data Tags Suite may be applied to encoding dataset information and consent.
Completeness and Community Intent	DATS provides a metadata model for describing biomedical and healthcare datasets, including information about the data, its provenance, accessibility, and usage.
Logical Consistency and Coherence	DATS uses a relatively well-defined metadata model and controlled vocabularies to ensure consistency across implementations.

Utility Criteria	Response
Accessibility	The original biocaddie.org website is defunct. The development work has moved to GitHub.
Active Use and Community Adoption	Limited adoption and community use.
Maturity	DATS can be roughly placed between the Initial (Level 1) and Repeatable (Level 2) stages of the Capability Maturity Model. The lack of recent updates and limited community engagement raise concerns about its ongoing maintenance and maturity.
Recommendation	This standard is recommended for use in the metadata schema. DATS demonstrates sufficient completeness, logical consistency, coherence, accessibility, and active use, but has limited community adoption. DATS covers various aspects of data governance, such as licensing, storage location, access, and adherence to data standards. However, it does not address policy or more detailed rules like those found in ODRL, such as prohibitions or duties/obligations. Organizations may need to consider additional standards or custom solutions to address those aspects of data governance. DATS also has limitations in the consent domain, as it only models the participant and consent dates, lacking other common consent elements such as status and scope of the consent.
Reference Links	https://pubmed.ncbi.nlm.nih.gov/32031623/ https://github.com/datatagsuite/schema

Data Use Ontology (DUO)

Description

The Data Use Ontology is a machine-readable standard to express data use conditions in the biomedical domain that could be applicable to the governance metadata schema's need to express data use conditions. DUO has been used to match specific datasets against the data access requests. DUO's originating domain is health, clinical, and biomedical research, particularly focusing on human subject's datasets and their data use conditions. DUO allows semantically tagging datasets with restriction about their usage, making them discoverable automatically based on the consent, authorizations, rules, and controls domains. However, it's essential to continually monitor its development and community adoption to ensure it remains suitable and beneficial.

Date of Last Update

February 23, 2021

Affiliation

DUO is affiliated with the Global Alliance for Genomics and Health (GA4GH).

Recency of Support

Last issue was closed September 22, 2022. Two issues were opened since and remain open as of the date of this report.

Intended User and Community

1. Data Access Committees
2. Researchers in the health/clinical/biomedical domain
3. Large genomics and health data repositories
4. Authors of informed consent forms for human subject's datasets
5. Commercial entities involved in studying health-related datasets

License

Creative Commons Attribution 4.0 International License

DUO Utility Assessment

Utility Criteria	Response
Application	DUO may be applied to encode information about IRB, rules, and authorization and could potentially be useful for dataset information as well.

Utility Criteria	Response
Completeness and Community Intent	DUO has a well-defined purpose, which is to describe data use conditions, particularly for research data in the health, clinical, and biomedical domain, and to provide a standard universal system for categorizing these conditions. It includes ontology terms needed to represent queries and the ontology hierarchy necessary for algorithms to determine compatibility between research purposes and dataset restrictions. DUO is based on the Open Biological and Biomedical Ontologies (OBO) Foundry principles and is developed using the W3C Web Ontology Language. It is already used in production by the European Genome-phenome Archive (EGA) and the Broad Institute for the Data Use Oversight System (DUOS). The community intent of DUO can be described as follows: 1) It aims to facilitate data sharing among large genomics and health data repositories by providing a standardized terminology for describing data use conditions; 2) it seeks to encourage the authors of new consent forms to align consent language with the terms used by DUO to speed up the availability of data for secondary use; and 3) the DUO Workstream, other contributors, and funding organizations are working together to develop and improve DUO.
Logical Consistency and Coherence	DUO is developed using the W3C Web Ontology Language, which ensures logical consistency and coherence in its structure. The ontology terms are organized hierarchically, allowing algorithms to determine compatibility between research purposes and dataset restrictions.
Accessibility	DUO can be accessed through the Ontology Lookup Service or Ontobee and is registered with the OBO Foundry. The ontology is distributed under a Creative Commons Attribution 4.0 International License, ensuring that it is accessible and available for use by the broader research community.
Active Use and Community Adoption	DUO has been implemented in several projects' production pipelines, such as the Broad Institute's DUOS, the EGA, and the Data Information System (DAISY). It is unclear how DUO is used by EGA based on the link provided by the project. The GitHub repository has 56 stars and 17 forks as of 10/27/23.

Utility Criteria	Response
Maturity	DUO can be roughly placed between the Managed (Level 2) and Defined (Level 3) stages of the Capability Maturity Model, with aspects of continuous improvement and evolution that may lead it toward higher levels of maturity in the future. No active development or additional community adoption in the past 12 months may further indicate that DUO's maturity level is closer to Managed (Level 2) stage.
Recommendation	This standard is recommended for use in the metadata schema. The DUO standard could be applied to the metadata schema to encode data use conditions, potentially in the authorizations, rules, and controls. Overall, DUO appears to be a comprehensive and community-driven effort to standardize data use conditions in the health/clinical/biomedical domain. However, it's essential to continually monitor its development and community adoption to ensure it remains suitable and beneficial in the long run.
Reference Links https://github.com/EBISPOT/DUO	https://GitHub.com/EBISPOT/DUO http://purl.obolibrary.org/obo/duo.owl https://www.ga4gh.org/news_item/data-use-ontology-approved-as-a-ga4gh-technical-standard/ https://www.semantic-web-journal.net/content/enhancing-data-use-ontology-duo-health-data-sharing-extending-it-odrl-and-dpv-1

Dublin Core (DCMI)

Description

Dublin Core™ Metadata Element Set (also known as "the Dublin Core" or DCMI) includes fifteen (15) core metadata terms plus several dozen properties, classes, datatypes, and vocabulary encoding schemes. DCMI represents the latest set of metadata terms in RDF and XML versions. DCMI is a useful standard set of data elements for consideration and support as it is considered to be the industry standard for dataset information organization and reference. Dublin Core™ metadata, or perhaps more accurately metadata "in the Dublin Core™ style," is metadata designed for interoperability based on Semantic Web or Linked Data Principles. Metadata in this style uses Uniform Resource Identifiers (URIs) as global identifiers both for the things described by the Metadata and for the terms used to describe them (vocabularies). This style is distinguished by the application profile—a specification detailing how well-known generic vocabularies such as the Dublin Core are used, constrained, or combined with more specialized vocabularies to meet the requirements of specific applications.

Date of Last Update

January 20, 2020

Affiliation

Dublin Core

Recency of Support

Support is current.

Intended User and Community

Application profiles have been the focus of the Dublin Core™ community since they first trended in 2000. The Dublin Core, a set of fifteen (15) generic, widely used elements—Creator, Contributor, Publisher, Title, Date, Language, Format, Subject, Description, Identifier, Relation, Source, Type, Coverage, and Rights—was first drafted at a 1995 meeting in Dublin, Ohio, initially to facilitate information discovery on an explosively growing Web by embedding simple, card-catalog-like metadata in its pages. A diverse community of librarians, technologists, and researchers rallied to the idea, pursued, and refined through a series of lively workshops and conferences, to achieve rough interoperability across languages and disciplines through a core of shared semantics.

License

Creative Commons Attribution 4.0 International License unless otherwise stated

DCMI Utility Assessment

Utility Criteria	Response
Application	Dublin Core may be applied to encoding dataset information, governing body, and party.

Utility Criteria	Response
Completeness and Community Intent	The DCMI core 15 data elements are intentionally set to represent the minimum and most efficient set of descriptive data elements to annotate dataset information and support. DCMI is the de facto standard and is widely used across multiple dataset information domains including health data.
Logical Consistency and Coherence	DCMI is logical and coherent, a well-vetted set of core standard data elements—consisting of Creator, Contributor, Publisher, Title, Date, Language, Format, Subject, Description, Identifier, Relation, Source, Type, Coverage, and Rights
Accessibility	DCMI is easily accessible and published on the Dublin Core hosted website and GitHub.
Active Use and Community Adoption	Actively used and supported since the 1990s.
Maturity	DCMI is a mature and stable standard and is at Level 5 of the Capability Maturity Framework. DCMI has a core set of optimized data elements that are actively used and supported.
Recommendation	This standard is recommended for use in the metadata schema. DCMI is a useful de facto standard set of data elements for consideration and support. DCMI is an industry standard for dataset information, organization, and reference. DCMI is a mature and stable standard.
Reference Links	https://www.dublincore.org/about/copyright/#documentnotice https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ https://www.dublincore.org/specifications/dublin-core/dcmi-terms/release_history/ https://www.dublincore.org/about/copyright/#documentnotice https://github.com/dcmi

Fast Health Information Resources (FHIR) Consent Resource

Description

The purpose of the Fast Health Information Resources Consent Resource is to express a consent regarding healthcare. There are four anticipated uses for the FHIR Consent Resource, all of which are written or verbal agreements by a healthcare consumer (grantor) or a personal representative, made to an authorized entity (grantee) concerning authorized or restricted actions with any limitations on purpose of use, and handling instructions to which the authorized entity must comply:

- Privacy Consent Directive: Agreement, Restriction, or Prohibition to collect, access, use, or disclose (share) information
- Medical Treatment Consent Directive: Consent to undergo a specific treatment (or record of refusal to consent)
- Research Consent Directive: Consent to participate in research protocol and information sharing required

This resource is scoped to cover all three uses, but currently, only the privacy use case is fully modeled; others are being used but no formal modeling exists.

FHIR Consent is highly relevant for data governance in healthcare information systems. It provides a comprehensive framework for managing patient consent, ensuring that sensitive healthcare data is shared and used according to the patient's preferences and in compliance with regulatory requirements and organizational policies.

However, it is important to note that FHIR Consent may not cover all use cases of clinical research outside of a healthcare system, such as data collected using Electronic Data Capture systems in a clinical trial setting.

Date of Last Update

FHIR Release 5 (R5) was published in March 2023. Consent is part of the FHIR R5 release.

Affiliation

HL7 (Health Level Seven International)

Recency of Support

Support is current.

Intended User and Community

The intended community includes healthcare organizations, developers, and policy administrators who need to manage patient consent for data sharing and privacy in healthcare information systems.

License

HL7's FHIR license Creative Commons "No Rights Reserved" (CC0)

FHIR Consent Utility Assessment

Utility Criteria	Response
Application	FHIR Consent may be applied to encoding consent information.
Completeness and Community Intent	FHIR Consent provides a comprehensive framework for managing patient consent in healthcare information systems using the FHIR standard. The FHIR Consent resource is labeled as a FHIR maturity level two standard for trial use. The community is continuing to develop the standard led by the HL7 Community Based Collaborative Care Work Group. The anticipated use cases for the consent resource include both written and verbal agreements between a grantor and grantee.
Logical Consistency and Coherence	FHIR Consent is designed to be logically consistent and coherent, providing a clear and unambiguous way to represent and manage patient consent.
Accessibility	FHIR Consent profile, examples, search parameters, operations, and related documents are freely available on the HL7 FHIR website.
Active Use and Community Adoption	At maturity level 2, FHIR Consent is still undergoing wider adoption. The standard is continuing to be tested at events such as HL7 Connectathon to accelerate its maturity and adoption.
Maturity	Maturity level 2: FHIR Consent has been tested and successfully supports interoperability among at least three independently developed systems leveraging most of the scope (e.g., at least 80% of the core data elements) using semi-realistic data and scenarios based on at least one of the declared scopes of the artifact (e.g., at a FHIR Connectathon). These interoperability results must have been reported to and accepted by the FHIR Management Group.
Recommendation	This standard is recommended for use in the metadata schema. FHIR Consent is highly relevant for data governance in healthcare information systems. It provides a comprehensive framework for managing patient consent, ensuring that sensitive healthcare data is shared and used according to the patient’s preferences and in compliance with regulatory requirements and organizational policies. However, it is important to note that FHIR Consent may not cover all use cases of clinical research outside of a healthcare system, such as data collected using Electronic Data Capture systems in a clinical trial setting.

Utility Criteria	Response
Reference Links	https://www.hl7.org/FHIR/consent.html

Fast Health Information Resource (FHIR) Data Segmentation for Privacy (DS4P) and Security Labeling

Description

Fast Health Information Resource Data Segmentation for Privacy and Security Labeling is a standard used in access control systems governing the collection, access, use, and disclosure of the target information to which they are assigned, as required by applicable organizational, jurisdictional, or personal policies related to privacy, security, and trust.

Date of Last Update

FHIR R5 was published in March 2023. DS4P and Security Labeling are part of the FHIR R5 release.

Affiliation

HL7

Recency of Support

Support is current.

Intended User and Community

The intended community includes healthcare organizations, developers, and policy administrators that need to implement data privacy and security policies in healthcare information systems.

License

HL7’s FHIR license, Creative Commons “No Rights Reserved” (CC0)

Fast Health Information Resource Data Segmentation for Privacy and Security Labeling DS4P Utility Assessment

Utility Criteria	Response
Application	FHIR DS4P can be applied to encoding controls and policy.
Completeness and Community Intent	FHIR DS4P and Security Labeling provide a comprehensive framework for implementing data privacy and security policies in healthcare information systems using the FHIR standard. There is active participation from healthcare organizations, developers, and policy administrators to support and improve the standard.
Logical Consistency and Coherence	FHIR DS4P and Security Labeling are designed to be logically consistent and coherent, providing a clear and unambiguous way to express data privacy and security policies. They use FHIR resources, profiles, and extensions to ensure consistency across implementations.

Utility Criteria	Response
Accessibility	FHIR DS4P and Security Labeling specifications and related documents are freely available on the HL7 FHIR website. There are also numerous resources, tutorials, and open-source implementations available to help developers and healthcare professionals understand and implement these standards.
Active Use and Community Adoption	FHIR DS4P and Security Labeling have been adopted by various healthcare organizations and are being actively used in healthcare information systems. The FHIR community is actively working on the development and improvement of these standards.
Maturity	Maturity Level 3: FMM2 + the artifact has been verified by the work group as meeting the Conformance Resource Quality Guidelines icon; has been subject to a round of formal balloting; has at least 10 distinct implementer comments recorded in the tracker drawn from at least three organizations resulting in at least one substantive change.
Recommendation	This standard is recommended for use in the metadata schema. FHIR DS4P and Security Labeling are highly relevant for data governance in healthcare information systems. They provide a comprehensive framework for implementing data privacy and security policies, ensuring that sensitive healthcare data is protected and managed according to regulatory requirements and organizational policies.
Reference Links	https://github.com/HL7/fhir-security-label-ds4p https://build.fhir.org/ig/HL7/fhir-security-label-ds4p/ http://hl7.org/fhir/security-labels.html

Informed Consent Ontology (ICO)

Description

The Informed Consent Ontology is an ontology that represents processes and information pertaining to obtaining informed consent in medical investigations.

Date of Last Update

April 1, 2021

Affiliation

University of Michigan and OBO Foundry

Recency of Support

The ontology contains 63 administrative notes: “This class is under group discussion as of 03/26/2019.”, is an example of a note.

Intended User and Community

ICO aims to support informed consent data integration and reasoning in the clinical research space, targeting researchers, organizations, and projects working with informed consent and human subject’s research.

License

Creative Commons Attribution 4.0 International License

ICO Utility Assessment

Utility Criteria	Response
Application	ICO can be applied to encoding consent and IRB metadata, specifically consent forms, policies governing informed consent, agents working with patients and biospecimens accompanied by consent, and the process of informed consent itself.
Completeness and Community Intent	ICO represents universals and relations in the domain of informed consent, following the OBO Foundry principles and extending from the top-level ontology Basic Formal Ontology (BFO). It aims to support informed consent data integration and reasoning in the clinical research space and has received contributions from multiple researchers and funding resources.
Logical Consistency and Coherence	ICO is based on OBO Foundry principles and the top-level ontology BFO, ensuring logical consistency and coherence in its structure and representation of the informed consent domain.

Utility Criteria	Response
Accessibility	ICO can be accessed through the Ontology Lookup Service or Ontobee and is registered with the OBO Foundry. The ontology is distributed under a Creative Commons Attribution 4.0 License, ensuring that it is accessible and available for use by the broader research community.
Active Use and Community Adoption	The availability of multiple publications that describe real-world applications of ICO demonstrate adoption and active use. Multiple recent publications describe extensions and applications of ICO. The GitHub repository has 6 stars and 7 forks as of 10/27/23.
Maturity	ICO can be roughly placed between the Initial (Level 1) and Managed (Level 2) stages of the Capability Maturity Model. The lack of recent updates and limited community engagement raise concerns about its ongoing maintenance and maturity.
Recommendation	This standard is recommended for use in the metadata schema. The ICO could be applied to governance to represent various aspects of consent. However, the lack of recent updates and limited community engagement raise concerns about ICO ongoing maintenance and maturity.
Reference Links	https://GitHub.com/ICO-ontology Lin Y, Zheng J, He Y. VICO: Ontology-based representation and integrative analysis of Vaccination Informed Consent forms. J Biomed Semantics. 2016 Apr 19;7:20. Doi: 10.1186/s13326-016-0062-4. PMCID: 27099700. Amith M, Harris MR, Stansbury C, Ford K, Manion FJ, Tao C. Expressing and Executing Informed Consent Permissions Using SWRL: The All of Us Use Case. AMIA Annu Symp Proc. 2022 Feb 21;2021:197-206. PMCID: 35309008. Umberfield EE, Stansbury C, Ford K, Jiang Y, Kardia SLR, Thomer AK, Harris MR. Evaluating and Extending the Informed Consent Ontology for Representing Permissions from the Clinical Domain. Appl Ontol. 2022;17(2):321-336. Doi: 10.3233/ao-210260. Epub 2022 May 4. PMCID: 36312514.

National Cancer Institute Thesaurus (NCIt)

Description

The National Cancer Institute Thesaurus provides reference terminology for many NCI and other systems. The NCIt includes vocabulary for clinical care, translational and basic research, and public information and administrative activities. NCIt defines definitions for over 10,000 cancers and diseases.

Date of Last Update

September 25, 2024

Affiliation

National Cancer Institute (NCI)

Recency of Support

Support is current.

Intended User and Community

The intended user includes researchers, clinicians, and professionals in the fields of oncology, biomedical research, and healthcare.

License

Creative Commons Attribution 4.0 International license (CC BY 4.0)

NCIt Utility Assessment

Utility Criteria	Response
Application	The NCIt may be applied to encoding dataset information and linkage metadata.
Completeness and Community Intent	The NCIt is complete and well-documented, with clear guidelines and specifications for implementation. The community's intent is to provide a comprehensive and standardized terminology for cancer research and clinical care. The NCIt continues to evolve and has 700 new entries each month.
Logical Consistency and Coherence	The NCIt is logically consistent and coherent, with a well-defined structure and organization for cancer-related concepts and their properties.
Accessibility	The NCIt is openly accessible and available for free download from the NCI Thesaurus website. It is also well-documented, with a comprehensive user guide and reference materials for developers.

Utility Criteria	Response
Active Use and Community Adoption	The NCIt is openly accessible and available for free download from the NCI Thesaurus website. It is also well-documented, with a comprehensive user guide and reference materials for developers.
Maturity	The NCIt is mature, having undergone multiple revisions and updates since its initial release. It has been tested and proven in various real-world scenarios and applications in cancer research and clinical care. The NCIt can be considered a maturity level between Level 3 (Defined) and Level 4 (Managed) according to the Capability Maturity Model (CMM) framework.
Recommendation	This standard is recommended for use in the metadata schema. The NCIt is a relevant and recommended standard for those seeking a comprehensive and standardized terminology for cancer research and clinical care. Its extensive structure and coverage make it suitable for a wide range of applications in oncology and biomedical research.
Reference Links	https://ncithesaurus.nci.nih.gov/

OASIS LegalRuleML TC

Description

The Organization for the Advancement of Structured Information Standards (OASIS) LegalRuleML TC defines a rule interchange language for the legal domain. The work enables modeling and reasoning that allows implementers to structure, evaluate, and compare legal arguments constructed using the rule representation tools provided. Legal texts (e.g., legislation, regulations, contracts, and case law) are the source of norms, guidelines, and rules. It is difficult to exchange specific information content contained in the texts between parties, to search for and extract structured content from the texts, or to automatically process it further. Legislators, legal practitioners, and business managers are, therefore, impeded from comparing, contrasting, integrating, and reusing the contents of the texts, since any such activities are manual. In the current Web-enabled context, where innovative eGovernment and eCommerce applications are increasingly deployed, it has become essential to provide machine-readable forms (generally in XML) of the contents of the text. The objective of the LegalRuleML Core Specification Version 1.0 is to define a standard (expressed with XML-schema and Relax NG and based on Consumer RuleML 1.02) that is able to represent the particularities of the legal normative rules with a rich, articulated, and meaningful markup language.

Date of Last Update

April 21, 2020

Affiliation

OASIS Open

Recency of Support

Support is current.

Intended User and Community

The intended user community includes regulators, researchers, legislators, legal practitioners, and business managers.

License

Available to use, if the use of LegalRuleML adheres to 1) Intellectual Property Rights (IPR) Policy, 2) Technical Committee (TC) Processes, 3) Bylaws, and 4) IPR declarations/modes that are further covered in the IPR. This process may represent a challenge to use, especially if users wish to contribute or become members, which automatically must be vetted by the TC. Further licenses may be required as a contributor.

OASIS LegalRuleML Utility Assessment

Utility Criteria	Response
Application	OASIS LegalRuleML may be applied to encoding agreement, policy, law, rules, and authorizations.

Utility Criteria	Response
Completeness and Community Intent	The OASIS LegalRuleML defines a rule interchange language for the legal domain. The work enables modeling and reasoning that allows implementers to structure, evaluate, and compare legal arguments constructed using the rule representation tools provided. Comprehensive of legal norms and standards—markup language typically structured as XML.
Logical Consistency and Coherence	Vocabulary including node elements are consistently and coherently listed with descriptors for markup language use. Specification is available in multiple file types and supported versions.
Accessibility	Spec is easily available and published on OASIS Open hosted site—with public announcements including several formatted listings of the full specification.
Active Use and Community Adoption	In the eHealth domain, LegalRuleML can be used to model privacy issues and security policies for managing document access according to the profile and the view publication stats authorizations of the operator. By using LegalRuleML, it is possible to filter sensitive data, according to the law/regulation, and to create views of the same health record or document based on the role of the querying agent.
Maturity	OASIS LegalRuleML is a mature and stable standard considered between levels 4 and 5 of the CMM framework. The OASIS project has developed extensive documentation to support the adoption and implementation of the standard found on the OASIS website. The standard has significantly evolved through technical work group and community feedback.
Relevance and Recommendation	This standard is recommended for use in the metadata schema. The LegalRuleML is relevant to the governance information domain of laws, agreement, policy, rules, and authorizations. It enables markup language and tagging of specific legal concepts listed in node elements and vocabulary—this will benefit the classification and interpretation of legal agreements pertinent to pediatric data and supported use cases. LegalRuleML is a mature and stable standard.

Utility Criteria	Response
Reference Links	https://docs.oasis-open.org/legalruleml/legalruleml-core-spec/v1.0/cs02/legalruleml-core-spec-v1.0-cs02.html https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=legalruleml https://www.oasis-open.org/2021/09/08/legalruleml-core-specification-v1-0-oasis-standard-published/ https://www.oasis-open.org/news/announcements/legalruleml-core-specification-v1-0-from-the-oasis-legalruleml-tc-approved-as-com/ https://wiki.oasis-open.org/legalruleml/FrontPage#preview https://github.com/oasis-tcs/legalruleml https://www.oasis-open.org/policies-guidelines/ipr/#introduction https://www.oasis-open.org/policies-guidelines/tc-process-2017-05-26/tc-process-16-september-2002/ https://www.oasis-open.org/policies-guidelines/bylaws/ https://www.oasis-open.org/licenses/ https://github.com/oasis-tcs/legalruleml/blob/master/CONTRIBUTING.md https://www.researchgate.net/publication/256536695_LegalRuleML_From_Metamodel_to_Use_Cases_-_A_Tutorial https://www.oasis-open.org/news/announcements/legalruleml-core-specification-v1-0-from-the-oasis-legalruleml-tc-approved-as-com/

Observational Medical Outcomes Partnership Common Data Model (OMOP)

Description

The Observational Medical Outcomes Partnership Common Data Model (OMOP) is an open community data standard, designed to standardize the structure and content of observational data and to enable efficient analyses that can produce reliable evidence. A central component of the OMOP CDM is the Observational Health Data Sciences and Informatics (OHDSI) standardized vocabularies. The OHDSI vocabularies allow organization and standardization of medical terms to be used across the various clinical domains of the OMOP Common Data Model and enable standardized analytics that leverage the knowledge base when constructing exposure and outcome phenotypes and other features within characterization, population-level effect estimation, and patient-level prediction studies.

Date of Last Update

V5.4 – September 24, 2021

Affiliation

OHDSI

Recency of Support

Support is current.

Intended User and Community

OMOP was built by and for researchers from industry, government, and academia.

License

All OMOP and OHDSI artifacts are open source and free. The only exceptions are proprietary vocabularies.

OMOP Utility Assessment

Utility Criteria	Response
Application	OMOP may be applied to encoding dataset information.
Completeness and Community Intent	OMOP has been widely adopted and used by researchers with extensive use in health research using observational data. With robust engagement of the user community, OMOP has been updated to meet the emerging needs of researchers.
Logical Consistency and Coherence	The OMOP Common Data Model prioritizes logical consistency between its 37 tables and 395 fields, also committing to backward compatibility so that updates and versioning do not impact a researcher. OMOP uses standard vocabularies further ensuring coherence using standards.

Utility Criteria	Response
Accessibility	The OMOP common data model specification is available publicly with extensive available documentation. Additionally, a free R package is available to support its use.
Active Use and Community Adoption	The OMOP Common Data Model is being actively used. The 5.4 version on GitHub has 428 forks and 788 stars as well as several open and recently closed issues. OHDSI maintains over 20 forums that are actively used to discuss use of the OMOP Common Data Model, some with thousands of topics and engaged users.
Maturity	OMOP is a mature and stable standard and would be considered as Level 4 of the capability model. OMOP has been widely adopted and implemented throughout the research community. The standard is managed and continues to be supported and updated through OHDSI community calls, the OHDSI steering committee, and affected work groups.
Recommendation	This standard is recommended for use in the metadata schema. The OMOP Common Data Model has two metadata tables that can capture metadata concepts and dataset information. OMOP is a mature and robust standard but may offer limited utilities across the breadth of governance metadata domains. Considering the engagement of the user community, OMOP may be a place to make recommendations about additions to their existing metadata capture.
Reference Links	https://ohdsi.github.io/CommonDataModel/

Open Digital Rights Language (ODRL)

Description

Open Digital Rights Language is an ontology for representing rights and conditions, including permissions, prohibitions, and duties. It's used for digital content, but the principles can be adapted for datasets.

Date of Last Update

February 15, 2023

Affiliation

W3C. This standard is supported by W3C's Permissions & Obligations Expression Working Group.

Recency of Support

Support is current, v2.2.

Intended User and Community

The intended community is nonspecific. ODRL applies to all data rights use cases, including photo and digital assets management, database information collection, and health data and research use.

License

W3C Software and Document license – 2015 version

ODRL Utility Assessment

Utility Criteria	Response
Application	ODRL may be applied to encoding metadata for dataset information, rules, consent, governing body, Law, Policy, Party, and Authorizations.
Completeness and Community Intent	ODRL is a policy expression language that provides a flexible and interoperable information model, vocabulary, and encoding mechanisms for representing statements about the usage of content and services. The ODRL Information Model describes the underlying concepts, entities, and relationships that form the foundational basis for the semantics of the ODRL policies. Policies are used to represent permitted and prohibited actions over a certain asset, as well as the obligations required to be met by stakeholders. In addition, policies may be limited by constraints (e.g., temporal, or spatial constraints) and duties (e.g., payments) may be imposed on permissions.
Logical Consistency and Coherence	W3C maintains its Technical Reports Index to publish the latest ODRL versions, the latest being 2018. The ODRL model presents elements that are consistent and coherent.

Utility Criteria	Response
Accessibility	ODRL is easily accessible through the W3C website and includes documentation in multiple languages.
Active Use and Community Adoption	ODRL is a widely used standard, with Google Scholar returning over 2000 references to Open Digital Rights Language (ODRL).
Maturity	ODRL is fully mature and usable and has evolved from a version 1.1 in 2002 to version 2.2 in 2018. It can be considered between the Managed (Level 4) and Optimizing (Level 5) stages of the Capability Maturity Model. This standard is managed and continues to be supported by the W3C Permissions & Obligations Expression Working Group.
Recommendation	This standard is recommended for use in the metadata schema. The ODRL Information Model provides a standard description model and format to express permission, prohibition, and obligation statements that are directly applicable to governance metadata.
Reference Links	https://www.w3.org/TR/odrl/ https://www.w3.org/TR/?filter-tr-name=odrl https://www.w3.org/TR/odrl-model/ https://www.w3.org/TR/odrl-vocab/ https://w3c.github.io/odrl/profile-bp/

Operational Data Model (ODM)

Description

Operational Data Model is a data exchange standard—vendor-neutral, platform-independent, suited for exchanging and archiving clinical and translational research data, along with their associated metadata...

Date of Last Update

ODM v2.0 is last version; update August 23, 2023

Affiliation

CDISC—the Clinical Data Interchange Standards Consortium

Recency of Support

Support is current.

Intended User and Community

This standard is intended for the clinical medicine and the health domain, specifically for dataset management. ODM-XML is a data exchange standard—vendor-neutral, platform-independent, suited for exchanging and archiving clinical and translational research data, along with their associated metadata, administrative data, reference data, and audit information. ODM-XML facilitates the regulatory-compliant acquisition, archival, and exchange of metadata and data.

License

MIT License, Copyright 2022 CDISC – free of charge

ODM Utility Assessment

Utility Criteria	Response
Application	ODM may be applied to encoding dataset information.
Completeness and Community Intent	This standard has become the language of choice for representing case report form content in many electronic data capture tools. The ODM v2.0 vision is to build on ODM's proven strength and improved support for automation. This will include improved alignment with CDISC Foundational Standards as well as healthcare standards such as HL7 FHIR. New ODM v2.0 features include a RESTful (Representational State Transfer) Application Programing Interface (API) specification for exchanging ODM clinical data and metadata, support for multiple media types (XML and JavaScript Object Notation [JSON]), enhanced semantics, the Study Design Model, data queries, more flexible data structure representations, and operational datasets.

Utility Criteria	Response
Logical Consistency and Coherence	ODM v2.0 can be serialized as XML, JSON, or other formats. An ODM XML schema is currently available. ODM provides a common base structure for standard extensions easing the learning curve and implementation complexity. Several CDISC standards have been developed by extending ODM-XML including Define-XML, SDM-XML, Dataset-XML, Dataset-JSON, CTR-XML, and CT-XML.
Accessibility	Easily accessible through the CDISC website and includes all previous versions, related Implementation Guides (IGs), and conformance rules.
Active Use and Community Adoption	No current open projects noted on GitHub. Due to ODM's generalizability and range of study information, it is compatible with most existing clinical data management systems.
Maturity	ODM is a managed and repeatable standard that can be considered as Level 4 maturity of the Capability Maturity Model. ODM has been widely adopted throughout the community and often used in many electronic data capture tools. Additionally, the CDISC website provides extensive documentation on the ODM specification and guidelines for implementation.
Recommendation	This standard is recommended for use in the metadata schema. ODM is a mature and relevant standard to encode dataset information. However, license information from CDISC may have further restrictions for use similar to other CDISC maintained standards.
Reference Links	https://rdamsc.bath.ac.uk/msc/m106 https://rdamsc.bath.ac.uk/msc/m106 https://www.cdisc.org/standards/data-exchange/odm https://www.cdisc.org/standards/data-exchange/odm-xml/odm-v2-0 https://github.com/cdisc-org/DataExchange-ODM https://github.com/cdisc-org/DataExchange-ODM/blob/main/LICENSE https://www.cdisc.org/odm-v2-0

5.3 Appendix C Summary of Other Standards

The rationale for standards exclusion from the landscape analysis and utility assessment non-recommendation are included below. Table 6 describes the 14 standards excluded from the landscape analysis based on exclusion criteria. Table 7 describes rationale for the 20 standards not recommended for use in the data governance metadata schema based on the utility criteria.

Table 6. Rationale for Standard Exclusion from the Landscape Analysis

	Standard Name	Brief Description and Rationale for Exclusion
1	Abbreviated Language for Authorization	Abbreviated Language for Authorization (ALFA) is a domain-specific language for a high-level description of XACML policies. Among its features, it presents domain-specific information such as attribute identifiers in compact form and it can be compiled into XACML 3.0. ALFA simplifies the authoring process for authorization policies, helping developers tackle authorization quicker than ever before. The language uses a syntax that closely resembles common programming languages such as Java and C#, making it much easier to read and work with than the verbose XML of the standard XACML policy model. By integrating the extension for ALFA into the VS Code environment, policy authoring becomes easier and faster as the XML syntax and encoding are abstracted away.[72] ALFA was excluded from the landscape analysis on relevance because it is a domain-specific language for policy authoring. ALFA is not a standard for governance information and not intended for biomedical research.
2	Attribute Based Access Control	Attribute Based Access Control (ABAC) is an authorization model that evaluates attributes (or characteristics), rather than roles, to determine access. The purpose of ABAC is to protect objects such as data, network devices, and IT resources from unauthorized users and actions—those that don't have "approved" characteristics as defined by an organization's security policies.[73] ABAC was excluded from the landscape analysis on relevance because it represents an access control paradigm similar to Security Descriptor Definition Language (SDDL). ABAC is not a standard for governance information and not intended for biomedical research.
3	Automatable Discovery and Access Matrix	Automatable Discovery and Access Matrix (AdaM) provides a standardized way to unambiguously represent the conditions related to data discovery and access.[74] AdaM was excluded from the landscape analysis because it is no longer maintained and supported by GA4GH. AdaM has been incorporated into Data Use Ontology.

	Standard Name	Brief Description and Rationale for Exclusion
4	B2FIND	B2FIND is a metadata indexing service (based on a comprehensive metadata catalog of research data collections stored in data centers and community repositories) that provides a discovery portal used to find data collections across various domains.[75] B2FIND was excluded from the landscape analysis because it is a search tool and not a standard. This resource is not used by the health or biomedical research field; the available domains are limited to ancient cultures, archeology, and the humanities.
5	Crossref	Crossref is a tool that allows users to find, cite, link, assess, and reuse research objects. Crossref is a not-for-profit membership organization whose goal is to improve scholarly communications. Crossref provides a schema library that collects metadata and provides a structure and set of guidelines to ensure all collected data remains consistent and interoperable.[76] Crossref was excluded from the landscape analysis because it is a repository of mappings (cross references) between journal articles, books, standards, and datasets rather than a standard itself.[77]
6	Dataverse	The Microsoft Dataverse Project created a metadata crosswalk that contains mappings for the most recently released version of the Dataverse software.[78] Dataverse metadata crosswalk was excluded from the landscape analysis because it is an open-source Web application, not a standard. This crosswalk tool consolidates other metadata standards and tools identified for consideration in the landscape analysis.
7	European Open Science Cloud Datasets Minimum Information	European Open Science Cloud Datasets Minimum Information (EOSC-EDMI) offers information metadata guidelines to help users and services find and access datasets by reusing existing data models and interfaces.[79] EOSC-EDMI was excluded from the landscape analysis due to lack of recent activity. This resource does reference some standards included in the standards inventory but does not appear to be currently supported. While this resource is posted to GitHub, there is very limited available documentation and no activity since 2019.
8	Europena	Europena is a search tool that provides access to European cultural heritage artifacts including images, texts, sounds, and videos. Europena provides a collection of digital cultural artifacts. Europena was excluded from the landscape analysis because it is not a standard and lacks relevance to biomedical research or governance information.

	Standard Name	Brief Description and Rationale for Exclusion
9	Integrating Data for Analysis, Anonymization, and Sharing	Integrating Data for Analysis, Anonymization, and Sharing (iDASH) was one of the National Centers for Biomedical Computing under the NIH Roadmap for Bioinformatics and Computational Biology. iDASH was excluded from the landscape analysis because the iDASH record was deprecated on January 30, 2018, when funding ended. As of 2017, all data within any of the communities in iDASH was no longer accessible.
10	Privacy Preference Ontology	Privacy Preference Ontology (PPO) is a lightweight vocabulary that enables users to create fine-grained privacy preferences for their data. The vocabulary is designed to restrict any resource to certain attributes that a requester must satisfy. PPO was excluded from the landscape analysis because no resources or documentation were identified beyond the original 2011 publication.[80] Additionally, there is no recent evidence of an active user community.
11	Research Data Alliance Metadata Interest Group	The RDA Metadata Interest Group concerns itself with all aspects of metadata for research data. It attempts to coordinate the efforts of the working groups concerned with metadata to produce a coherent approach to metadata covering metadata modalities of description, restriction, navigation, provenance, preservation, and the use of metadata for the purposes of discovery, contextualization, validation, analytical processing, simulation, visualization, and interoperation. The RDA Metadata Interest Group was excluded as it is not a standard, but rather a working group interested in defining and sharing research metadata. The Interest Group has not produced any formal standards to the public.
12	Requirements for Establishing Ground Truth in Observational Data	Requirements for Establishing Ground Truth in Observational Data (REGO) is a general-purpose policy language by Open Policy Agent. The primary purpose of REGO is to accept JSON or Yet Another Markup Language (YAML) inputs and data that are evaluated to make policy-enabled decisions about infrastructure resources, identities, and operations. REGO enables users to write policy about any layer of a stack or domain without requiring a change or extension of the language.[81] REGO was excluded as it is a policy language, not relevant for data governance metadata.
13	Security Descriptor Definition Language	SDDL is a security descriptor language that defines string format, which allows storing and transporting information.[82] SDDL was excluded from the landscape analysis because it is not a standard with relevance to data governance metadata. SDDL is used to implement and operate access controls and is not designed to annotate governance information about dataset access.

	Standard Name	Brief Description and Rationale for Exclusion
14	Zenodo Crosswalk of Resources	The Zenodo Crosswalk of Resources provides a list of the most used metadata schemas and guidelines to achieve metadata interoperability. This is a tool that describes a crosswalk between vocabularies, classes, groups, and types of 18 standards. Many of these standards were considered or included in the landscape analysis.[83] Zenodo was excluded from the landscape analysis because it is a tool for analyzing standards rather than a standard itself. Additionally, this resource was extended in OpenAIRE.

Table 7. Rationale for Utility Assessment Determination

	Standard Name	Maintaining Organization	Rationale for Utility Assessment Determination
1	Clinical Data Acquisition Standards Harmonization	The Clinical Data Interchange Standards Consortium	Clinical Data Acquisition Standards Harmonization (CDASH) establishes a standardized way to collect data consistently across studies and sponsors so that data collection formats and structures provide clear traceability of submission data into the Study Data Tabulation Model (SDTM), delivering more transparency to regulators and others who conduct data review. CDASH is part of the Clinical Data Interchange Standards Consortium (CDISC). CDASH is intended to provide more transparency to regulators and data reviewers by standardizing data collection formats and structures that make it clearly traceable to data submission into SDTM.[84] This standard is not recommended for use in the data governance metadata schema because the CDISC license prohibits derivative work, and therefore makes utilization infeasible.

	Standard Name	Maintaining Organization	Rationale for Utility Assessment Determination
2	Control Objectives for Information and Related Technologies	Information Systems Audit and Control Association	Control Objectives for Information and Related Technologies (COBIT) is a framework that aims to help organizations that are looking to develop, implement, monitor, and improve IT governance and information management. COBIT is not recommended for use in the data governance metadata schema as it offers overarching governance best practices for management rather than applicable attributes or concepts to encode governance metadata.[85] COBIT is a business philosophy similar to Lean Six Sigma and other practices. COBIT would only be relevant if the project team was developing high-level governance best practices to benefit NICHD ODSS management activities. COBIT provides management processes and practices but does not provide vocabularies or concepts that may be applied to governance metadata.
3	DataCite 4.4	DataCite Community Metadata Working Group	DataCite provides a consistent approach to access, share, identify, and re-use research datasets. Key to the DataCite service is the concept of a long-term or persistent identifier, making scholarly references easily and persistently identifiable according to designated values such as Journal, BookChapter, and Dissertation, with relational values between resources such as relatedItemType, relatedItemIdentifier, publicationYear, and more.[86] The primary DataCite application is research and scholarly references rather than data governance information. While DataCite can focus on scientific research and relevant resources, it is not recommended for use in the data governance metadata schema as its primary application is academic in nature and for literature discovery rather than supporting data governance metadata.

	Standard Name	Maintaining Organization	Rationale for Utility Assessment Determination
4	Datasheets for Datasets	Microsoft Research	Datasheets for Datasets aim to provide additional metadata about a given dataset to benefit both dataset creators and dataset consumers to understand context, limitations, and potential applications. Descriptors such as Motivation, Composition, Collection Process, Uses, and more are tagged in datasets.[87] By creating and sharing Datasheets for Datasets, researchers and organizations can improve the transparency, accountability, and reusability of their data, leading to more reliable and robust data-driven applications.[88] Although there are relevant applications and rationale is sound, Datasheets for Datasets is too immature as a standard for consideration and not recommended for use in the data governance metadata schema. Lack of maturity is evident through a lack of examples of use, and published resources with accompanying documentation (GitHub etc.) limited to graduate level work that is also dated.
5	Extensible Access Control Markup Language	Organization for the Advancement of Structured Information Standards (OASIS)	XACML is an attribute-based access control policy language or XML-based language, designed to express security policies and access requests to information.[89] XACML is a highly relevant standard for implementing access control policies in XML-based applications and systems. The standard is not recommended for use in the data governance metadata schema as XACML has limited applicability for data governance metadata.
6	Fast Healthcare Interoperability Resource Provenance	Health Level Seven International (HL7)	The Provenance FHIR resource is a record that describes entities and processes involved in producing and delivering or otherwise influencing a given resource, documenting history and contributors to a resource's creation. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility.[90] This standard is not recommended for use in the data governance metadata schema as the FHIR Provenance resource is based on the W3C PROV-O specification and is therefore duplicative, and similarly has limited coverage of governance information domains and applicability to dataset-level metadata information.

	Standard Name	Maintaining Organization	Rationale for Utility Assessment Determination
7	Fast Healthcare Interoperability Resource US Core	Health Level Seven (HL7)	The US FHIR Core IG is a labeled subset of all HL7 US Realm produced FHIR profiles and is designed to provide the community with a single point of reference to foundational US FHIR profiles. These profiles can then be used by US stakeholders when implementing FHIR and act as a basis for creating further US Realm profiles.[91] The FHIR US Core IG is a mature and robust set of standards that represent a wide range of healthcare data classes and elements identified within the United States Core Data for Interoperability (USCDI). While mature and industry recognized, US Core is not recommended for use in the data governance metadata schema as it lacks standardization on common research governance metadata. For example, the US Core IG includes an element for Organization.meta, the metadata about a given resource; however, as a standard it does not further define given metadata types and therefore offers limited applicability to encoding data governance metadata.
8	Information Artifact Ontology	OBO Foundry	The Information Artifact Ontology (IAO) is an ontology of information entities, originally facilitated by the Ontology for Biomedical Investigations (OBI) digital entity and realizable information entity branch.[92] IAO is a relevant standard for those seeking an efficient ontology for representing information artifacts and representation. However, its application to data governance information domain is limited, and therefore it is not recommended for use in the data governance metadata schema. For example, although IAO does incorporate certain data process-related terms, such as *database extract, transform, and load process*, these terms are not only defined at a high level, but are also sparse and scattered, thereby limiting their applicability in a comprehensive data governance context.[93]

	Standard Name	Maintaining Organization	Rationale for Utility Assessment Determination
9	ISO/IEC 38500:2015- Governance of IT for Organization	International Organization for Standardization	The Governance of IT for Organization standard provides guiding principles for members of governing bodies of organizations (which can comprise owners, directors, partners, executive managers, or similar) on the effective, efficient, and acceptable use of information technology within their organizations.[94] This standard is not recommended for use in the data governance metadata schema as it is a guiding philosophy for IT Governance similar to Lean Six Sigma or other managerial best practices and trainings. The standard would only be relevant in providing overarching governance best practices to benefit an organization's management activities. This standard has limited applicability as it does not provide vocabularies for metadata, only overarching management processes and standardized practices.
10	Ontology for Biomedical Investigations	Ontology for Biomedical Investigations Consortium	OBI communicates scientific investigation information by defining more than 2,500 terms for assays, devices, and objectives.[95] OBI is a relevant standard for those seeking a comprehensive and standardized ontology for representing biomedical investigations. For example, the Core Classes of OBI include two main classes for Continuant, which focuses on information content including specimen type, organism, data item returns from testing, and similar, while the second class Occurrent focuses on biological processes including assay results and specimen collection.[96] The ideal use case given by OBI is for biobanking, exemplifying that this standard is designed for a subspecialty of biomedical research and has limited applicability to the broader research and data governance metadata domains.[97] OBI is not recommended for use in the data governance metadata schema.

	Standard Name	Maintaining Organization	Rationale for Utility Assessment Determination
11	Ontology of Information Security	Linkoping University	Ontology of Information Security is an OWL (Web Ontology Language)-based ontology of information security that models assets, threats, vulnerabilities, countermeasures, and their relations. The ontology can be used as a general vocabulary, roadmap, and extensible dictionary for the domain of information security.[98] This standard is not recommended for use in the data governance metadata schema because it has limited application to data governance metadata and is overly information security focused. For example, the ontology includes several layers of information security concepts including Countermeasures and Cryptography, Vulnerability, Threat, and Defense Strategy that are critical to keeping data secure but not applicable to describing metadata.[99]
12	Open Access Infrastructure for Research in Europe Guidelines for Other Research Products	Open Access Infrastructure for Research in Europe (OpenAIRE)	The Open Access Infrastructure for Research in Europe (OpenAIRE) Guidelines for Other Research Products (ORP) Repository Managers 1.0 provide orientation for repository managers to define and implement their local management policies according to the requirements of the OpenAIRE standard. The OpenAIRE standard is specifically designed to support and meet the Open Access strategy and requirements of the European Commission. These guidelines are intended to provide instruction on how to cite ORPs; for example, peer-reviewed scientific publications, academic journals, research data, and more with an intended open access.[100] While relevant for research data, this standard is not recommended for use in the data governance metadata schema as its primary intent is to support the European Commission's requirements, and is potentially duplicative as it borrows from Dublin Core, DataCite, and other standards already included in this landscape analysis.

	Standard Name	Maintaining Organization	Rationale for Utility Assessment Determination
13	Provenance, Authoring, and Versioning	Massachusetts General Hospital; Harvard Medical School; Balboa Systems; University of Manchester	Provenance, Authoring, and Versioning (PAV) is an ontology for tracking provenance, authoring, and versioning with a dedicated vocabulary, designed to address more specific needs than Dublin Core or PROV-O, for example. PAV specializes the W3C provenance ontology PROV-O to describe authorship, curation, and digital creation of online resources.[101] This standard is not recommended for use in the data governance metadata schema as PAV has limited application in the data governance domain. While PAV can distinguish between contributors, authors and content creators, curators, and more, it is evident through example projects such as Semantic Web Applications in Neuromedicine that the context in which PAV is employed, namely for research annotation and publication, is not applicable to data governance metadata.
14	Provenance Ontology	World Wide Web Consortium	The Provenance Ontology (PROV-O) expresses the PROV Data Model using the OWL2 Web Ontology Language and provides a set of classes, properties, and restrictions that can be used to represent and interchange provenance information generated in different systems and under different contexts. It can also be specialized to create new classes and properties to model provenance information for different applications and domains. The Provenance Ontology was developed by W3C and can be used to trace data origins and transformations.[102] PROV-O is relevant for organizations that publish or consume provenance information, as it provides a standardized way to describe and discover provenance data. However, this standard is not recommended for use in the data governance metadata schema as it does offer an approach to annotating the origin of authorizations, rules, or controls suggesting limited applicability.

	Standard Name	Maintaining Organization	Rationale for Utility Assessment Determination
15	Science on Schema.Org	Earth Science Information Partners (ESIP) Schema.org Cluster	Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on Web pages, in email messages, and beyond.[103] This effort is founded and well supported by Google, Microsoft, Yahoo, and Yandex, with current use by many of the same companies as well as Pinterest. The primary use of the Schema.org published schemas is to mark up web pages and email messages, designating entities within content and relationships between entities and actions. Schema.org is well supported and extensible with regular updates released. While well supported, this standard is not recommended for use in the data governance metadata schema as its applicability to data governance metadata is low. Schema.org is also potentially duplicative as it implements other standards (e.g., PROV-O).
16	Study Data Tabulation Model	The Clinical Data Interchange Standards Consortium (CDISC)	Study Data Tabulation Model (SDTM) provides a standard for organizing and formatting data to streamline processes in collection, management, analysis, and reporting. Implementing SDTM supports data aggregation and warehousing, fosters mining and reuse, facilitates sharing, helps perform due diligence and other important data review activities, and improves the regulatory review and approval process.[104] Despite relevance, SDTM is not recommended for use in the data governance metadata schema as the CDISC license prohibits derivative works and makes utilization infeasible.

	Standard Name	Maintaining Organization	Rationale for Utility Assessment Determination
17	terms4FAIRskills	Committee on Data International Science Council (CODATA)	The terms4FAIRskills (T4FS) project aims to create a formalized terminology that describes the competencies, skills, and knowledge associated with making and keeping data FAIR (Findable, Accessible, Interoperable, Reusable).[105] T4FS is a relevant standard for those seeking a modern, efficient, and reliable terminology solution for FAIR data skills. However, its application to the data governance domain is limited as application of FAIR terminology falls outside the project team's scope for schema development, and therefore is not recommended for use in the data governance metadata schema. Examples of use cases for the T4FS terminology include creation and assessment of stewardship curricula, trainings, and resources for enabling FAIR practices, and formalization of job descriptions with FAIR competencies.[106]
18	Unified Medical Language System	U.S. National Library of Medicine (NLM)	Unified Medical Language System (UMLS) provides a set of files and software that brings together health and biomedical vocabularies and standards to enable interoperability between computer systems. One of the primary UMLS knowledge sources, the UMLS Metathesaurus, links synonymous terms from over 200 vocabulary sources and identifies useful relationships between terms.[107] UMLS is a relevant standard for those seeking a comprehensive and standardized system for integrating and harmonizing various biomedical and health-related terminologies.[108] However, this standard is not recommended for use in the data governance metadata schema as its license terms may require additional considerations before any Unified Modeling Language implementation. While the UMLS Metathesaurus is a useful composite of 200 source vocabularies, these vocabularies are individually and separately referenceable.

	Standard Name	Maintaining Organization	Rationale for Utility Assessment Determination
19	US Core Data for Interoperability	Office of the National Coordinator	USCDI is a standardized set of health data classes and constituent data elements for nationwide, interoperable health information exchange.[109] This standard is not recommended for use in the data governance metadata schema. USCDI is a mature and robust standard that is actively being updated and evolving to meet the needs of health IT stakeholders. However, USCDI offers limited applicability across the breadth of data governance metadata domains.
20	Web Access Controls	W3C	Web Access Control (WAC) is a decentralized cross-domain access control system providing a way for Linked Data systems to set authorization conditions on HTTP resources using the Access Control List model. This RDF vocabulary can be used to describe access control lists. It's primarily used in the solid framework but can be applied more generally for dataset permissions. The WAC standard is similar to the access control system used within many file systems except that the documents controlled, the users, and the groups are all identified by URIs.[110] This approach covers some aspects of data governance, but not beyond read/write access control making its applicability limited. This standard is not recommended for use in the data governance metadata schema.

5.4 Appendix D Review of Additional Resources

The landscape analysis search efforts and Technical Experts Panel recommendations yielded 23 projects, consortiums, initiatives, frameworks, and principles that were reviewed to identify additional standards for consideration or relevant guidance for the formation of a governance metadata schema. Summary descriptions and relevant findings for the project are provided in Table 8.

Table 8. Findings from Review of Additional Resources

Resource	Description and Relevance to the Project
Anonymization Decision Making Framework	The UK Anonymization Network publishes the Anonymization Decision Making Framework (ADF) to address a need for a practical guide to General Data Protection Regulation (GDPR)-compliant anonymization that gives more operational advice than other publications.[111] ADF is primarily intended for those who have microdata that they need to anonymize with confidence, typically in order to share it for some purpose in some form compliant with GDPR and the UK Data Protection Act. While ADF primarily focuses on the anonymization of personal data, its principles can be applied to a governance metadata schema to ensure that sensitive information is protected while maintaining the utility of the metadata. ADF recommends limiting the collection of sensitive information in the metadata schema to the minimum necessary for the needed purpose and conducting a risk assessment to identify potential privacy risks including unauthorized access, re-identification, or misuse. ADF suggests applying appropriate privacy-enhancing techniques to the governance metadata schema such as data masking, pseudonymization, or generalization while balancing privacy protection with data utility. Employing access controls and defining governance policies for the governance metadata are recommended to protect its integrity and confidentiality as appropriate.
Biomedical and Healthcare Data Discovery Index Ecosystem	Biomedical and Healthcare Data Discovery Index Ecosystem (bioCADDIE), funded by a U24 resources grant, is a consortium led by the University of California San Diego that has created an ecosystem in which all details, from object unique identifiers to metadata specifications, allow for easy sharing and formatting for citing data.[112] The bioCADDIE project created DataMed, a prototype biomedical data search engine that allows searching across data repositories and data aggregators supporting the FAIR principles. bioCADDIE is associated with the Data Access Tag Suites to describe the datasets being ingested into DataMed.

Resource	Description and Relevance to the Project
Bioschemas	Bioschemas aims to improve the Findability on the Web of life sciences resources such as datasets, software, and training materials.[113] It does this by encouraging people in the life sciences to use Schema.org markup in their websites so that they are indexable by search engines and other services. Bioschemas encourages the consistent use of markup to ease the consumption of the contained markup across many sites. This structured information then makes it easier to discover, collate, and analyze distributed resources. Bioschemas is making two main contributions: proposing new types and properties to Schema.org to allow for the description of life science resources and defining usage profiles over the Schema.org types that identify the essential properties to use in describing a resource. To simplify the marking up of web resources, and to provide consistency of markup within the life sciences community, Bioschemas is defining profiles over types that state which properties must be used (minimum), should be used (recommended), and could be used (optional). The profiles also state the cardinality of usage of a property and identify domain ontologies to use for the value of properties. For example, the schema.org/Dataset type has over 100 properties available to use. The Bioschemas profile over Dataset brings this down to a more manageable number, with 5 mandatory properties and 8 recommended properties. Many of the other properties have little relevance for a Dataset. The dataset markup properties that Bioschemas specifies as mandatory will also make them findable by Google's Dataset Search tool. The Bioschemas community is defining profiles over relevant existing Schema.org types (e.g., DataCatalog, Course, and SoftwareApplication) and over the new types being defined for the life sciences (e.g., Gene, Protein, and Taxon).

Resource	Description and Relevance to the Project
Cancer Biomedical Informatics Grid	Cancer Biomedical Informatics Grid (caGrid) began as a US government program to develop an open-source, open access information network for secure data exchange on cancer research.[114] The initiative was developed by the National Cancer Institute (part of the National Institutes of Health) and was maintained by the Center for Biomedical Informatics and Information Technology. The cancer Biomedical Informatics Grid (caBIG) project charged with developing caGrid was officially retired in 2011; however, the resources and guidance it provided for governance metadata schema can still be valuable for researchers and organizations working in cancer research and data sharing. The National Cancer Informatics Program was created as caBIG's successor program. caBIG developed caGrid for secure data exchange on cancer research. caGrid evaluated the state of existing technology frameworks and the availability of tools and middleware systems in each framework. The caBIG compatibility guidelines stated that the caBIG services expose "Gold Level" analytical and data resources to the Grid environment. Gold systems are defined as the information models, terminologies, ontologies, and common data elements that were accepted as standards within the caBIG community. They created object-oriented service interfaces in the form of Grid services and use XML for data exchange. They leveraged the NCI Enterprise Vocabulary Services, cancer Data Standards Registry and Repository (caDSR), and the Mobius Global Model exchange, for ontology and metadata schema management. These standards can be reviewed for the development of a metadata schema.

Resource	Description and Relevance to the Project
Clinical Data Interchange Standard Consortium	The Clinical Data Interchange Standards Consortium (CDISC) is an organization that develops and supports global data standards to streamline clinical research and enable the exchange of high-quality clinical data.[115] CDISC creates clarity in clinical research by bringing together a global community of experts to develop and advance data standards of the highest quality, with a focus on accessibility, interoperability, and reusability of data for more meaningful and efficient research to impact global health. CDISC provides standardized terminologies, such as CDISC Controlled Terminology, which can be used in a data governance metadata schema to promote consistent definitions, classifications, and descriptions of data elements. CDISC offers several data models and standards, including SDTM, Analysis Data Model (AdaM), and Biomedical Research Integrated Domain Group Model. These models can guide the design and structure of a governance metadata schema, ensuring interoperability and consistency across different systems and datasets. The Shared Health and Research Electronic Library is a metadata repository that stores, manages, and shares standardized metadata definitions and could be used to leverage existing metadata definitions and understand best practices for managing governance metadata. CDISC has guidelines and best practices for data governance, including data quality, data stewardship, and data lifecycle management. These principles can be applied to the data governance metadata schema to ensure effective management and control of metadata. CDISC standards promote seamless data integration across different systems, platforms, and organizations. Incorporating these standards into a data governance metadata schema can help facilitate data exchange and collaboration.
Creative Commons Licenses	Creative Commons (CC) is an international nonprofit organization that empowers people to grow and sustain the thriving commons of shared knowledge and culture needed to address the world's most pressing challenges and create a brighter future for all.[116] Creative Commons licenses give everyone from individual creators to large institutions a standardized way to grant the public permission to use their creative work under copyright law. From a reuse perspective, the presence of a Creative Commons license on a copyrighted work answers the question, "What can I do with this work?" There are six types of licenses—CC BY, CC BY-SA, CC-BY-NC, CC BY-NC-SA, CC BY-ND, and CC BY-NC-ND—that address credit to the creator, commercial uses, derivatives, and adaptations of the original work. Though the CC license is not designed for use in licensing reuse of dataset per se, the CC license offers a framework that could be applicable to describing a dataset's authorization(s) for reuse. The issues that the CC license types address are relevant to data use and linkage.

Resource	Description and Relevance to the Project
Data Management Body of Knowledge	The Data Management Body of Knowledge (DMBOK) is a comprehensive framework that provides a holistic approach to data management by defining different knowledge areas and best practices.[117] For governance metadata schema, DMBOK covers various aspects that can guide organizations in designing and implementing an effective schema. DMBOK emphasizes the importance of establishing a data governance framework that defines the roles, responsibilities, policies, and processes for managing metadata. DMBOK identifies metadata management as a critical knowledge area and recommends organizations have a comprehensive metadata management strategy. This strategy should cover the governance metadata schema, including its design, implementation, and maintenance, as well as the tools and technologies used for metadata management. When designing a governance metadata schema, DMBOK suggests considering the organization's data architecture, including the data models, data flows, and data integration patterns. The schema should be designed in a way that supports and aligns with the organization's overall data architecture. The governance metadata schema should facilitate data integration across different systems and platforms within the organization. DMBOK recommends using standard metadata models, vocabularies, and ontologies to promote interoperability and seamless data integration. DMBOK highlights the role of data stewards in managing and maintaining governance metadata. The schema should be designed in a way that enables data stewards to easily understand, update, and manage the metadata, as well as to enforce data governance policies and procedures.
The database of Genotypes and Phenotypes	The database of Genotypes and Phenotypes (dbGaP) is a public repository developed by the National Center for Biotechnology Information that archives and distributes the results of studies investigating the interaction of genotype and phenotype in humans.[118] Such studies include genome-wide association studies, medical sequencing, molecular diagnostic assays, as well as association between genotype and non-clinical traits. The principles, standards, and resources can offer guidance for a governance metadata schema. dbGaP's standardized definitions, classifications, and descriptions of data elements can be a helpful reference when defining elements of a governance metadata schema. dbGaP defines best practices for data submission, organization, and formatting that could be applicable to a governance metadata—ensuring a consistent and structured approach to data management. dbGaP recommends a controlled-access model to protect sensitive data that, if applied to governance metadata, could promote data sharing and collaboration among researchers. dbGaP has strict guidelines for data security and privacy, including de-identification, data encryption, and secure data transfer methods. dbGaP adheres to various standards and regulations, such as the NIH Genomic Data Sharing Policy that has relevant implications for this work. dbGaP emphasizes the importance of comprehensive documentation and adherence to 99 metadata standards.

Resource	Description and Relevance to the Project
Global Alliance for Genomics and Health consortium	Global Alliance for Genomics and Health consortium (GA4GH) is a community that develops open-source products to facilitate the request, access, and storage of study data anywhere and help organizations become effective data stewards.[119] GA4GH has created a Common Framework that is a set of guidelines, policies, and technical standards developed by the alliance to enable responsible, voluntary, and secure sharing of genomic and health-related data. The common framework consists of the following components: Regulatory and Ethics Toolkit, Data Security Toolkit, Data Sharing and Access Policies, Technical Standards, and Work Streams. Relevant work products include the Data Access Committee Review Standards Toolkit, Data Use Ontology, and Machine-Readable Consent Guidance.
Globus Toolkit	Globus Toolkit GridFTP and Grid Security Infrastructure software have been widely used within the scientific community for data transfer and security.[120] Since 2010, developers have leveraged that experience to create the Globus cloud service, which provides enhanced capabilities for data transfer plus new identity and group management, data sharing, data publication, and other functions. Limited information about the Globus Toolkit is available as it was discontinued in 2018. The Globus Toolkit's various components and services could be useful in managing, transferring, and sharing metadata within distributed systems and data grids. Globus Toolkit provides reliable and high-performance data transfer capabilities using the GridFTP protocol. This can be used to transfer governance metadata between different systems and platforms securely and efficiently. The Grid Security Infrastructure and the Globus Online service could implement controlled access to governance metadata across different organizations and systems. The Replica Location Service can catalog and discover distributed data resources to maintain an inventory of governance metadata across multiple systems and facilitate the discovery of relevant metadata. The Globus Task Execution Service and the Globus Resource Allocation and Management service could support metadata management tasks and processes.

Resource	Description and Relevance to the Project
InCommon Identity Federation	The InCommon Federation by Educause is the signer and curator of US research and education trust registry information used in federated transactions globally. Think of the registry as a trust phone book.[121] The InCommon Trust Registry/Metadata Service allows Service Providers and Identity Providers to communicate with each other safely and securely. The InCommon Identity Federation is a trusted framework that enables secure access to online resources and simplifies collaboration across various organizations, such as universities, research institutions, and government agencies. InCommon Federation ensures a secure environment for exchanging identity information by adhering to strict security standards, such as Security Assertion Markup Language and the Identity Assurance Profiles. These security measures can be incorporated into the governance metadata schema to ensure data protection and secure access to resources. InCommon Federation offers guidance on managing digital identities, including user attributes, authentication, and authorization. This information can be used to develop a robust governance metadata schema that supports effective identity management and access control. InCommon Federation provides guidelines on privacy and data protection, ensuring compliance with relevant regulations, such as the Family Educational Rights and Privacy Act and GDPR. These guidelines can be integrated into the governance metadata schema to ensure proper handling of sensitive information and regulatory compliance. InCommon Federation maintains a centralized metadata repository containing information about participating organizations and their services. The InCommon metadata is a schema valid against OASIS Security Assertion Markup Language V20. This metadata can be used as a reference for developing the governance metadata schema, ensuring consistency and interoperability across different organizations. In summary, the InCommon Identity Federation offers valuable resources and guidance for developing a governance metadata schema by providing standardized guidelines, best practices, and security measures for identity and access management. This ensures trust, security, collaboration, privacy, and compliance across organizations participating in the federation.
Integrating the Healthcare Enterprise	Integrating the Healthcare Enterprise (IHE) is an initiative by healthcare professionals and industry to improve the way computer systems in healthcare share information.[122] IHE promotes the coordinated use of established standards such as Digital Imaging and Communications in Medicine and HL7 to address specific clinical needs in support of optimal patient care. Systems developed in accordance with IHE communicate with one another better, are easier to implement, and enable care providers to use information more effectively. IHE addresses the information exchange and electronic health record content standards necessary to share information relevant to quality improvement in patient care, clinical research, and public health monitoring. The Quality, Research and Public Health domain was formed in 2007 to address use cases related to repurposing of clinical data for these critical “secondary” uses.

Resource	Description and Relevance to the Project
Kidney Precision Medicine Project	The Kidney Precision Medicine Project (KPMP) is an ambitious, multi-year project funded by the National Institute of Diabetes and Digestive and Kidney Diseases with the purpose of understanding and finding new ways to treat chronic kidney disease and acute kidney injury.[123] The KPMP Consortium includes patient representatives, researchers, and clinicians to meet the goals of the study and needs of the community. The project involves collecting and analyzing human kidney tissue samples to identify novel therapeutic targets and biomarkers. While the KPMP is focused on kidney research, it does not specifically define a metadata schema or data model. However, it is part of the broader precision medicine ecosystem, which emphasizes the importance of data sharing, interoperability, and standardization. In the context of metadata schema and governance, the KPMP can provide insights and best practices for managing, sharing, and standardizing the data generated during the project. This may include data management and standardization. The KPMP follows standardized data formats and ontologies to ensure interoperability and data consistency. The project developed a metadata repository that generates a standard set of metadata for each dataset. These practices can be extended to create a governance metadata schema that ensures data quality and standardization across different research projects.
National Institute of Standards and Technology Cybersecurity Framework	The National Institute of Standards and Technology (NIST) cybersecurity framework is a set of guidelines developed by the NIST to help organizations manage and reduce their cybersecurity risks.[124] The framework consists of five core functions: Identify, Protect, Detect, Respond, and Recover. The NIST cybersecurity framework can be applied to governance metadata to ensure its confidentiality, integrity, and availability and would recommend the inventory and classification of governance metadata, regular risk assessments, defined roles and responsibilities, implementation of access controls, ensuring secure storage and transmission, establishment of governance policies for the governance metadata, monitoring and logging, engagement of alerts and notifications, creation of an incident response plan, and implementation of backup and recovery strategies.

Resource	Description and Relevance to the Project
Observational Health Data Sciences and Informatics	The Observational Health Data Sciences and Informatics (OHDSI) program is a multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics.[125] All its solutions are open source. OHDSI has established an international network of researchers and observational health databases with a central coordinating center housed at Columbia University. While OHDSI does not specifically offer a governance metadata schema, it provides various resources, tools, and guidance that could help in creating or managing a governance metadata schema. The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), to standardize observational healthcare data, can be used as a foundation for creating a metadata schema that supports data governance. The CDM allows researchers to transform data from various sources into a consistent format, enabling efficient analysis and research. A key component to OMOP is the standardized vocabularies. OHDSI provides a set of standardized vocabularies that help in harmonizing clinical data across various healthcare systems. These vocabularies can be used to define metadata elements for a governance metadata schema. OHDSI also has a Data Quality Dashboard that provides guidance on data quality checks and monitoring. This could help in developing a governance metadata schema that ensures data quality across the organization.
Open Biological and Biomedical Ontologies Foundry	The mission of Open Biological and Biomedical Ontologies (OBO) Foundry is to develop a family of interoperable ontologies that are both logically well-formed and scientifically accurate.[126] To achieve this, OBO Foundry participants follow and contribute to the development of an evolving set of principles including open use, collaborative development, non-overlapping and strictly scoped content, and common syntax and relations, based on ontology models that work well. OBO Foundry provides a library of structured, standardized ontologies, some of which could be used to define and organize governance metadata, promoting consistency and interoperability across different datasets and systems. OBO Foundry ontologies follow a common metadata framework that includes standardized annotation properties, relationships, and terms that could serve as a reference for creating a consistent and interoperable governance metadata schema. Relevant ontologies for consideration include IAO, BFO, OBI, and Semantic Science Integrated Ontology.

Resource	Description and Relevance to the Project
Principle of Least Privilege	The principle of least privilege (POLP) is a computer security concept in which a user is granted the minimum levels of access necessary to complete their job functions.[127] This principle can also be applied to governance metadata, which refers to the information that describes the structure, policies, and processes related to the management and control of data. Applying the principle of least privilege to governance metadata involves ensuring that access to metadata is restricted to only those individuals who require it to perform their duties. This can help prevent unauthorized access, tampering, or misuse of sensitive information related to data governance. Some ways to apply POLP to governance metadata include role-based access control, need-to-know basis, regular audits, segregation of duties, and monitoring and logging.
Research Data Alliance	Research Data Alliance or RDA is an international initiative aiming to build the social and technical bridges to enable open data sharing that hosts a variety of interest groups and working groups, some of which focus on data governance and metadata.[128] The Research Metadata Working Group maintains a crosswalk of 10 common metadata schemas that includes ISO, DCAT, DATS, and DDI. The Metadata Standards Catalog Working Group has produced a machine-actionable catalog of metadata standards submitted by all RDA working groups.
Schema.org	Schema.org is a collaborative project founded by major search engines, including Google, Bing, Yahoo, and Yandex, that aims to create and maintain a structured data vocabulary for the internet.[129] This vocabulary, known as schema markup, helps search engines understand the content of Web pages and deliver richer, more relevant search results to users. The schema.org vocabulary covers a wide range of entities, such as people, organizations, events, products, and reviews, and is constantly evolving to accommodate new types of structured data. Schema.org offers a general framework and a set of best practices for metadata schema generation including identifying relevant schema.org types and properties such as Organization, Person, Government, CivicStructure, and Event, using JSON-LD as the recommended format for structured data, providing human-readable labels and descriptions for types and properties, using URIs to identify entities, testing and validating the schema using Google's Structured Data Testing Tool or the Schema Markup Validator, and publishing and sharing the validated schema.

Resource	Description and Relevance to the Project
The Social Data Foundation	The SDF (Social Data Foundation for Health and Social Care) acts as a new form of data institution that proposes data trust services as a sociotechnical model for good data governance by acting as a Trusted Research Environment.[130] The new data institution builds on the Data Foundations Framework and strong citizen representation. SDF ensures the citizen voice is not lost in the data lifecycle process. The overall purpose of SDF governance model is to facilitate the safe (re)usage of data through "well-defined data governance roles and processes" that builds "prompt and on-going risk assessment and risk mitigation into the whole data lifecycle." SDF's standardized and comprehensive framework for organizing and managing social data that ensures data quality, integration, privacy, security, and compliance with relevant regulations and best practices could inform the development of a governance metadata schema.

Resource	Description and Relevance to the Project
Trusted Exchange Framework and Common Agreement	The Trusted Exchange Framework and Common Agreement (TEFCA) is a set of principles and requirements developed by the Office of the National Coordinator for Health Information Technology in the United States.[131] TEFCA aims to enable a more secure and interoperable exchange of electronic health information across different health networks. The Trusted Exchange Framework describes high-level principles that networks should adhere to for trusted exchange. The Common Agreement is a legal agreement that will enable network-to-network data sharing. The Common Agreement will set minimum requirements to enable the appropriate sharing of electronic health information between networks. While TEFCA primarily focuses on health information exchange, some of its guidance can be applied to a data governance metadata schema in the context of data interoperability and security. TEFCA emphasizes the importance of interoperability in data exchange, which can be applied to a governance metadata schema by using standardized metadata models, vocabularies, and ontologies that promote seamless data integration across different systems and platforms. TEFCA highly recommends adopting widely recognized data exchange standards, such as HL7 FHIR, to ensure consistency and compatibility in the representation and exchange of governance metadata. TEFCA provides guidance on implementing privacy and security measures to protect sensitive information during data exchange. For a governance metadata schema, this includes access controls, encryption, and secure communication protocols to ensure the confidentiality and integrity of metadata. TEFCA highlights the importance of data quality such that governance metadata schemas can benefit from inclusion of quality checks, validation rules, and processes to maintain the accuracy, completeness, and consistency of metadata. TEFCA recommends building trust among metadata exchange partners by promoting transparency in the governance metadata schema. This includes documenting and sharing information about the schema's design, implementation, and management, as well as any privacy and security measures in place. Finally, TEFCA recommends a commitment to continuous improvement by regularly monitoring, reviewing, and updating the governance metadata schema based on new requirements, best practices, and lessons learned, to ensure it remains effective and relevant.

Resource	Description and Relevance to the Project
Vulcan FHIR Accelerator	The Vulcan Accelerator serves the needs of the clinical and translational research communities through the implementation of HL7 FHIR standardized data exchange.[132] The Vulcan FHIR Accelerator offers valuable resources and guidance for developing a governance metadata schema by promoting the adoption of FHIR standards and fostering collaboration among healthcare stakeholders. This ensures standardization, interoperability, data organization, security, and privacy in healthcare data management. The Vulcan FHIR Accelerator is working on a project to support the development of FHIR to OMOP transfer for better analysis of clinical data for research. The project is currently developing a FHIR server implementation built on top of the OMOP Common Data Model designed to provide a FHIR clinical API to read and write data from the OMOP database. Key takeaways and best practices for mapping this data can be utilized for the metadata schema. The Vulcan FHIR Accelerator also developed a FHIR IG to map FHIR data into CDISC.
Zero Trust Architecture	The zero-trust security model, also known as Zero Trust Architecture (ZTA), and sometimes known as perimeterless security, describes an approach to the strategy, design, and implementation of IT systems.[133] While ZTA primarily focuses on security, its principles can be applied to a governance metadata schema to enhance its protection and ensure the confidentiality, integrity, and availability of metadata. ZTA recommends least privilege access (see POLP) using role-based access controls; strong identity and access management such as multi-factor authentication and single sign-on, micro-segmentation of the metadata so that each metadata segment may have dedicated access controls and security; continuous monitoring and validation; standard data security practices such as encryption and secure communication protocols to ensure confidentiality; and adaptive policies that can respond to changes in the security landscape and user behaviors.

6 References

[1] *Eunice Kennedy Shriver* National Institute of Child Health and Human Development Office of Data Science and Sharing. PCORTF Pediatric Record Linkage Governance Assessment. [Internet]. December 2023. Available from: https://www.nichd.nih.gov/about/org/od/odss

[2] *Eunice Kennedy Shriver* National Institute of Child Health and Human Development Office of Data Science and Sharing. Privacy Preserving Record Linkage (PPRL) for Pediatric COVID-19 Studies. [Internet]. September 2022. Available from: https://www.nichd.nih.gov/sites/default/files/inline-files/NICHD_ODSS_PPRL_for_Pediatric_COVID-19_Studies_Public_Final_Report_508.pdf

[3] *Eunice Kennedy Shriver* National Institute of Child Health and Human Development Office of Data Science and Sharing. PCORTF Pediatric Record Linkage Governance Assessment. [Internet]. December 2023. Available from: https://www.nichd.nih.gov/about/org/od/odss

[4] National Institutes of Health Office of Data Science Strategy. Streamlining Access to Controlled Data at the NIH [Internet]. 2022. [cited 2023 Dec 11]. Available from: https://datascience.nih.gov/streamlining-access-to-controlled-data

[5] Office of the Assistant Secretary for Planning and Evaluation, U.S. Department of Health and Human Services. Building Data Capacity for Patient-Centered Outcomes Research. Office of the Secretary Patient Centered Outcomes Research Trust Fund Strategic Plan: 2020-2029. [Internet]. 2022 September [cited 2023 Dec 11]. Available from: https://aspe.hhs.gov/sites/default/files/documents/b363671a6256c6b7f26dec4990c2506a/aspe-os-pcortf-2020-2029-strategic-plan.pdf

[6] National Institutes of Health Office of Data Science Strategy. Streamlining Access to Controlled Data at the NIH [Internet]. 2022. [cited 2023 Dec 11]. Available from: https://datascience.nih.gov/streamlining-access-to-controlled-data

[7] *Eunice Kennedy Shriver* National Institute of Child Health and Human Development Office of Data Science and Sharing. PCORTF Pediatric Record Linkage Governance Assessment. [Internet]. December 2023. Available from: https://www.nichd.nih.gov/about/org/od/odss

[8] *Eunice Kennedy Shriver* National Institute of Child Health and Human Development Office of Data Science and Sharing. Privacy Preserving Record Linkage (PPRL) for Pediatric COVID-19 Studies. [Internet]. September 2022. [cited 2023 Dec 11]. Available from: https://www.nichd.nih.gov/sites/default/files/inline-files/NICHD_ODSS_PPRL_for_Pediatric_COVID-19_Studies_Public_Final_Report_508.pdf

[9] *Eunice Kennedy Shriver* National Institute of Child Health and Human Development Office of Data Science and Sharing. PCORTF Pediatric Record Linkage Governance Assessment. [Internet]. December 2023. Available from: https://www.nichd.nih.gov/about/org/od/odss

[10] OSTP (Office of Science and Technology Policy). 2013. Increasing Access to the Results of Federally Funded Scientific Research. Memorandum for the Heads of Executive Departments and Agencies from John P. Holdren. Washington, DC: OSTP.

[11] National Academies of Sciences, Engineering, and Medicine; Policy and Global Affairs; Board of Research Data and Information; Committee on Toward an Open Science Enterprise Open science by design: Realizing a vision for 21st century research. 2018 Jul 17; https://pubmed.ncbi.nlm.nih.gov/30212065/

[12] Sansone, SA, McQuilton, P, Rocca-Serra, P *et al.* FAIRsharing as a community approach to standards, repositories and policies. Nat Biotechnol. 2019;37:358–367. https://doi.org/10.1038/s41587-019-0080-8; FAIRsharing: https://fairsharing.org/

[13] Ulrich H, Kock-Schoppenhauer AK, Deppenwiese N, Gött R, Kern J, Lablans M, Majeed RW, Stöhr MR, Stausberg J, Varghese J, Dugas M, Ingenerf J. Understanding the Nature of Metadata: Systematic Review. J Med Internet Res. 2022 Jan 11;24(1):e25440. doi: 10.2196/25440. PMID: 35014967; PMCID: PMC8790684.

[14] Dahlquist JM, Nelson SC, Fullerton SM. Cloud-based biomedical data storage and analysis for genomic research: Landscape analysis of data governance in emerging NIH-supported platforms. Human Genetics and Genomics Advances. 2023 Jul 13;4(3).

[15] Batista D, Gonzalez-Beltran A, Sansone SA, Rocca-Serra P. Machine actionable metadata models. Sci Data. 2022 Sep 30;9(1):592. doi: 10.1038/s41597-022-01707-6. PMID: 36180441; PMCID: PMC9525592.

[16] Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, Appleton, G. Axton, M, Baak, A, ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3(1), 1-9. https://doi.org/10.1038/sdata.2016.18

[17] *Eunice Kennedy Shriver* National Institute of Child Health and Human Development Office of Data Science and Sharing. PCORTF Pediatric Record Linkage Governance Assessment. [Internet]. December 2023. Available from: https://www.nichd.nih.gov/about/org/od/odss

[18] Hillmann D, Guenther R, Hayes A. Metadata Standards and Applications Trainee Manual. [Internet]. Washington, D.C.; 2008. [cited 2023 Dec 11]. Available from: https://www.loc.gov/catworkshop/courses/metadatastandards/pdf/MSTraineeManual.pdf

[19] ISACA CMMI Performance Solutions. CMMI Institute Levels of Capability and Performance. [Internet]. [cited 2023 Dec 11]. Available from: https://cmmiinstitute.com/learning/appraisals/levels

[20] IT Governance. IT Governance Software Capability Maturity Model. [Internet]. [cited 2023 Dec 11]. Available from: https://www.itgovernance.co.uk/capability-maturity-model

[21] Pistola Alliance FAIR Toolkit. Data Capability Maturity Model. [Internet]. [cited 2023 Dec 11]. Available from: https://fairtoolkit.pistoiaalliance.org/methods/data-capability-maturity-model/

[22] U.S. Department of Labor Office of Data Governance. Data Maturity Model. [Internet]. [cited 2023 Dec 11]. Available from: https://www.dol.gov/agencies/odg/data-management-maturity-model

[23] UC San Diego Biomedical Informatics, Biomedical and Healthcare Data Discovery Index Ecosystem (bioCADDIE). [Internet]. [cited 2023 Dec 11]. Available from: https://dbmi.ucsd.edu/projects/biocaddie.html

[24] Bioschemas. What is Bioschemas. [Internet]. [cited 2023 Dec 11]. Available from: https://bioschemas.org

[25] Von Eschenbach AC, Buetow K. Cancer informatics vision: caBIG™. Cancer informatics. 2006; 2: 22–24. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2675495/

[26] Clinical Data Interchange Standards Consortium. The Clinical Data Interchange Standards Consortium Home Page. [Internet]. [cited 2023 Dec 11]. Available from: https://www.cdisc.org

[27] Creative Commons. About CC Licenses. [Internet]. [cited 2023 Dec 11]. Available from: https://creativecommons.org/share-your-work/cclicenses/

[28] National Library of Medicine: National Center for Biotechnology Information. The Database of Genotypes and Phenotypes (dbGaP). [Internet]. [cited 2023 Dec 11]. Available from: https://www.ncbi.nlm.nih.gov/gap/

[29] Global Alliance for Genomics & Health. Global Alliance for Genomics & Health. [Internet]. [cited 2023 Dec 11]. Available from: https://www.ga4gh.org

[30] Globus. Globus Toolkit. [Internet]. [cited 2023 Dec 11]. Available from: http://toolkit.globus.org

[31] InCommon. InCommon Federation. [Internet]. [cited 2023 Dec 11]. Available from: https://incommon.org/federation/

[32] Integrating the Healthcare Enterprise. Integrating Healthcare Enterprise. [Internet]. [cited 2023 Dec 11]. Available from: https://www.ihe.net

[33] The Kidney Precision Medicine Project. The Kidney Precision Medicine Project. [Internet]. [cited 2023 Dec 11]. Available from: https://www.kpmp.org

[34] Observational Health Data Sciences and Informatics. Observational Health Data Sciences and Informatics. [Internet]. [cited 2023 Dec 11]. Available from: https://www.ohdsi.org

[35] Open Biological and Biomedical Ontology Foundry. OBO Foundry. [Internet]. [cited 2023 Dec 11]. Available from: http://obofoundry.org

[36] The Research Data Alliance. The Research Data Alliance. [Internet]. [cited 2023 Dec 11]. Available from: https://www.rd-alliance.org

[37] Schema.org. Schema.org. [Internet]. [cited 2023 Dec 11]. Available from: https://schema.org

[38] Social Data Foundation. The Social Data Foundation. [Internet]. [cited 2023 Dec 11]. Available from: https://www.socialdatafoundation.org

[39] Vulcan Health Level Seven FHIR. Vulcan. [Internet]. [cited 2023 Dec 11]. Available from: https://hl7vulcan.org

[40] Elliot M, Mackey E, O'Hara K. The Anonymisation Decision-Making Framework: European Practitioners' Guide. 2nd Edition. [Internet]. Manchester (UK): UKAN; 2020 [cited 2023 Dec 11]. Available from: https://msrbcel.files.wordpress.com/2020/11/adf-2nd-edition-1.pdf

[41] The Global Data Management Community. The Global Data Management Community Publications. [Internet]. [cited 2023 Dec 11]. Available from: https://www.dama.org/cpages/body-of-knowledge

[42] National Institute of Standards and Technology. Cybersecurity Framework. Version 2.0 Draft. [Internet]. 2023. [cited 2023 Dec 11]. Available from: https://www.nist.gov/cyberframework

[43] Ross R, Pillitteri V, Dempsey K, Riddle M, Guissanie G. Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations. [Internet]. National Institute of Standards and Technology; 2020 Mar; Special Publication 800-171, Revision 2. [cited 2023 Dec 11]. Available from: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-171r2.pdf

[44] Health IT.gov. Trusted Exchange Framework and Common Agreement (TEFCA). [Internet]. [cited 2023 Dec 11]. Available from: https://www.healthit.gov/topic/interoperability/policy/trusted-exchange-framework-and-common-agreement-tefca

[45] Rose S, Borchert O, Mitchell S, Connley S. Zero Trust Architecture. [Internet]. NIST Special Publication 800-207. [cited 2023 Dec 11]. Available from: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-207.pdf

[46] Health Level Seven. FHIR Consent Resource. [Internet]. [cited 2023 Dec 11]. Available from: https://build.fhir.org/consent.html

[47] National Library of Medicine Unified Language System. UMLS Frequently Asked Questions. [Internet]. [cited 2023 Dec 11]. Available from: https://www.nlm.nih.gov/research/umls/faq_main.html

[48] National Library of Medicine. Unified Language System Metathesaurus License Agreement. [Internet]. [cited 2023 Dec 11]. Available from: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/license_agreement.html

[49] OMOP Common Data Model. Metadata Tables. [Internet]. [cited 2023 Dec 11]. Available from: OMOP CDM v5.4 (ohdsi.github.io)

[50] DCMI Usage Board Dublin Core. DCMI Type Vocabulary. [Internet]. 2014. [cited 2023 Dec 11]. Available from: https://www.dublincore.org/specifications/dublin-core/dcmi-type-vocabulary/

[51] Brickley D, Miller L. FOAF Vocabulary Specification. [Internet]. 2004 Aug. [cited 2023 Dec 11]. Available from: http://xmlns.com/foaf/0.1/

[52] Lin Y, Harris MR, Manion FJ, Eisenhauer E, Zhao B, Shi W, Karnovsky, He Y. Development of a BFO-based Informed Consent Ontology (ICO). Proceedings of the 5th International Conference on Biomedical Ontologies (ICBO). [Internet]. Houston, Texas, USA. 2014. Page 84-86. [cited 2023 Dec 11]. Available from: [http://ceur-ws.org/Vol-1327/icbo2014_paper_54.pdf]

[53] Lawson J, Cabili MN, Kerry G, Boughtwood T, Thorogood A, Alper P, Bowers SR, Boyles RR, Brookes AJ, Brush M, Burdett T, Clissold H, Donnelly S, Dyke SOM, Freeberg MA, Haendel MA, Hata C, Holub P, Jeanson F, Jene A, Kawashima M, Kawashima S, Konopko M, Kyomugisha I, Li H, Linden M, Rodriguez LL, Morita M, Mulder N, Muller J, Nagaie S, Nasir J, Ogishima S, Ota Wang V, Paglione LD, Pandya RN, Parkinson H, Philippakis AA, Prasser F, Rambla J, Reinold K, Rushton GA, Saltzman A, Saunders G, Sofia HJ, Spalding JD, Swertz MA, Tulchinsky I, van Enckevort EJ, Varma S, Voisin C, Yamamoto N, Yamasaki C, Zass L, Guidry Auvil JM, Nyrönen TH, Courtot M. The Data Use Ontology to streamline responsible access to human biomedical datasets. Cell Genom. 2021 Nov 10;1(2):None. doi: 10.1016/j.xgen.2021.100028. PMID: 34820659; PMCID: PMC8591903.

[54] Iannella R, Villata S. ODRL Information Model 2.2. [Internet]. W3C; 2018. [cited 2023 Dec 11]. Available from: https://www.w3.org/TR/odrl-model/

[55] Dublin Core. Creating Metadata User Guide. [Internet]. [cited 2023 Dec 11]. Available from: https://www.dublincore.org/resources/userguide/creating_metadata/#RightsHolder

[56] Palmirani M, Governatori G, Athan T, Boley H, Paschke A, Wyner A. OASIS LegaRuleML Core Specification. [Internet]. Version 1.0. OASIS; 2020. Available from: https://docs.oasis-open.org/legalruleml/legalruleml-core-spec/v1.0/cs02/legalruleml-core-spec-v1.0-cs02.html

[57] Iannella R, Villata S. ODRL Information Model 2.2. [Internet]. W3C; 2018. [cited 2023 Dec 11]. Available from: https://www.w3.org/TR/odrl-model/

[58] Palmirani M, Governatori G, Athan T, Boley H, Paschke A, Wyner A. OASIS LegaRuleML Core Specification. [Internet]. Version 1.0. OASIS; 2020. Available from: https://docs.oasis-open.org/legalruleml/legalruleml-core-spec/v1.0/cs02/legalruleml-core-spec-v1.0-cs02.html

[59] Palmirani M, Governatori G, Athan T, Boley H, Paschke A, Wyner A. OASIS LegaRuleML Core Specification. [Internet]. Version 1.0. OASIS; 2020. Available from: https://docs.oasis-open.org/legalruleml/legalruleml-core-spec/v1.0/cs02/legalruleml-core-spec-v1.0-cs02.html

[60] Iannella R, Villata S. ODRL Information Model 2.2. [Internet]. W3C; 2018. [cited 2023 Dec 11]. Available from: https://www.w3.org/TR/odrl-model/

[61] Health Level Seven. FHIR Data Segmentation for Privacy Implementation Guide. [Internet]. [cited 2023 Dec 11]. Available from: https://build.fhir.org/ig/HL7/fhir-security-label-ds4p/

[62] Health Level Seven. FHIR Core Security Labels. [Internet]. [cited 2023 Dec 11]. Available from: http://hl7.org/fhir/security-labels.html#core

[63] Palmirani M, Governatori G, Athan T, Boley H, Paschke A, Wyner A. OASIS LegaRuleML Core Specification. [Internet]. Version 1.0. OASIS; 2020. Available from: https://docs.oasis-open.org/legalruleml/legalruleml-core-spec/v1.0/cs02/legalruleml-core-spec-v1.0-cs02.html

[64] Iannella R, Villata S. ODRL Information Model 2.2. [Internet]. W3C; 2018. [Internet]. [cited 2023 Dec 11]. Available from: https://www.w3.org/TR/odrl-model/

[65] Global Alliance for Genomics & Health. Global Alliance for Genomics & Health Community Resources. [Internet]. [cited 2023 Dec 11]. Available from: https://www.ga4gh.org/product/data-use-ontology-duo/#:~:text=The%20GA4GH%20DURI%20Data%20Use,to%20a%20particular%20data%20set.

[66] Iannella R, Villata S. ODRL Information Model 2.2. [Internet]. W3C; 2018. [Internet]. [cited 2023 Dec 11]. Available from: https://www.w3.org/TR/odrl-model/https://www.w3.org/TR/odrl-vocab/

[67] Health Level Seven. FHIR Data Segmentation for Privacy Implementation Guide. [Internet]. [cited 2023 Dec 11]. Available from: https://build.fhir.org/ig/HL7/fhir-security-label-ds4p/

[68] Health Level Seven. FHIR Core Security Labels. [Internet]. [cited 2023 Dec 11]. Available from: http://hl7.org/fhir/security-labels.html#core

[69] DCMI Usage Board Dublin Core. DCMI Type Vocabulary. [Internet]. 2014. [cited 2023 Dec 11]. Available from: https://www.dublincore.org/specifications/dublin-core/dcmi-type-vocabulary/

[70] The Data Documentation Initiative. The Data Documentation Initiative Lifecycle 3.3. [Internet]. 2023. [cited 2023 Dec 11]. Available from: https://ddialliance.org/Specification/DDI-Lifecycle/3.3/

[71] National Institutes of Health Office of Data Strategy. Streamlining Access to Controlled Data at the NIH. [Internet]. [cited 2023 Dec 11]. Available from: https://datascience.nih.gov/streamlining-access-to-controlled-data

[72] Lockhart H, Parducci B. Abbreviated Language for Authorization. [Internet]. OASIS; 2015. [cited 2023 Dec 11]. Available from: https://github.com/axiomatics/alfa-vscode-doc/blob/main/index.html

[73] Casey K. What is Attribute Based Access Control. [Internet]. Okta; 2020. [cited 2023 Dec 11]. Available from: https://www.okta.com/blog/2020/09/attribute-based-access-control-abac/

[74] Ga4GH. The Automatable and Discovery Index. [Internet]. 2019. [cited 2023 Dec 11]. Available from: https://github.com/ga4gh/ADA-M

[75] EUDADT Collaborative Data Infrastructure. B2FIND. [Internet]. 2023. [cited 2023 Dec 11]. Available from: https://b2find.eudat.eu/

[76] Crossref. Crossref. [Internet]. 2023. [cited 2023 Dec 11]. Available from: https://www.crossref.org/

[77] Feeney, P. Crossref Documentation Schema Library. [Internet]. 2021. [cited 2023 Dec 11]. Available from: https://www.crossref.org/documentation/schema-library/required-recommended-elements/#005

[78] Microsoft. Dataverse. [Internet]. 2023. [cited 2023 Dec 16]. Available from: https://learn.microsoft.com/en-us/power-apps/maker/data-platform/data-platform-intro

[79] European Commission. European Open Science Cloud. [Internet]. 2023. [cited 2023 Dec 16]. Available from: https://digital-strategy.ec.europa.eu/en/policies/open-science-cloud

[80] Sacco O, Passant A. A Privacy Preference Ontology (PPO) for Linked Data. [Internet]. Semantic Scholar; 2011. [cited 2024 Jan 2]. Available from: https://ceur-ws.org/Vol-813/ldow2011-paper01.pdf

[81] Open Policy Agent. Rego. [Internet]. 2023. [cited 2023 Dec 16]. Available from: https://www.openpolicyagent.org/docs/latest/#rego

[82] UW Information Technology. Understanding SDDL Language. [Internet]. 2023. [cited 2023 Dec 16]. Available from: https://itconnect.uw.edu/tools-services-support/it-systems-infrastructure/msinf/other-help/understanding-sddl-syntax/

[83] Corcho O, Magnus E. Crosswalk of Most Used metadata schemes and guidelines for metadata interoperability. [Internet]. Zenodo; 2021. [cited 2023 Dec 16]. Available from: https://zenodo.org/records/4420116

[84] Clinical Data Interchange Standards Consortium. CDASH. [Internet]. 2023. [cited 2023 Dec 16]. Available from: https://www.cdisc.org/standards/foundational/cdash

[85] ISACA. COBIT an ISACA Framework. [Internet]. 2023. [cited 2023 Dec 16]. Available from: https://www.isaca.org/resources/cobit

[86] Datacite Schema. Datacite Metadata Schema. [Internet]. 2021. [cited 2023 Dec 16]. Available from: https://schema.datacite.org/meta/kernel-4.4/

[87] Garbin, Christian. Datasheet for Dataset Template. [Internet]. GitHub; 2020. [updated 2022 Sep 25; cited 2024 Jan 2]. Available from: https://github.com/fau-masters-collected-works-cgarbin/datasheet-for-dataset-template

[88] Gebru T, Morgenstern J, Vecchione B. Datasheets for Datasets. [Internet]. Cornell University; 2018. [updated 2021 Dec 1; cited 2023 Dec 16]. Available from: https://arxiv.org/abs/1803.09010

[89] Parducci B, Lockhart H. OASIS extensible Control Markup Language TC. [Internet]. OASIS. [cited 2023 Dec 16]. Available from: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml

[90] Health Level Seven FHIR. Resource: Provenance. [Internet]. 2023. [cited 2023 Dec 16]. Available from: https://hl7.org/fhir/R5/provenance.html

[91] Health Level Seven FHIR. US Core Implementation Guide. [Internet]. 2023. [cited 2023 Dec 16]. Available from: https://www.hl7.org/fhir/us/core/

[92] OBO Foundry. Information Artifact Ontology. [Internet]. 2022. [cited 2023 Dec 16]. Available from: https://obofoundry.org/ontology/iao.html

[93] OBO Foundry. Information Artifact Ontology, OntologyMetadata. [Internet]. GitHub; 2019. [cited 2024 Jan 2]. Available from: https://github.com/information-artifact-ontology/IAO/wiki/OntologyMetadata

[94] International Organization for Standardization. ISO/IEC 38500:2015 Information technology Governance of IT for the organization. [Internet]. 2015. [cited 2023 Dec 16]. Available from: https://www.iso.org/standard/62816.html

[95] OBI. Ontology for Biomedical Investigations. [Internet]. 2019. [cited 2023 Dec 16]. Available from: https://obi-ontology.org/

[96] OBI. Ontology for Biomedical Investigations, Core Classes. [Internet]. GitHub; 2018. [updated 2018 Oct 19; cited 2024 Jan 2]. Available from: https://github.com/obi-ontology/obi/wiki/Core-Classes

[97] OBI. Ontology for Biomedical Investigations, Example Use Cases. [Internet]. GitHub; 2018. [updated 2018 Oct 12; cited 2024 Jan 2]. Available from: https://github.com/obi-ontology/obi/wiki/Example_Use_Cases

[98] Herzog A, Shahmehri N, Duma C. An Ontology of Information Security. [Internet]. 2007. [cited 2023 Dec 16]. Available from: https://www.researchgate.net/publication /220065815_An_Ontology_of_Information_Security

[99] Herzog A, Shahmehri N, Duma C. An Ontology of Information Security, Main Security Ontology, Overview Illustration. [Internet]. 2007. [cited 2024 Jan 2]. Available from: https://www.ida.liu.se/divisions/adit/security/projects/secont/

[100] OpenAIRE Guidelines for Literature, institutional and thematic repositories. [Internet]. 2023. [cited 2024 Jan 3]. Available from: https://guidelines.openaire.edu/en/latest/literature/introduction.html

[101] Ciccarese P, Soiland-Reyes P, Belhajjame K, JG Gray A, Goble C, Clark T. PAV ontology: Provenance, Authoring and Versioning. [Internet]. Journal of Biomedical Semantics; 2013;4:37. [cited 2023 Dec 16]. Available from: https://github.com/pav-ontology/pav/wiki/Publications

[102] World Wide Web Consortium. PROV-O: The PROV Ontology. [Internet]. 2013. [cited 2023 Dec 16]. Available from: https://www.w3.org/TR/prov-o/

[103] SOSO. Science on Schema.Org. [Internet]. 2022. [cited 2023 Dec 16]. Available from: https://github.com/ESIPFed/science-on-schema.org

[104] Clinical Data Interchange Standards Consortium. Study Data Tabulation Model. [Internet]. 2023. [cited 2023 Dec 16]. Available from: https://www.cdisc.org/standards/foundational/sdtm

[105] Committee on Data International Science Council. Terms4FairSkills FAIR Data Stewardship Terminology. [Internet]. 2023. [cited 2023 Dec 16]. Available from: https://codata.org/initiatives/data-skills/wg-terms4fairskills/

[106] Committee on Data International Science Council. Terms4fairskills, FAIRterminology. [Internet]. GitHub; 2022. [updated 2023 Mar 2; cited 2024 Jan 2]. Available from: https://github.com/terms4fairskills/FAIRterminology/

[107] National Library of Medicine. UMLS Metathesaurus Browser. [Internet]. 2024. [cited 2024 Jan 2]. Available from: https://uts.nlm.nih.gov/uts/umls

[108] National Library of Medicine. UMLS Metathesaurus License Agreement. [Internet]. 2023. [cited 2023 Dec 16]. Available from: https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/license_agreement.html

[109] Health IT Interoperability Standards Advisory. US Core Data for Interoperability. [Internet]. Office of the National Coordinator; 2023. [cited 2023 Dec 16]. Available from: https://www.healthit.gov/isa/united-states-core-data-interoperability-uscdi#uscdi-v4

[110] Berners-Lee T, Story H, Capadisili S. Web Access Control. [Internet]. Web Access Control; 2023. [cited 2023 Dec 16]. Available from: https://solid.github.io/web-access-control-spec/

[111] Elliot M, Mackey E, O'Hara K. The Anonymisation Decision-Making Framework: European Practitioners' Guide. 2nd Edition. [Internet]. Manchester (UK): UKAN; 2020. [cited 2023 Dec 11]. Available from: https://msrbcel.files.wordpress.com/2020/11/adf-2nd-edition-1.pdf

[112] UC San Diego Biomedical Informatics, Biomedical and Healthcare Data Discovery Index Ecosystem (bioCADDIE). [Internet]. [cited 2023 Dec 11]. Available from: https://dbmi.ucsd.edu/projects/biocaddie.html

[113] Bioschemas. What is Bioschemas. [Internet]. [cited 2023 Dec 11]. Available from: https://bioschemas.org

[114] Von Eschenbach AC, Buetow K. Cancer informatics vision: caBIG™. Cancer informatics. 2006; 2: 22–24. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2675495/

[115] Clinical Data Interchange Standards Consortium. The Clinical Data Interchange Standards Consortium Home Page. [Internet]. [cited 2023 Dec 11]. Available from: https://www.cdisc.org

[116] Creative Commons. About CC Licenses. [Internet]. [cited 2023 Dec 11]. Available from: https://creativecommons.org/share-your-work/cclicenses/

[117] The Global Data Management Community. The Global Data Management Community Publications. [Internet]. [cited 2023 Dec 11]. Available from: https://www.dama.org/cpages/body-of-knowledge

[118] National Library of Medicine: National Center for Biotechnology Information. The Database of Genotypes and Phenotypes (dbGaP). [Internet]. [cited 2023 Dec 11]. Available from: https://www.ncbi.nlm.nih.gov/gap/

[119] Global Alliance for Genomics & Health. Global Alliance for Genomics & Health. [Internet]. [cited 2023 Dec 11]. Available from: https://www.ga4gh.org

[120] Globus. Globus Toolkit. [Internet]. [cited 2023 Dec 11]. Available from: http://toolkit.globus.org

[121] InCommon. InCommon Federation. [Internet]. [cited 2023 Dec 11]. Available from: https://incommon.org/federation/

[122] Integrating the Healthcare Enterprise. Integrating Healthcare Enterprise. [Internet]. [cited 2023 Dec 11]. Available from: https://www.ihe.net

[123] The Kidney Precision Medicine Project. The Kidney Precision Medicine Project. [Internet]. [cited 2023 Dec 11]. Available from: https://www.kpmp.org

[124] National Institute of Standards and Technology. Cybersecurity Framework. Version 2.0 Draft. [Internet]. 2023. [cited 2023 Dec 11]. Available from: https://www.nist.gov/cyberframework.

[125] Observational Health Data Sciences and Informatics. Observational Health Data Sciences and Informatics. [Internet]. [cited 2023 Dec 11]. Available from: https://www.ohdsi.orgl

[126] Open Biological and Biomedical Ontology Foundry. OBO Foundry. [Internet]. [cited 2023 Dec 11]. Available from: http://obofoundry.org

[127] Ross R, Pillitteri V, Dempsey K, Riddle M, Guissanie G. Protecting Controlled Unclassified Information in Nonfederal Systems and Organizations. [Internet]. National Institute of Standards and Technology; 2020 Mar; Special Publication 800-171, Revision 2. [cited 2023 Dec 11]. Available from: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-171r2.pdf

[128] The Research Data Alliance. The Research Data Alliance. [Internet]. [cited 2023 Dec 11]. Available from: https://www.rd-alliance.org

[129] Schema.org. Schema.org. [Internet]. [cited 2023 Dec 11]. Available from: https://schema.org

[130] Boniface M, Carmichael L, Hall W, Pickering B, Stalla-Bourdillon S, Taylor S. The Social Data Foundation model: Facilitating health and social care transformation through datatrust services. [Internet]. Cambridge University Press; 2022. [cited 2024 Jan 2]. Available from: https://www.cambridge.org/core/journals/data-and-policy/article/social-data-foundation-model-facilitating-health-and-social-care-transformation-through-datatrust-services/CD882977DA412B4020945C3FFE8725A0

[131] Health IT.gov. Trusted Exchange Framework and Common Agreement (TEFCA). [Internet]. [cited 2023 Dec 11]. Available from: https://www.healthit.gov/topic/interoperability/policy/trusted-exchange-framework-and-common-agreement-tefca

[132] Vulcan Health Level Seven FHIR. Vulcan. [Internet]. [cited 2023 Dec 11]. Available from: https://hl7vulcan.org

[133] Rose S, Borchert O, Mitchell S, Connley S. Zero Trust Architecture. [Internet]. NIST Special Publication 800-207. [cited 2023 Dec 11]. Available from: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-207.pdf