



Metadata Governance Framework for Public Health: A Practical Guide

**Prepared by Audacious Inquiry for the Council of State and
Territorial Epidemiologists (CSTE)**

Final Version

July 30, 2025

This publication was supported by cooperative agreement NU38PW000005. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the CDC.

Contents

Notes	3
Executive Summary	4
Key Highlights	4
Integration with Existing Governance	4
Next Steps & Recommendations.....	5
1 Introduction.....	6
2 Approach	6
3 Definitions	7
4 Before Starting.....	7
5 Assessment.....	7
5.1 Whom to Assess.....	7
5.2 Metadata Governance Assessment Levels	8
5.2.1 Level 0: No Capability	8
5.2.2 Level 1: Initial / Ad Hoc.....	8
5.2.3 Level 2: Repeatable	8
5.2.4 Level 3: Defined	8
5.2.5 Level 4: Managed	9
5.2.6 Level 5: Optimization.....	9
5.3 Assessment Questions	9
5.3.1 Agency Governance.....	9
5.4 Governance Processes, Policies, and Procedures	9
5.4.1 Policies and Procedures.....	9
5.4.2 Agency Metadata Assets	10
6 Evaluating the Assessment.....	10
6.1.1 Governance Engagement	11
6.1.2 Policy Development.....	12
6.1.3 Procedure Documentation	12
6.1.4 Version Control Practices.....	12
7 Metadata Governance Framework.....	13
7.1 Integrate with Existing Governance and Agency Priorities	13
7.2 Establish a Metadata Repository	15
7.2.1 Vital Capabilities for a Metadata Repository	15
7.2.2 Desirable Capabilities of a Metadata Repository	16
7.2.3 Using a Document Management Solution as a Metadata Repository.....	16
7.2.4 Metadata Artifact Conventions	19
7.2.5 Agency and Individual Directory.....	19
7.2.6 Change Control Management System	20

7.2.7	Shared Artifacts	20
7.3	Templates, Templates and More Templates	20
7.4	Incorporate Existing Knowledge into Infrastructure and Processes.....	20
7.5	Develop Data Set Catalog Entries	21
7.6	Develop Data Dictionaries.....	21
7.6.1	Constraints.....	22
7.6.2	Constraints Across Fields	23
7.7	Extract Data Schemas and Models from Existing Data Infrastructure	23
7.8	Gather Implementation Guide Resources	23
7.9	Transformation Scripts	23
7.10	Link Policy to Metadata Artifacts.....	24
7.11	Validation Scripts	24
7.12	Quality Measurement Scripts.....	25
Appendix A – Glossary of Terms		25
Metadata		25
Metadata Artifacts		25
Roles		26
Privacy		26
Policy Implementation		27
Additional Definitions		27

Notes

Technology products mentioned in this document do not indicate an endorsement of those products by CSTE or its contractors and may be trademarks of their respective owners.

Executive Summary

The Metadata Governance Framework for Public Health provides a comprehensive guide for state, Tribal, local, and territorial (STLT) public health agencies to implement and manage metadata governance practices effectively. It addresses the management of metadata, which is critical for ensuring data accessibility, security, and compliance with federal and state privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA), Confidentiality of Substance Use Disorder Patient Records (42 CFR Part 2), and the Family Educational Rights and Privacy Act (FERPA).

Key Highlights

- **Purpose:** The framework helps public health agencies establish a solid foundation for metadata governance, improving data interoperability and facilitating secure data sharing across public health jurisdictions.
- **Audience:** The framework is designed for agencies at various stages of metadata governance maturity, from early-stage implementations to mature agencies seeking continuous improvement.
- **Approach:** Using an assess, evaluate, and act approach, this framework provides an easy-to-understand roadmap to assess current metadata governance practices, evaluate results, and implement necessary changes to close identified gaps.
- **Metadata Governance Levels:** The framework includes five maturity levels, ranging from "No Capability" to "Optimization," offering agencies a way to gauge where they stand in terms of their metadata governance practices.
- **Key Components:** The framework incorporates essential elements, such as:
 - Defining roles like data owners and stewards.
 - Identifying and developing a metadata repository for organizing and managing metadata artifacts.
 - Ensuring proper version control and security measures for metadata management.
 - Integrating metadata governance into existing governance structures and agency priorities.
 - Offering best practices for data catalogs, dictionaries, and schemas to support metadata management and improve data quality.

Integration with Existing Governance

The document emphasizes the need for aligning metadata governance with the agency's broader data governance practices, ensuring that metadata processes and policies are documented, standardized, and continually refined. It also stresses the importance of involving key stakeholders, including senior leadership, compliance officers, IT staff, and data stewards, to drive effective governance processes.



Next Steps & Recommendations

- **Assess Current Practices:** Agencies should conduct an initial assessment of their metadata governance practices, evaluate existing resources, and identify areas that need improvement.
- **Develop a Plan:** Based on the assessment, agencies should develop a roadmap for establishing metadata governance, integrating with broader data governance efforts, and deploying necessary technologies.
- **Utilize Available Resources:** [CSTE Connect](#) offers a resource library that includes templates, tools, and additional guidelines to support the implementation of this framework.



1 Introduction

This document outlines a framework for state, Tribal, local, and territorial (STLT) public health agencies to implement effective metadata governance practices. The framework ensures that metadata such as data origin, access controls, privacy requirements, and usage constraints is appropriately managed and protected as part of broader data governance efforts. It is specifically designed to support public health agencies at varying stages of metadata governance maturity.

Effective metadata governance is essential for ensuring compliance with federal and state data privacy laws, including the Health Insurance Portability and Accountability Act (HIPAA), Confidentiality of Substance Use Disorder Patient Records (42 CFR Part 2), the Family Educational Rights and Privacy Act (FERPA), and state-specific health information privacy statutes. It also plays a crucial role in enhancing data interoperability and facilitating secure data sharing across public health jurisdictions.

Building a solid foundation for metadata governance involves adopting key governance practices that align with broader data governance frameworks. The CSTE Metadata Governance Workgroup has played a pivotal role in shaping this framework, tailoring recommendations to support jurisdictions at varying stages of metadata governance maturity. The Data Management Association International's (DAMA) Data Management Body of Knowledge (DMBOK) serves as a foundational reference for best practices and tools, guiding the implementation of metadata management, data stewardship, and policy development. Throughout this document, various sections of the DMBOK will be referenced using the format (DMBOK: 12:1.3), where the number before the colon is the chapter number, and what follows is the section within the chapter.

To assist public health agencies in applying this framework, some of the tools, templates, and examples referenced throughout the document are housed in the [CSTE Connect](#) Resource Library. This repository offers additional resources and templates that can be used as part of the implementation process.

The integration of metadata governance into existing governance systems is critical for success. It is essential to assess and integrate current data governance practices within the agency, including identifying roles such as data owners, data stewards, and metadata custodians. Governance policies and procedures should be documented, standardized, and aligned with agency goals. This document will also examine the roles of data catalogs, data dictionaries, and business glossaries in managing metadata, enabling jurisdictions to develop comprehensive records that document data quality, accessibility, and data lineage.

2 Approach

This framework is designed for agencies that are just starting to establish metadata governance. It helps to quickly reach the 'Defined' stage of governance, as described in DMBOK 12:1.3 and refined in Metadata Governance Assessment Levels. The framework assumes agencies have limited resources and can use their existing or easily accessible technology to manage their metadata, supporting wider data governance efforts. The framework does not preclude agencies that choose to invest in enterprise data



catalog management tools from applying the techniques it describes. However, those techniques may require modifications tailored to the selected tool.

The approach used in this framework is assess, evaluate, and act. This document outlines an approach that enables agencies to determine their metadata governance maturity, consider the assessment results, and then take action to close identified gaps. This process is repeated periodically (e.g., annually, or at the inception and completion of each project) to ensure the continuous evolution of metadata governance activities.

3 Definitions

Before beginning a process to establish a framework for metadata governance, it is essential to define a few key terms. Please see [Appendix A](#) for a glossary of terms. Additionally, the Public Health Informatics Institute (PHII) has also established several definitions, which can be found at <https://bit.ly/phii-data-governance-definitions>.

4 Before Starting

Before establishing a Metadata Data Governance practice within an agency, division, or department, it is essential to assess the existing practices within the agency. Ideally, governance practices for data management, of which metadata management is a key component, should be managed at the agency level and be commonly understood throughout the agency. Key individuals within the agency, such as the Chief Data Officer, Chief Information Officer, or Chief Technology Officer, may provide valuable insights. Key participants in this inquiry should include legal and compliance leaders, data modernization leaders, program or division subject matter experts, and lead data analysts.

5 Assessment

Assessing an agency's current metadata governance is essential before launching a governance program, as it helps identify areas that need attention and enables the agency to prioritize its efforts effectively. Regular assessments (e.g., annually) should be conducted to support continuous improvement of metadata governance. Additionally, it's recommended to perform evaluations at key project milestones (e.g., at the start and completion of metadata governance initiatives) to evaluate how agency governance impacts the project.

5.1 Whom to Assess

The assessment instrument, described in more detail below, is intended for use by individuals responsible for metadata governance. This includes the chief officer roles and compliance leaders named in the previous section, as well as individuals in defined data governance roles, such as data owners, stewards, and custodians. These individuals should be familiar with the agency's metadata governance processes, resources, and capabilities.



Assessing individuals who are not directly involved in metadata governance processes provides valuable insights into the agency's staff's awareness of its governance efforts. This feedback can help identify gaps in communication or education about metadata governance processes and offer a reality check on how effectively those processes are being executed.

If the agency does not already have a well-established data governance process, this initial reality check is not essential in the initial assessment. However, if the agency believes it already has a well-established data governance program, this perception-reality check may be a crucial step to complete. When there is a significant difference (e.g., more than 1 point score) in the responses between those who should know (e.g., directly involved in metadata governance), and those who need not know (e.g., stakeholders benefiting from metadata governance but not directly involved in the details), more communication may be necessary, or the execution of the affected metadata governance processes should be more closely examined.

5.2 Metadata Governance Assessment Levels

DMBOK 12:1.3 describes five levels of data governance maturity, which this framework adopts with modest changes to address metadata governance specifically. These are further described below.

5.2.1 Level 0: No Capability

There are no organized metadata management practices or formal processes for managing data. The agency may be familiar with the concept of “metadata,” but it is unsure where to start managing it.

5.2.2 Level 1: Initial / Ad Hoc

The agency utilizes a general-purpose metadata management system with a limited toolset, lacking effective governance. Metadata creation and management are highly reliant on a small number of experts. Roles and responsibilities in metadata governance are often defined within silos, such as divisions, departments, or programs, with limited connections to the agency's broader data governance processes. Solutions for managing metadata are limited. Metadata is defined inconsistently across the agency.

5.2.3 Level 2: Repeatable

Emergence of consistent tools and role definitions to support process execution. In level 2, the agency provides central oversight and tools for metadata management and governance. Specific roles and responsibilities are assigned regarding metadata governance, and processes are no longer entirely dependent on particular experts. There is awareness of metadata and related concepts. At this level, the agency may begin to define, document, govern, and manage metadata processes and tools.

5.2.4 Level 3: Defined

Emerging metadata governance capability. Level 3 introduces scalable metadata management processes and tools, and views metadata management as a key component supporting data governance. At this stage, processes become more uniform across the agency, and they can begin to leverage automation capabilities.

5.2.5 Level 4: Managed

At this stage, the agency incorporates metadata governance processes into new projects that establish new datasets or update existing ones. Metadata is routinely used in automated tasks to support the delivery of data catalogs, dataset documentation, interface validation, and the creation of data subsets (with their metadata) for specific research programs or investigations.

5.2.6 Level 5: Optimization

At this stage, metadata governance processes are optimized and continuously improved. They are routinely measured to identify areas for further improvement.

5.3 Assessment Questions

The following questions enable an assessment of an agency's metadata governance capabilities. These questions support answers on a no-to-yes scale with intermediate degrees of uncertainty. The suggested responses and scoring are:

- 0. No
- 1. Likely Not
- 2. Likely So
- 3. Yes

Unanswered questions or questions answered as Not Applicable should not be used to compute an overall assessment score. Note that "Unknown" is not a permitted answer. This is intentional, as these questions require knowledge and perception. When the information is unknown, agencies need to understand how metadata governance is perceived by its stakeholders.

5.3.1 Agency Governance

(DMBOK: 1.3.2)

- 1. Does the agency (e.g., state government, public health department, or other relevant entity) have a data governance steering committee, council, or office?
- 2. Is the leadership of the committee, council, or office defined?
- 3. Are the agency's data stewards, who are responsible for creating, managing, and maintaining metadata, identified?
- 4. Does the program, division, or department (e.g., within the health department or another unit) have its data governance committee or workgroup?
- 5. Are the individuals involved in or leading these activities known?

5.4 Governance Processes, Policies, and Procedures

5.4.1 Policies and Procedures

- 1. Are the agency's data governance or metadata governance processes and policies written down?
- 2. Are they accessible to everyone who needs them?
- 3. Do all relevant staff members know where to find them?

4. Are they managed; in other words, is there a process for updating and approving them?
5. Are they version-controlled; can previous revisions of processes or policies be located?

5.4.2 Agency Metadata Assets

The following questions explore the metadata assets an agency uses to support its data governance activities. These activities are the core components of metadata governance.

Business Glossary (*DMBOK 3:2.14, 3:3.2, 12:1.3.5.2*)

1. Does the agency have a business glossary?

Business Metadata (*DMBOK 12:1.3.2.1*)

1. Does the agency maintain a data catalog?
2. If the agency does not have a catalog, are there written records describing the data maintained?
3. Do the records, data catalog, or associated policies document who can access data (Data Users) and the policies and procedures for obtaining access?
4. Do the records or data catalog provide any description of the sensitivity of the data?
5. Do the records indicate who provides access (Data Custodian)?
6. Do the records specify who is responsible for maintaining the data's accuracy and quality (Data Stewards)?
7. Do the records describe the data source or sources, including where the data originated and/or which agencies it came from? This is also known as the provenance of the data.
8. Are Data Catalog entries version-controlled; can previous revisions of data catalog entries be located?

Technical Metadata (*DMBOK 12:1.3.2.2*)

1. Does the agency maintain data dictionaries for the data being managed?
2. Does the agency distinguish between a data dictionary and a data model or schema?
3. Do the agency's data dictionaries describe how data values are established (e.g., what interface fields are used to populate them), derived, or computed, potentially from other data elements?
4. Is the location and format of the agency's data dictionaries known?
5. Are there important artifacts the agency uses to manage its data or metadata under version control?

Technology Infrastructure (*DMBOK 12:1.3.5.1*)

1. Does the agency deploy any metadata or data management technologies?
2. Does the agency have an enterprise data catalog application?
3. Document management and control systems can be a necessary technology infrastructure for agencies without an enterprise data catalog application. Examples include SharePoint, OneDrive, Google Docs or Folders, Box, GitHub, Confluence, or an agency wiki. Does the agency use a document management or document control system to manage documentation related to metadata?

6 Evaluating the Assessment

Once relevant stakeholders have completed the assessment instrument, it will be evaluated. The output of the evaluation process is an improvement plan that addresses the gaps identified during the assessment.



The Assessing Metadata Governance spreadsheet contains assessment questions, the answer key, and an evaluation tab to facilitate assessment. The questions are divided into three groups, and the Metadata Assets group is further subdivided into sections for the Business Glossary, Business Metadata, and Technical Metadata artifacts.

The evaluation process is described below.

1. Collect the responses.
2. Count the answers of each type.
3. Score them as described in section 5.3 above.
4. Compute the final score as the total score for each answer divided by the number of answers. The Evaluation tab in the spreadsheet will automatically compute these values.
5. Average the scores within each section to obtain a section score.
6. Evaluate the scores.

Analyze each section looking for significant gaps in either capabilities or awareness. If some responses are “no” or “likely not”, and others are “yes” or “likely so”, reconcile these different answers. Suppose the answer indicates a lack of awareness of a capability. Determine the best way to address that lack of understanding, e.g., through education, documentation, or a combination of both. If, instead, it identifies a gap in a capability, address that by creating a plan to develop it.

The assessment questions are prioritized in a way that facilitates the development of a plan of action. Questions coming earlier in the instrument address capabilities that should be improved before working on later ones.

It is not uncommon for organizations to have some gaps in capabilities in earlier sections but have well-evolved capabilities in later sections. For example, an organization may have well-structured data dictionaries but may be lacking a comprehensive data catalog or business glossary. In these cases, there are likely missing links in the metadata between later and earlier capabilities.

In the example above, given a data dictionary, a user should be able to identify the data catalog entries to which it applies, and given a data catalog entry, a user should be able to identify the relevant data dictionaries that apply to it. This requires the addition of a link from one artifact (e.g., the data dictionary) to the other (e.g., the data catalog). We recommend including those links in the later artifact. For example, the data dictionary should contain metadata that links to the data catalog entry metadata artifact. This allows for the evolution of metadata in the order in which it would typically be developed in a well-managed metadata development process. In another example of a missing business glossary, data dictionaries and data catalog entries are likely to use different terms to describe available data. These artifacts can be mined for key terms that should be included in a business glossary and may need to be revised to ensure that key business terms are used correctly. Data dictionaries often employ terminology that differs from that used in a data catalog. Rather than modify the data dictionary to use the business glossary term, provide additional information (metadata) in the data dictionary to link a data element definition to a business glossary entry to show this relationship.

The sections below outline actions based on the identified gaps in each section.

6.1.1 Governance Engagement

Purpose: Engage with existing agency data governance or establish it if necessary

Actions:

- Question 1.1 and 1.2: Identify and engage with existing agency data governance leadership and stakeholders.
- Question 1.3: Identify data stewards within the department, program, or agency.
- Question 1.4 and 1.5: This is not essential and depends upon the agency's governance model. If the agency has individual governance groups within divisions, programs, or other subdivisions, either engage with or create such a governance group within the subdivision.

6.1.2 Policy Development

Purpose: Define the policies needed to govern the agency's management and use of metadata.

Actions:

- Question 2.1: Identify the key areas of metadata governance that require formal policies (e.g., data access, retention, privacy).
- Question 2.1: Ensure policies align with legal and regulatory requirements (e.g., HIPAA, 42 CFR Part 2, FERPA).
- Question 2.2 and 2.3: Communicate policies to all relevant staff and ensure they are easily accessible.
- Question 2.4: Engage relevant stakeholders (e.g., data owners, data stewards, legal, compliance) to draft and review policies.

6.1.3 Procedure Documentation

Purpose: Document the implementation processes for policies describing how they are performed within the agency.

Actions:

- Question 2.1: For each policy, develop clear, step-by-step procedures for performing tasks (e.g., data entry, data sharing, data validation).
- Question 2.1 and 2.4: Ensure procedures are aligned with workflows and responsibilities.
- Question 2.2 and 2.3: Maintain procedure documentation in a centralized location for easy access.
- Question 2.4: Regularly review and update procedures to reflect changes in policies or regulations.

6.1.4 Version Control Practices

Purpose: Track changes and update metadata artifacts (business glossary, policies, procedures, business and technical artifacts) to ensure the agency always uses the most current version and can track the changes and reasons for change.

Actions:

Note: actions within this section address gaps found through responses to Questions 2.5, 3.3.4, and 3.3.5:

- Implement a version control system to track all changes made to artifacts.
- Assign version numbers and provide change logs with details of updates. Note, a good change log includes not just what changed, but why it changed, either directly or through a link to relevant material (e.g., a change request).

- Ensure that all staff are using the most recent version of documents.
- Provide access to previous versions when needed for audit or review purposes.

7 Metadata Governance Framework

7.1 Integrate with Existing Governance and Agency Priorities

Establish a metadata governance framework within the existing structures of the agency's broader data governance program.

Identify the level of the agency that will be targeted for establishing the metadata governance framework. Ideally, the target will be the entire public health agency. However, this may not be feasible for various reasons. Generally, it is recommended to target the highest level in the agency hierarchy as possible. Metadata governance is a critical component of data governance and is typically an enterprise-wide activity.

Identify and engage key individuals in the agency's data governance to establish the metadata governance program. When direct access to senior leadership is not possible (e.g., Chief Executive Officer), it is essential to engage leaders at the division, departmental, or program level who are responsible for overseeing and directing data governance. Begin with the individual who has been designated as accountable for data governance when known.

The key individuals to include are:

- Senior leadership roles (e.g., Chief Executive Officer, Executive Directors, Managers, or other designated leaders)
- Compliance and legal roles
- IT staff responsible for the implementation of key information systems, providing data, or managing metadata
- Data Owners
- Data Stewards
- Data Custodians
- Key subject matter experts and Data or Business Analysts
- Communications and Education staff

Identify the executive sponsor of the metadata governance activity. This may be the Chief Data Officer, Chief Technology Officer, Chief Information Officer, or another executive (e.g., Data Modernization Leader) with a strong interest in the success of metadata governance activities. The executive sponsor should be accountable for the metadata governance program. They should have the authority and access to the necessary resources to implement it, as well as the associated policies and procedures needed for its establishment.

In DMBOK 3: 1.3.5, the authors recognize and quote the first edition: "the best stewards are often found, not made." They further clarify that "... in most agencies, there are people who steward data, even in the absence of a formal Data Governance function." Identify those individuals where they exist. Ensure that

you engage them as data stewards in the metadata governance process. In DMBOK 3:1.3.4, the authors identify four key functions of data stewards:

1. Creating and managing core metadata
2. Documenting rules and standards (more metadata)
3. Managing data quality issues
4. Executing operation data governance activities

Thus, data stewards are the key drivers in metadata governance. In a RACI matrix, the executive sponsor may be the accountable party, but the data stewards are the parties responsible for ensuring that metadata governance is implemented.

The RACI matrix below suggests the type of engagement for each stakeholder.

Individuals	Accountable	Responsible	Consulted	Informed
Senior leadership roles	√		√	√
Compliance and legal			√	√
IT Staff		√		√
Data Owners	√		√	√
Data Stewards		√		
Data Custodians		√	√	
Subject Matter Experts and Analysts		√	√	
Communication and Education			√	√

Align data governance initiatives with key agency priorities. If the agency focuses on electronic case reporting, focus on that area or determine how to integrate it with that priority.

Assess metadata governance for the program. The Assessment instrument described above can be used, or the agency can use a different instrument, or engage with data governance experts to provide a more formal assessment. This framework assumes that the evaluation will focus on metadata governance.

Evaluate and prioritize which of the key areas of metadata governance to focus on based on the assessment results. The sections below outline tasks to consider in each area within this framework.

7.2 Establish a Metadata Repository

A metadata artifact is a file or document stored on a computer. The term artifact is used here because not all “documents” can be read easily by a human without using appropriate technology. A metadata repository is a system designed to store, organize, locate, edit, and manage a collection of artifacts. It should not be assumed that all metadata artifacts consist of human-readable words. The repository serves as a tool for the efficient management of metadata, which may include a variety of formats beyond just readable text.

Identify, acquire, or create/configure a technology base to serve as a repository of metadata artifacts, including the business glossary, dataset descriptions, policies and procedures, data dictionaries, data schemas, data models, value sets (constraints on data elements), and national and local implementation guides.

If you already have an enterprise data catalog, use it. If you do not, and one is not on the horizon to be acquired soon (e.g., less than a year), consider document management and control systems as an alternative.

Existing applications within your agency, such as OneDrive, SharePoint, Confluence, Git (including hosted and Software-as-a-Service (SaaS) versions of GitHub and GitLab), Google Drive, and Google Docs, can serve as a basis for a repository of metadata artifacts. Below is a list of key technical capabilities for your agency’s metadata repository.

7.2.1 Vital Capabilities for a Metadata Repository

7.2.1.1 Version Control

It is essential to track individual versions of metadata artifacts and retrieve older versions when necessary. Systems supporting version control ensure that each document change is saved, allowing easy access to prior versions when needed. A version control system is crucial for maintaining a historical record of metadata changes.

Security

- **Integration with Authorization and Authentication Processes:** The metadata repository should integrate seamlessly with your agency’s existing authorization and authentication systems. Many agencies use Microsoft Active Directory, and most cloud-based document management solutions (e.g., SharePoint, Box, Google Drive) can integrate with Active Directory and other Identity and Access Management (IAM) solutions.
- **Access Control:** Managing access to specific metadata artifacts is crucial. Metadata, like any other data, can be sensitive. For instance, information detailing which agencies supply or use aggregated data may be confidential and should not be publicly disclosed. Access control features ensure that only authorized personnel can view or modify sensitive metadata.
- **Audit:** Auditing is essential for tracking updates to metadata artifacts. Knowing who updated an artifact and when is crucial for maintaining data integrity and ensuring accountability in the metadata management process.

7.2.1.2 Integration with Document Editing Tools

Enabling users to edit metadata artifacts within the system that stores them directly is crucial. This minimizes barriers to maintaining metadata, making the governance process more efficient. The ability to seamlessly edit documents within the repository simplifies ongoing management tasks.

7.2.1.3 Hierarchical Structures

A metadata repository should employ a clear hierarchical structure (e.g., folders or similar) that classifies metadata artifacts based on relevant criteria (e.g., programs, data sets, systems). This facilitates the grouping and retrieval of related artifacts. The folder structure must be intuitive and easy to navigate.

7.2.2 Desirable Capabilities of a Metadata Repository

7.2.2.1 Linking

Linking metadata artifacts across multiple folders or datasets enhances their accessibility. For example, users may need to group shared artifacts or relate multiple datasets to one another. Linking allows these relationships to be made visible, streamlining the process of managing and accessing metadata across the system.

7.2.2.2 Tagging and Metadata

Many document management solutions (e.g., SharePoint, Google Docs, Box) allow metadata to be tagged or labeled with key-value pairs. This tagging helps classify artifacts, making them easier to find through search functionality. Additionally, custom metadata can be integrated to enrich search and categorization further, improving data discoverability.

7.2.2.3 Change Approval Processes

It is beneficial for the metadata repository to support an approval workflow before artifacts are updated or changed. Many document management systems do not directly automate change control processes; however, these processes can be facilitated by effectively managing access controls. Document management systems specifically designed for this purpose often include built-in features for change approval.

7.2.3 Using a Document Management Solution as a Metadata Repository

When using a document management solution, consider how to organize the metadata that will be contained within it to:

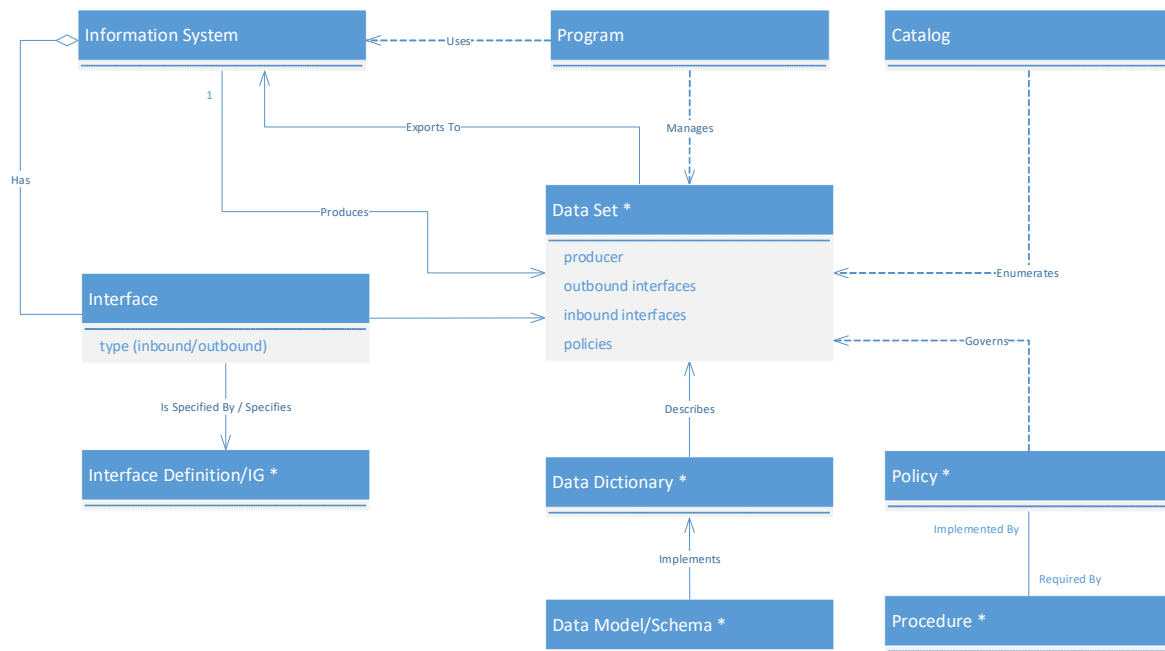
- Enable easy access to metadata artifacts
- Group related artifacts together
- Tag or label artifacts
- Associate additional metadata with artifacts

7.2.3.1 Metadata Hierarchical Structure

One key element in this process will be determining the overall hierarchical (e.g., folder) structure for the repository. Folders should neither be nested too deeply nor contain too many individual artifacts. Folder nesting structures are recommended to be no more than three to five levels deep, starting from the root

folder location in a metadata repository. It is also recommended that folders contain no more than 20 artifacts. When a folder contains more than that number, the artifacts could be grouped into subfolders underneath the original folder.

The image below provides a high-level model for metadata. Artifacts in blue boxes with an * after their name are metadata artifacts (others are entities). This diagram shows that the data set is central to all relationships.



7.2.3.2 A Suggested Hierarchy

The constraints and observations above lead to the following recommendation for a metadata folder structure.

- Metadata
 - Data Sets
 - Data Set 1 Name
 - Data Set Description
 - Data Dictionary Artifact
 - Data Model Artifact
 - Data Schema Artifact
 - Value Sets
 - Implementation Guides
 - Transformation Scripts
 - Transformation 1 Folder
 - Data Validation Scripts
 - Quality Measurement Scripts
 - Data Set 2 Name
 - ⋮
 - Shared Artifacts
 - Data Models and Schemas
 - Value Sets
 - Transformations
 - Policies
 - Policy 1
 - Policy 2
 - ⋮
 - Procedures
 - Procedure 1
 - Procedure 2
 - ⋮
 - Templates
 - Data Set Templates
 - Data Set Description Template
 - Data Dictionary Template
 - Value Set Template
 - Transformation Template
 - Policy Template
 - Procedure Template
 - Change Control Request Template



The top-level "Metadata" folder contains subfolders for Data Sets, Policies, Procedures, and Templates.

- Policies and Procedures folders house official documents, which may include Word-processor files, Markdown, wiki text, or application-specific markup (e.g., Confluence). Using markup formats like Markdown can allow for metadata extraction (e.g., approval dates, effective dates, applicable data sets).
- For word-processor documents, it is recommended to use templates that incorporate document properties, enabling automatic generation of metadata. This ensures easy access to document details and supports advanced search capabilities.
- These folders may also contain working copies of policies and procedures under revision, with restricted access.
- Data Sets folders contain individual subfolders for each dataset, including a description, data dictionary, data models, and referenced value sets. These value sets may exist in different formats, such as CSV for human readability or as FHIR resources for interoperability. Managing multiple artifact representations requires coordinated processes, such as creating transformation artifacts to maintain consistency.

7.2.4 Metadata Artifact Conventions

Metadata governance should define the types of artifacts for each metadata folder, such as word processing documents, CSV files, SQL table definitions, or UML diagrams. For SQL files, the specific dialect should be identified. It is also essential to set standards for the structure of these artifacts, such as column headings for CSV files or formatting for SQL queries.

The most crucial aspect of artifact conventions is naming. Clear, consistent naming ensures that metadata artifacts are easily identifiable and retrievable. Artifact names should be unique within each folder, with minimal exceptions. This helps avoid confusion, especially when only the file name is available. For example, when an artifact is found on a system and only the file name is visible, it should be easy to understand its content.

Names should be short, descriptive, and distinct. "Distinct" means the name cannot be easily confused with others, even if it is unique. For example, "Johnston" and "Jonston" are unique but not distinct.

To balance descriptiveness with brevity, longer names may be needed. A good naming format for metadata artifacts related to datasets is: Dataset Name, Artifact Component Name, Artifact Type, followed by the file extension (e.g., "Vaccination Data Patient Table.sql"). This ensures that artifacts are linked to their datasets.

To address the challenge of naming while keeping names manageable, shorter names for datasets can be used and incorporated into metadata artifact names. This adds another layer of metadata in the dataset description.

7.2.5 Agency and Individual Directory

Metadata often identifies individuals and agencies associated with the data. This information, along with their relationships with datasets, should be maintained in a directory. While systems like LDAP and Active

Directory help store metadata, they are not ideal for non-IT staff. For easier management, a Customer Relationship Management (CRM) system can be an effective tool for organizing these relationships.

7.2.6 Change Control Management System

Metadata governance involves managing changes over time. A ticket management system (e.g., Jira, Zoho) is essential for controlling these changes, as it enables the documentation, review, and approval of changes. Changes are tracked through a ticketing process, with restricted transitions to ensure proper approval before implementation.

7.2.7 Shared Artifacts

Effective metadata governance ensures consistency across datasets. Shared artifacts, such as value sets, are key examples. Common data elements, such as patient demographics (e.g., race, ethnicity, gender), are often used across datasets with consistent constraints. Managing these shared artifacts separately ensures uniformity across multiple datasets.

7.3 Templates, Templates and More Templates

To ensure consistency in creating, managing, and using metadata, templates are essential. A template provides a framework for starting an artifact, ensuring consistency in structure. The creator can then focus on filling in the necessary details. Templates are often implemented as forms within systems, such as ticketing or CRM platforms.

In the early stages of a governance program, agencies may spend considerable time developing and refining templates. However, it's important not to wait until templates are finalized to begin creating metadata artifacts. Templates should be developed in conjunction with their implementation, allowing issues to be identified and resolved early. Work is not complete until it is tested, and templates should be used in real-world scenarios for refinement.

Templates play a critical role in ensuring key metadata is included in artifacts. Most office applications and ticket management systems support templates, making them a standard tool in well-executed metadata governance programs. These templates streamline the change management process by ensuring all necessary details are entered upfront, improving approval speed, and reducing rework.

The templates folder serves as a repository for these templates, organized similarly to other metadata. Often, one or a few templates will replace multiple artifacts in the top-level metadata repository.

7.4 Incorporate Existing Knowledge into Infrastructure and Processes

Existing documents, publications, website content, spreadsheets, schema, and data model documentation, scripts, and other materials used to work with datasets within an agency should be collected, inventoried, and analyzed during the development of an initial metadata governance effort. These files will often be a considerable source of metadata about agency datasets. Collecting this data can provide benefits to an organization in the initial stages of establishing metadata governance. The files collected not only can provide metadata for the repository, but are also tied to existing organizational practices that produced them. These practices help to identify organizational procedures and enable alignment with metadata governance. They further identify practices that should be documented as procedures, and which may need adjustment to adapt to governance activities.

7.5 Develop Data Set Catalog Entries

The data set catalog entry links the various metadata components. It ties them back to data owners, stewards, and maintainers, as well as the information systems and agency structures that manage and utilize the systems that hold the data (e.g., division, department, program). It describes the data and links to the policies and procedures used to manage and access them. A well-defined data set description is an agency's entry in its data catalog.

Metadata Field	Description
Data Set Name	The official name of the dataset as it appears in a data catalog.
Short Name	A short, unique name for the dataset
Custodian	The individual to whom access requests to the dataset should be directed.
Owner	The agency subdivision that is responsible for the dataset.
Steward	The steward assigned to the dataset.
Authors	Authors of the data catalog entry.
Web Site Links	Links to websites related to the dataset.
Sources of Information	A description of the source or sources of data in this dataset.
Description of the Data	A detailed description of the data.
Strengths	A description of the strengths of this dataset.
Limitations	A description of the limitations of this dataset.
Uses of the Data	Ways that the data in this dataset is used.
Availability	The availability of this dataset.
Data Formats	The formats in which the dataset is available.
Policies	The policies associated with this dataset for release, linking, or other policy.
Publications	Publications about this dataset.
Suggested Citation	A suggested format for citations in publications referencing this dataset.
References	References associated with the description of this dataset.

7.6 Develop Data Dictionaries

The data dictionary sits in the middle zone. It is intended to be a human-readable artifact that describes the data contained within the dataset. It should also be readily consumed by an information system and easily support automation and validation activities. Data dictionaries are typically in tabular format, where each row of the table describes the contents of a row in a table, or the attributes of an object in the information system that stores it.

An example data dictionary is provided below. A detailed data dictionary appears in the Data Dictionary.xlsx file accompanying this document.

Field Name	Data Type	Size	Min	Max	Conf	Value Set	Constraint
Field Name	String	16	1	1	N		Table + Field Name must be unique.
Data Type	String	16	1	1	N	Data Types	
Field Size	Non-Negative Number	1	0	1	N		
Min	Non-Negative Number	1	0	1	N		
Max	Non-Negative Number	1	0	1	N		
Confidentiality	String	1	1	1	N	Confidentialities	
Units	String	32	0	1	N		
Value Set	String	32	0	1	N	Value Set Names	
Constraints	String	80	0	*	N		
Definition	String	4096	1	1	N		
Description	String	4096	1	1	N		
Rationale	String	4096	1	1	N		

From a best practices perspective, a data dictionary will provide a concise name, a descriptive name, a detailed description of the data, the rationale for its collection, and any other relevant constraints. These constraints include the number of times the data element can appear (its cardinality or repeatability), the type of data item it stores, the maximum length of the data, the range of legal values, or when coded data is used, the set of valid values which can appear (ideally with the meaning of each code), and other constraints or business rules which may apply.

7.6.1 Constraints

To ensure clarity and usability, constraints should be expressed both in human-readable and machine-interpretable formats. HL7 and other standards use the terms "SHALL", "SHOULD", and "MUST" in human-readable descriptions when defining constraints in data dictionaries.

It is strongly recommended to use restricted constraint languages for human readers, appropriate to the implementation:

1. For constraints on data definitions in systems using HL7 V2 messaging, use the conformance language defined for the HL7 V2 project [described here](#).
2. Consider using the conformance language [described here](#) to constrain data definitions in systems exchanging HL7 Clinical Document Architecture (CDA) documents.



3. Consider using the conformance language [described here](#) when no other conformance language is defined.

Computer-readable constraints can appear in any programming language. Common approaches in HL7 specifications utilize FHIRPath for constraints on the HL7 FHIR and the most recent CDA and Consolidated Clinical Document Architecture (C-CDA) specifications, which leverage FHIR tooling. FHIRPath can also be applied to other information models. Still, applying it to them without sufficient skills in using the FHIRPath language and its implementation in readily available open-source technologies is not recommended.

Schematron is strongly recommended as the computer-readable constraint language in systems using XML-based information models.

7.6.2 Constraints Across Fields

Some constraints (called co-occurrence constraints) can only be evaluated across multiple data elements in a data dictionary. An example of a co-occurrence constraint is found in the CDC HL7 V2.5.1 Immunization Guide, which examines four different fields within an OBX segment. It says:

IZ-35: If OBX-3.1 is “64994-7” and OBX-2 is “CE” then the value set for OBX-5 shall be HL70064

Another example can be found in the HL7 FHIR Patient resource. It says that the Patient.Contact SHALL at least contain the contact's details or a reference to an agency resource. Expressed in code, this is represented in an expression as `name.exists()` or `telecom.exists()` or `address.exists()` or `agency.exists()`, which means that four different fields in the contact must be examined: name, telecom, address, or agency.

7.7 Extract Data Schemas and Models from Existing Data Infrastructure

Data schemas or data models can often be extracted from existing IT infrastructure within a database. All SQL-based databases provide technology that allows the extraction of the SQL database schema. Other database technologies offer a means to extract the metadata that describes the database. These artifacts can be readily collected from existing infrastructures and stored in the metadata repository.

7.8 Gather Implementation Guide Resources

Many public health information systems provide inbound or outbound interfaces that describe the format and protocols used to exchange data with public health agencies. A list of nationally recognized and alternative implementation guides for public health reporting can be found on the Assistant Secretary for Technology Policy's (ASTP) [Interoperability Standards Advisory \(ISA\)](#) Website. Many public health agencies publish localized implementation guides that incorporate additional constraints into the national guides, addressing local constraints or specific vocabulary. These guides and the technical artifacts used to create them should be collected and stored in the metadata repository, linked to the datasets and information systems that implement them.

7.9 Transformation Scripts

Transformation scripts are automated processes or code written to convert or reformat data from one format to another. These scripts are essential for ensuring that data conforms to the proper structure,

format, and standards as required by various public health systems and stakeholders. Transformation scripts are commonly used in scenarios such as:

- Mapping data between different data models or formats (e.g., from legacy systems to modern metadata frameworks).
- Standardizing data to ensure consistency across jurisdictions or systems.
- Cleaning data by applying transformation rules that ensure the data meets required constraints (e.g., removing invalid entries, correcting formatting errors).

These scripts are crucial for metadata management because they enable public health agencies to process raw data from various sources (e.g., hospitals, public health departments) into standardized formats that can be easily used and shared.

Examples of Transformation Script Use Cases:

- Converting HL7 messages into FHIR-compliant formats.
- Mapping data from one system's schema (e.g., legacy SQL database) into the structure of a new metadata repository.

Transformation scripts may be developed using standard scripting languages such as Python, SQL, or specialized tools like Apache NiFi or Talend, depending on the complexity and scale of the task. These scripts are stored as part of the metadata repository, ensuring they are version-controlled and updated as necessary.

Key Considerations:

- Ensure that metadata transformations are carefully documented to avoid misinterpretation of data and ensure traceability of data lineage.
- Implement a change control management system to manage and document updates to transformation scripts.
- Utilize validation scripts to ensure transformed data meets the specified rules and standards.

7.10 Link Policy to Metadata Artifacts

Datasets and other artifacts should contain metadata that links them to applicable policy.

7.11 Validation Scripts

Validation can often be automated through systems that process data dictionaries, models, or schemas to generate validation scripts. These scripts are typically written in custom languages that enable validation of data flowing through information systems and their interfaces. Data validation scripts are specialized transformation scripts that produce a validation report, measuring the validity of the incoming data.

The technical artifacts used to create implementation guides or accompany them often include validation tools or scripts, or can be used to make them. Several open-source tools support the development of validation scripts. Some of these are described in the table below.

Standard	Validation Script Development Tools
----------	-------------------------------------

HL7 Version 2	The NIST Implementation Guide Authoring and Management Tool (IGAMT) site includes tools for developing implementation guides, which can also be used to develop test cases and validation scripts for HL7 Version 2 messaging.
HL7 FHIR	Implementation guides supporting HL7 FHIR can be validated using HL7 and open-source tools from the implementation guide content (see Validation - FHIR v4.0.1) and through the MITRE-developed Inferno Framework and Test Kits .
Other Standards	The ASTP's Standards and Interoperability Testing Environment (SITE) website provides validation tools and scripts for public health reporting guides and for other interoperability standards .

7.12 Quality Measurement Scripts

Extending beyond data validation, quality measurement scripts are yet another form of transformation that measures the quality of data, rather than its validity. The distinction between validation and quality measurement is that validation identifies insufficient data. Still, it does not assess the distinction between good, better, and best data based on some defined quality standard. That standard may be based on internal quality, whether all fields are populated, and whether the data contains essential information necessary to perform other tasks or assessments. Still, it can also be based on the quality of care or expected outcomes.

Appendix A – Glossary of Terms

Metadata

Often described as “data about data,” metadata refers to the information about data that is not typically present within the data itself. It provides essential details about the data's context, including the **Who**, **What**, **Where**, **Why**, **How**, and **When** related to the data.

Examples of metadata include:

- **Who** sends it, who manages it, who authorizes access to it
- **What** processes apply to it and what policies apply to it
- **Where** is it stored? Where did it come from?
- **Why** it is used
- **How** it may be sent, how values are computed
- **When** it might be sent (e.g., the triggers for sending it)

Metadata Artifacts

- **Data Catalog:** A list of datasets, descriptions, and relevant information to enable their use.

- **Data Dictionary:** A description of the data and the relationships between data elements within a single data set. It is generally intended for human readers but may also be structured for computer use. Data dictionaries are most often found in tabular formats or spreadsheets.
- **Data Element:** The smallest piece of data that can be exchanged or to which a policy or procedure can be applied.
- **Data Element Definition:** A description of the smallest unit of information in a data dictionary, such as a column in a table or an element in a schema.
- **Data Link:** A data element that links within or between datasets (e.g., a unique individual identifier).
- **Data Schema or Data Model:** A structured artifact primarily intended to communicate the data structure for automated systems to use, but also informative to developers or managers of those automated systems. Data schemas or models are often found in structured formats using JSON, SQL, XML, or other technical syntax, or may also be depicted graphically using specialized drawing tools.
- **Data Set:** A collection of organized data, possibly time-bound (e.g., 2024 wastewater reports).

Roles

These roles are often shared across teams and may include sub-roles such as product owners or source technical subject matter experts. Not all jurisdictions will have staff specifically dedicated to the roles provided below, the assignment to that role may arise from other responsibilities within the agency. For example, subject matter experts may take on a data stewardship role for data within their area of expertise.

- **Data Custodian:** Manages technical control and access to the data, including its storage, security, and backups.
- **Data Owner:** Holds accountability for a dataset and has the authority to make decisions regarding how the data is collected, shared, and used.
- **Data Steward:** Oversees the quality and integrity of data and metadata over time, ensuring compliance with governance standards.
- **Data User:** Accesses and utilizes data for analysis, reporting, research, or decision-making.

Privacy

These definitions are general and may not align precisely with those provided in HIPAA or 42 CFR Part 2.

- **Deidentified Data:** Deidentified data is data that cannot by itself be used to re-identify an individual or entity (e.g., organization). It originates from identifiable data (e.g., data containing PHI or PII) that has been processed to remove identifying information. That process may be irreversible, in which case the result can be described as Anonymized Data (see below), or it can be reversed



- **Re-identification Data:** Data that can be used to re-identify an individual or entity in previously pseudonymized (de-identified) datasets.
- **Personally Identifiable Information (PII):** A collection of data that can be used to identify an individual directly.
- **Protected Health Information (PHI):** A PII collection that includes health data.
- **Safe Harbor:** A HIPAA standard requiring the removal of specific identifiers for data to be considered de-identified under federal law.

Policy Implementation

- **Policy Classification:** A set of codes or tags applied to a dataset or data element to identify applicable policies (e.g., PHI, PII, 42 CFR Part 2).
- **Sensitivity:** The degree of importance or risk associated with a dataset or data element, often categorized into levels such as Low, Moderate, or High, to indicate how critical or restricted the data is in terms of handling and access.
- **Conditionally Sensitive:** A marker indicating sensitivity is based on content (e.g., HIV medications vs. other medications).

Additional Definitions

- **Policies and Procedures:** Policies and procedures establish the agency's processes to govern data and metadata. Policies outline the necessary actions (e.g., the processes that must exist and be performed), while procedures detail the specific steps for executing those policies. (DMBOK: 3:1.3.6)
- **Business Glossary:** A business glossary describes the meaning of key terms used within the agency.
- **Data Catalog:** A data catalog describes the data that an agency holds. A data catalog comprises written records that describe the agency's data. This could be as simple as a collection of documents describing each dataset or an application that maintains structured records for each dataset.
- **Technical Metadata:** Technical metadata describes the structure, format, and processing rules of data, including data dictionaries, schemas, and data models, essential for managing and operating on data within systems.
- **Technology Infrastructure:** Technology infrastructure refers to the hardware, software, and systems used to manage, store, and process metadata and data within an agency including data management platforms, enterprise data catalogs, and document control systems.