ITRC has developed a series of fact sheets that summarizes the latest science, engineering, and technologies regarding environmental data management (EDM) best practices. This fact sheet describes:

- data management principles and concepts applicable to environmental data
- documenting data for usability and discovery

Additional information related to data storage, documentation, and discovery is provided in the ITRC fact sheets on Data Management Planning; Data Governance; Data Lifecycle; Data Access, Sharing, and Security; and Data Disaster Recovery.

1 INTRODUCTION

Data storage and documentation are important components of any data management strategy. Organizations that manage environmental data should incorporate the elements in this document and data disaster recovery protocols into their data management framework.

2 DATA DICTIONARIES

Data dictionaries provide a description of data in business or operational terms useful to the organization maintaining the data, including information about the data types, data structure, definitions, and security restrictions. Data dictionaries help ensure usability and compatibility of data sets within an organization. Metadata creation is simplified if data dictionaries are in place because the data dictionary definitions can be used in the metadata.

The United States Geological Survey (USGS) Data Dictionaries webpage indicates that data dictionary contents can vary but typically include some or all of the following:

- a listing of data objects (names and definitions)
- detailed properties of data elements (data type, size, nullability, optionality, indexes)
- entity-relationship (ER) and other system-level diagrams
- reference data (classification and descriptive valid values)
- missing data and quality-indicator codes
- business rules, such as for verification of a schema or data quality

3 MAIN AND REFERENCE DATA MANAGEMENT

An organization that uses the data sets repeatedly will benefit from centrally managing the shared and repeatedly used data. Main data is data shared across an organization such as activity locations, contact/customer data, water levels, and sampling results. Reference data are shared or common attributes that change infrequently, but are often used, such as proper names of places, Chemical Abstracts Service Registry Numbers (CASRN), or units of measure that have well-defined valid values. The valid values adopted for use by an organization or project should also be constrained by adding them to domains within the data management tools used by the organization or project (see also Valid Values Fact Sheet).

Data stewards are ultimately responsible for maintaining main and reference data within their designated data sets. It is a good practice to create a core team of data stewards, information technology staff, and/or users who have a stake in keeping the data usable to maintain the main and reference data for the organization. The team structure encourages the communication needed so changes are effectively identified, implemented, and communicated to the affected users.

3.1 Persistent Identifiers

Persistent identifiers (which may also be key fields) uniquely identify a main or reference data attribute value (for example, a monitoring point name or system-generated internal identifier). Persistent identifiers improve the efficiency of data usage by enabling consistent and reliable linking and comparison of information across an organization. Data values that need to be identified across data sets should be given a persistent identifier (for example, locations, facilities, methods) that includes all main and reference data, as well as some project-specific data. A best practice is to use system-generated persistent identifiers to ensure the uniqueness of each identifier; however, persistent identifiers may be valid values representing real-world terms (for example, using “MG” or “mg” to represent “milligrams” as a unit of measure) and are usually short or

system-defined unique codes to conserve data system storage space and simplify using them as linking values. In this example, as long as the value “MG” or “mg” always represents “milligrams” throughout the data storage system, then it may be used as a persistent identifier.

3.2 Interoperability/Integration

A data set that can be used in more than one system or process is more valuable than single-use data sets. Interoperability can be improved by using existing schemas such as the Open Geospatial Consortium Standards or the United States Environmental Protection Agency (USEPA) Data Standards), or by following these principles when designing a data schema:

- use main and reference data where possible
- use valid values (the set of possible values for an attribute) to control the allowable valid values
- use an existing data dictionary or create one to define attributes to be used across the data set or organization
- use industry-standard definitions and attribute definitions

3.3 Procedures

Data management procedures should be defined and documented to encourage consistency, repeatability, and defensibility of the data. Processes, documents, and tools may vary across the data lifecycle as discussed in the following sections.

3.3.1 Acquisition and Collection (see **Field Data Collection** fact sheets)

No matter how data are acquired for inclusion in a data set or project, the methods and processes used to incorporate the data should be documented, including:

- agreements (for example, access agreements, consent/acceptable use agreements, data sharing agreements)
- field procedures and collection standards (see Defining Data Categories and Collection Methods Fact Sheet)
- source data documentation and attribution
- steps taken to transform and incorporate the data
- quality assurance procedures

3.3.2 Data Modification and Audit Tracking

Data in use are rarely static and standards should be developed to govern how data modifications are made and tracked so changes to the data can be monitored and recorded. A best practice is to track data modifications within the data set using audit tracking data elements and document the processing steps in the metadata associated with the data set. Audit tracking is record-level tracking of data changes and should be enabled if possible. Audit tracking elements such as dates and usernames should be system-generated, if possible, to reduce data entry time and ensure that the audit data will be populated. Table 1 lists examples of basic, record-level audit data elements.

Table 1. Examples of record-level audit data elements

Data Element Name	**Information Stored**
Created Date	Calendar date/time when the record was created
Created User	Username of the person or system that created the record
Last Modified Date	Calendar date/time when the record was last modified
Last Modified User	Username of the person or system that modified the record
Modification Description	Standardized short description of modification such as “Location corrected,” “Detection Limit updated,” “Migrated to new schema”

3.3.3 Archiving and Retention

After a data set is no longer needed for its intended purpose, it may be archived or deleted. It’s recommended to archive rather than delete data sets because it’s difficult to know if a data set may be useful in the future either by the organization

that owns the data or by secondary data users. A best practice is to develop organization-, project-, or client-level record retention policies that address the regulatory or legal retention requirements, the project/program lifespan, stakeholder requirements, and the usability of the data. For example, data that has been transformed into another data set could be deleted if the data lineage is adequately attributed and all usable elements of the initial data set are still available. Retention policies should be periodically reviewed and updated to ensure that all applicable data sets are covered by the policy. The organization should also document when data sets are deleted or no longer available to stakeholders (for example, a data set is no longer publicly accessible, but available to internal stakeholders).

3.4 Stewardship

Data stewardship is the accountability and responsibility for data and processes that ensure effective control and use of data assets and is enabled by defining data stewards (see Data Access, Sharing, and Security Fact Sheet) within the organization. Data stewards may be responsible for multiple data sets but defining stewards for all data sets is important for maintaining data usability.

4 DATA DOCUMENTATION AND DISCOVERY

A data set's utility depends on the ability of end users to understand what the data represents and to find the data needed. Proper documentation and tools empower data discoverability at all user levels, including within and outside the data's host organization. Data that is available but unknown to potential data users has little value to an organization. The following sections will describe the components recommended for documenting data and enabling its discovery.

4.1 Data Lineage

Data lineage documents the pathway of the data and any alterations from the original source to its current form. Lineage documentation can include descriptions of any or all of the following:

- source files
- exchange, transformation, and load (ETL) processes
- change and modification logs
- transformations of the data for analysis and calculations done to create derived data

Documented data lineage provides transparency, compliance with regulatory and other requirements, and reduces the risk that the data are not suitable for the intended use. Formatted metadata is the preferred method for storing attribution, but documents describing the data set and its history may also be used.

4.2 Data Catalogs

Data catalogs are inventories that inform users about available data sets and associated metadata. A well-designed catalog will also provide users with easy access to the data itself or provide a mechanism/process for accessing the data. Ideally, data catalogs should be maintained and published at an organizational level or higher to enable data discovery by as many potential users as possible. Data catalogs should capture the following data elements:

- data set title
- data steward(s) or contact(s)
- description of the contents
    - purpose
    - timeframe of relevance (for example, what period of time does the data represent). The end date may be left blank if the data are still current.
    - description of the contents such as "Ground water sample results collected for remediation/closure of site x."
    - description of data elements stored in the data set
- link or instructions on how to access the data
- acceptable use restrictions

- status of the data set (for example, active, archived, deleted, deprecated, etc.)
- data source
- data quality/usability

Data catalogs use many of the same elements that are used in metadata, so it is possible to use a metadata management system as a data catalog. It's also useful to store information about shareable products derived from the data, such as reports and exports, in a catalog to enable users to find and access the derivative products.

4.3 Metadata

Metadata is information that defines and describes the characteristics of other data and should be implemented to capture the business definition of the data and the technical features of the data. Metadata provides background information about the data and can make data discoverable when searching digital resources for information. Metadata should provide enough detailed information to facilitate data handling and should include any data use limitations. Updated and complete metadata are critical to maintaining data quality, and metadata documentation must be updated to reflect actions taken upon the data throughout the data lifecycle process (for example, data lineage).

4.3.1 Why Is Metadata Important?

In short, metadata provide data users with the business-related context (for example, what is the data used for) and technical features (for example, data element definitions, lineage, attribution, and intended use) needed to understand what the data represent.

Metadata are crucial for any use or reuse of data; no one can responsibly reuse or interpret data without metadata that explain how the data were created, why they were created, where they are geographically located, and details about the structure of the data (USGS 2021).

4.3.2 Who Manages Metadata?

Data stewards are responsible for creating, managing, updating and enhancing the metadata through the data lifecycle and should select the appropriate metadata standard and the detail levels needed for each data set. Some organizations may have standards for metadata schemas and required fields which are also implemented by the Data Steward. Metadata updates may also be completed by other data managers during the data lifecycle especially where data modifications are documented as part of the data lineage metadata.

4.3.3 Principles for Implementing Metadata

Some important principles to help with implementation of the metadata are as follows:

- Metadata should describe the data set's purpose, content, history, limitations, and structure.
- Any changes in the metadata content should be tracked and dated.
- Metadata should be updated as the data set moves through its data lifecycle.
- The metadata schema should conform to community standards and is appropriate to material in the collection, the users, and current or future uses. The schema should support interoperability.
- Metadata should provide context and assumptions regarding the data set, describe sharing between systems, and inform mapping to help conduct metadata searches.
- More than one metadata schema can be applied, such as Federal Geographic Data Committee (FGDC) for geographic data with another standard used for nongeographic data (see FGDC Geospatial Metadata Standards and Guidelines). The choice of metadata schema selected may depend on cost, expertise level, expected use, users of the data collection, goals set for sharing and interoperability, and the level of additional details of metadata needed at collection, group, and item levels.

- Metadata should list the conditions and terms of use for the objects, such as copyrights, license, publication limitations, and citation and attribution requirements, as well as comments on published status, so data users are able to find and use the associated data.
- Because metadata can also be considered as a data set, documentation for the metadata (that is, metadata for metadata) should also be tracked to document the metadata's provenance, integrity, and authority. Here

metadata documents the standard of completeness and quality, such as metadata update frequency.

4.3.4 Metadata Organizational Models

Because metadata support the management, use, and preservation of data collected for many different purposes, there are multiple models describing how metadata can be structured. Choosing a metadata model, or developing a custom model, depends on how the organization is structured and what metadata elements are most useful to the organization.http://framework.niso.org/24.html Each data category may have a different set of standards to follow depending on the metadata elements deemed most important by the organization (for example, GIS data have different requirements than field data and may follow the FGDC or International Standards Organization (ISO) metadata models [see Geospatial Metadata subtopic sheet], and non-GIS data may use a different model). The National Information Standards Organization (NISO) model classifies metadata based on the data's purpose (see NISO Metadata) and is well suited for data collected to support long-term studies and sharing because it provides detailed information about the history and lineage of the data, whereas the DAMA model (DAMA 2017) is good for documenting data for short-lived projects where data retention is not an important consideration.

4.3.5 Metadata Elements

Metadata elements are the data objects used to characterize the data set. ISO 15836 provides an example of core metadata elements that describe resources across domains. Table 2 provides an example of common metadata elements and associated descriptions.

Table 2. Example metadata elements
Source: Adapted from District of Columbia Metadata Submission Guide (Open Data DC 2021).

Element	Description
1.1	Item Description—title, tags, description summary, credits, limits
1.2	Topics & Keywords—topics, place, theme
1.3	Citation—created, published, revised
2.1	Resource Details—Status, supplemental information
2.2	Resource Extent—time period, current reference,
2.3	Resource Contacts—Contact, organization, position, role, address, email
2.4	Resource Maintenance—frequency, next update
2.5	Resource Constraints—legal, security
2.6	Spatial Reference
2.7	Resource Quality—Accurate, consistent, complete, spatial references
2.8	Resource Lineage—source media, citation, title, origin, publication, and other process
2.9	Distribution—contact info
2.10	Entity Attribute (Data Dictionary)
3.1	Metadata Reference

4.3.6 Metadata Creation

The USGS Metadata Questionnaire and Metadata in Plain Language documents can help in getting the metadata creation process started (USGS 2021).

Metadata creation tools typically use data dictionaries to store and communicate metadata information in a database, a system, or data (used by applications) and create metadata records. An example of environmental metadata showing identification, entity and attributes, and distribution elements can be found at the following New Jersey GIS link: Ambient Stream Quality Monitoring Sites (1998—2010).

4.3.7 Metadata and the Data Lifecycle

While it is important to review and update all metadata elements as a data set changes, some metadata elements become more important as data moves through the data lifecycle. Table 3 lists the metadata elements that are introduced in or that increase in importance based on the lifecycle phase.

Table 3. Metadata elements introduced by data lifecycle phase

Lifecycle Stage	Acquire: Information about the Source Data	Process/Maintain: Information about the Data	Share: Information Needed to Understand and Discover the Data
Metadata Elements	• Listing of source data objects (names and definitions) • Detailed properties of source data elements (data type, size, nullability, optionality, indexes) • Source entity-relationship (ER) and other system-level diagrams • Source spatial reference frame • Steps taken to transform or incorporate the source data	• Entity and attribute information—details about the information content of the data set, including the entity types, their attributes, and the valid values to which attribute values may be assigned • Spatial reference frame used • Reference data (classification and descriptive valid values) • Data organization and structure	• Distribution information—information about the distributor of and options for obtaining the data set • Metadata reference information—information on the correctness of the metadata information, and the responsible party • Citation and attribution standards for referencing or reusing the data

4.3.8 Validating Metadata Records

Metadata should be validated periodically to ensure that the metadata have been created properly and all required elements have been filled in. A best practice is to update and validate metadata whenever a data set is modified or moves to a different phase of the data lifecycle (for example, from Process/Maintain to Share).

5 REFERENCES AND ACRONYMS

The references cited in this fact sheet, and the other ITRC EDM Best Practices fact sheets, are included in one combined list that is available on the ITRC web site. The combined acronyms list is also available on the ITRC web site.