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:

- considerations when using electronic data deliverable (EDD) formats for data exchange
- transformations that may be required for transmittal of EDDs, and documentation thereof

Additional information related to data exchange is provided in the ITRC fact sheet on Valid Values and in Historical Data Migration Case Study: Filling Minnesota's Superfund Groundwater Data Accessibility Gap, and USGS Challenges with Secondary Use of Multisource Water Quality Monitoring Data Case Study.

1 INTRODUCTION

When addressing the exchange of data between source and target environmental data management systems (EDMS), a very powerful tool and medium for that exchange is the electronic data deliverable (EDD). EDDs serve as a method to collect, transform, and communicate information between parties and their EDMSs. EDDs may transmit new data into or between EDMSs, or they may be used to update or add information to what is already in the EDMS (such as adding data validation information to existing analytical data). A recent International Conference on Environmental Data Management (ICEDM) white paper defined EDDs as, among other things, "… a structured data file that places a particular data item in a particular column/order" (ICEDM 2017). More than just lab results, EDDs reveal the choices parties make on how data must be represented (Valid Values Fact Sheet) and what and how data must be preserved and communicated (Data Migration Best Practices Fact Sheet).

2 ELECTRONIC DATA DELIVERABLE FORMATS

EDDs contain data within a specific format. EDDs themselves can also take multiple forms—they may, for example, be a comma-separated values (.csv) file, a collection of distinct but referentially connected tables, or a raw text file. An important aspect to the success of EDDs as a tool for the exchange of information is the clear definition of what information each column in a record is supposed to communicate. This is critical for terms that might mean different things to different people who use different EDMSs. Clear definitions for the columns of records, as well as the metadata of valid values, communicate to EDD sources and targets what each field represents, how values are determined, and what vocabulary is implicated in the data.

An EDD's producer and recipient can also have distinct rules or constraints for EDD formats. EDD formats can have rule sets applied to them with rules that can be enforced at multiple stages and triggered by particular conditions. Some examples of types of enforcement may include rules at the level of the structure or schema of the format, rules set as macros, or rule sets enforced by vendor applications.

When using an EDD as the medium for exchanging data, the recipient will need to be familiar with the rules established for the EDD by the party that published the source format, and how those rules relate to the data structure and content of the producer and the recipient. Ideally the EDD's rules were crafted with FAIR principles in mind (Findable, Accessible, Interoperable, and Reusable), which are described in greater detail in the Public Communication and Stakeholder Engagement White Paper. Understanding the constraints enforced by the EDD format, and how those do or do not match the data systems of the source or target EDMS, is crucial to preventing the loss or misrepresentation of data being exchanged. Consequently, the format of an EDD produced by either party may present challenges for communicating information. Navigating differences in EDD formats can impact the effort to exchange data between parties that have made different choices regarding how they expect their data to be preserved and communicated. The burden of acquiring this understanding may fall most heavily on the recipient of the EDD. Additional communication between the producer and recipient can be important to ensure an accurate understanding.

Examples of EDD Media

- text file (.txt)
- comma-separated value (.csv) file
- spreadsheet
- proprietary file type

Examples of Exchange Methods

- application programming interface (API)
- web service
- email
- proprietary software

Differences in the use and interpretation of EDD data by the source and target may result from:

- differences in the structures or constraints between data systems. For example, one system may require that sample identifiers be unique only within each event, whereas another requires sample identifiers to be unique across all events.
- differences in the level of detailed information stored in each system. Examples include: 1) one system stores attributes that the other does not, such as penetration depth for sediment samplers; and 2) one system contains all analytical results from all control samples, whereas another system contains only one result for each sample, with control samples omitted or averaged.
- differences in the level of detail for individual data types. Examples include: 1) one system includes time as part of the sampling date, whereas another system does not; 2) one system requires a lower sample depth if there is an upper depth, and another system does not; and 3) one system records the number of significant digits for each measurement, and another system does not.
- conceptual differences in the meaning of terms. For example, “sample type” may be used to distinguish original samples and split samples in one system, natural and quality control samples in another system, and different sample mediums in another system.
- differences in the values reported for undetected chemical measurements, specifically method detection limits or quantitation limits.
- differences in the scope of valid values between systems. For example, one system may use different analyte codes for dissolved and total metals, whereas another system uses two different sets of valid values, one for the analyte and one for the fraction.
- differences in the allowable maximum length of identifiers or differences in data types. For example, one system may require a column to be an integer while another system requires a similarly defined column to be text.

The documentation for an EDD format should address as many of these types of issues as possible. That documentation may take the form of tabular information that defines and describes the tables and columns of the EDD format, annotations in an EDD template or an EDD product, and/or narrative descriptions of requirements and constraints. Detailed documentation may be available for widely used EDD formats but be minimal or lacking for custom or ad hoc EDD formats. In the latter cases, multiple interactions between the source and target may be needed to clarify how the EDD should be, or has been, used. Whenever a custom EDD is to be used for repeated transfer of data between two parties, joint development of the EDD format and documentation is recommended to minimize data representation issues.

Even detailed EDD specifications cannot address all differences that may arise in actual usage. For example, the recipient of an EDD may note that there are apparently no results for control samples, but may not know whether this is because

- no control samples were collected
- data from only one of each set of control samples is included, either deliberately or because of an error in EDD preparation
- data from control samples were averaged, and the average is reported
- the EDD structure, and the sample identifiers used by the data preparer, do not allow control samples to be easily identified

3 DATA TRANSFORMATIONS

Preparing to share data from the source to the target means extracting the data from a source EDMS into an EDD that conforms with the data needs of the target EDMS. EDD sources and targets may have different means of recording information. One party may use text or commentary to narrate information that’s collected by another party as a choice from a set of valid values. One party may use a character string where another uses a numeric value set. One party may preserve distinct objects for values that the other preserves as attributes of another object. When these differences arise, if information is to be exchanged effectively between parties, the data elements must be mapped from where and how they are stored in the source EDMS to the target. Once the fields and valid values have been mapped between the two systems, they can be transformed *or* remapped to create EDD(s) for the target EDMS that contain the data from the source. Sometimes, manual changes with visual inspection may be required. Other times, it may be sufficient to use formulas and software tools on either end to transform the data with less direct human input. Yet other times, revisiting the metadata choices of how the data are represented (choices on how to represent compounds, units of measurement, etc.) may prove worthwhile as a means to ensure both parties agree on what is being reported.

On the level of valid values choices, two parties may differ regarding how to represent a compound (see USGS Challenges with Secondary Use of Multisource Water Quality Monitoring Data Case Study). These differences may be crucial for accurate reporting or may be easier to reconcile. The ICEDM white paper addresses the impact of consequential vs. nonconsequential valid values and possible steps that can be taken with an EDMS to address those consequences (ICEDM 2017). Some decisions about valid values may be easier than others—choices for how a company should be represented in a data set don't have the same impact as how a method, a chemical, or a sample material might be represented.

Different types of EDD formats communicate different values and decisions. Some may aim for simplicity in communicating information, and others for complexity, so it is crucial where possible that the source (that is, the data generator, which could be a laboratory, consultant, responsible party) and target (that is, the data recipient, such as a regulatory agency) strive for an agreement—a balance that addresses the interests of both preserving the work at the source and serving the requirements of the target. Some information may prove more or less consequential. Some data may prove to be unusable in the form in which they were received and require transformation (see Historical Data Migration Case Study: Filling Minnesota's Superfund Groundwater Data Accessibility Gap). Sometimes metadata may prove to be worth distilling into common data elements (see USGS Challenges with Secondary Use of Multisource Water Quality Monitoring Data Case Study). There may be no one answer as to how and to what degree data must be transformed in cases where source and destination differ on the matter of how much complexity is necessary in describing an event for reporting purposes. But in preparing an exchange of data, it is crucial to address the question of how complex and specific does the EDD need to be to communicate and document the required information.

Both simple and complex data transformations may be necessary to transfer data from a database to an EDD or vice versa. For consistency, reliability, and efficiency, these operations should be automated as much as possible. This automation may be accomplished with spreadsheet macros and formulas, structured query language (SQL) scripts, or programs written in languages such as Python or R. Comments in scripts and programs are another means by which the need and rationale for data transformations can be documented. Scripts and programs can also include quality assurance checks of the data and can produce documentation of all the data transformation changes that are made.

An EDD that is transmitted from one party to another should ideally be accompanied by metadata about the EDD itself. This is particularly important when there is an iterative cycle of corrections and resubmissions of the same data, and when the file names used for the EDD files are fixed. A non-exhaustive list of critical EDD metadata may include:

- a data set identifier that uniquely identifies the data that are included
- a narrative description of the data that are included
- a version number for the submittal, to distinguish successive submittals of the same data
- a version number or other identifier for the EDD format, if available
- the date at which the data set was last updated prior to preparation of the EDD
- the date of EDD preparation
- contact information for the person and organization that prepared the EDD
- the data source, as applicable

Negotiating the differences and gaps that arise from how different parties choose to preserve data can be expected to involve marshaling available resources. To negotiate the decisions made by multiple parties regarding how to preserve and communicate information, the primary resources would be information presented by those parties on the structure of their databases and their guidance on how to submit or report data. In addition, there's a wealth of published information on EDD formats, guidance, and decisions made by states and industries regarding how they want to receive information (see the Data Exchange and Valid Values Resource List). Reviewing the decisions made by other parties regarding which information should be prioritized and how the information may be preserved can bring clarity to the needs of the parties exchanging information.

4 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.