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 methodologies for:

- extracting data from commonly encountered formats
- transforming data for loading into a new EDMS, often a relational database

Additional information related to data migration best practices is provided in the Historical Data Migration Case Study: Filling Minnesota's Superfund Groundwater Data Accessibility Gap and the fact sheet Defining Data Categories and Collection Methods.

1 INTRODUCTION

Data migration is a subset of data exchange where the source and target do not have a pre-existing data export/import process.

The two top reasons for conducting data migrations are:

- moving data from legacy sources or systems into a new or different system
- transferring data from an existing data owner's system to a new data owner's system

This fact sheet presents best practices and considerations for:

- planning for a data migration
- manual vs. automated migration
- migration quality control

2 PLANNING FOR A DATA MIGRATION

Developing a well-documented approach for each step of a data migration is essential. The long-term benefits to data quality are substantial.

2.1 Determine Migration Objectives

Before beginning the migration, determine the objectives and milestones. Identifying clear objectives and milestones for the migration provides a tangible benefit to the time and money invested, even if the migration is ultimately suspended due to budget constraints. There are numerous reasons why data migration may or may not proceed. Some important questions to consider include:

- Will the data be used for a specific purpose, such as modeling or risk assessment?
- Do you want or need all legacy records in a single system?
- Are you being asked to submit legacy data to an environmental data management system (EDMS)?
- What are your data quality objectives?
- Can you accommodate missing metadata?

2.2 Conduct an Audit

A data audit involves looking at key metrics to create conclusions about the characteristics and properties of a data set. Specifically, an environmental data audit is an assessment of existing project data to verify quality and completeness. Conducting an audit of the source data to be migrated is essential. The International Conference on Environmental Data Management (ICEDM) developed a white paper, Environmental Data Quality Audit: Foundation and Framework (http://www.icedm.net/s/ICEDM-Historical-Data-Migration-Audit-Final.pdf), that provides guidance on this process.

2.3 Identify Common Data Sources Based on the Audit

Examples of data sources from which data might be migrated include:

- spreadsheets
- relational databases

- document databases (for example, JavaScript Object Notation [JSON], XML, or NoSQL like MongoDB)
- collections of files that may span several additional formats, including PDF, paper copies, or delimited text files

While building an inventory of source data to migrate based on the results of the audit, identify and group data sources with similar formats or structures. Grouping similar formats helps when developing a strategy for mapping source data to the target system. This is especially important when considering automated data migration.

2.3.1 Data Structure

For data migrations, consider the overall data structure. The three categories of data structure are structured, semi-structured, and unstructured. A general relationship between these categories is illustrated in Figure 1 and described below.

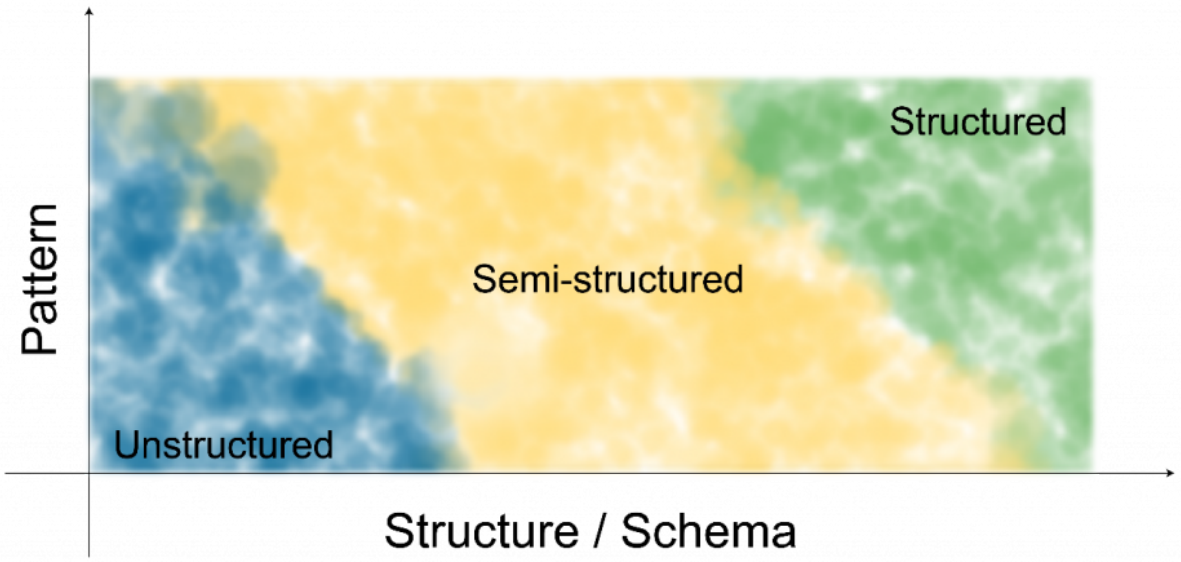

Figure 1. Spectrum of data structure

Structured data have a predefined structure or schema, which results in them being consistently organized and parsable. Once the individual elements of the data set are identified, data users can easily filter, search, and visualize the information. Examples of structured data include:

- relational databases
- document databases (for example, JSON, XML) in which data are encoded by a set of rules
- spreadsheets or delimited text files that follow consistent procedures for data organization

Semi-structured data don't conform to a rigid, predefined structure, but are generally parsable. For a given set of data and patterns, a semi-structured data source can have many possible structures. Although it is preferable to work with structured data, semi-structured formats may be more easily attained than structured data formats. Examples of semi-structured data include:

- Data maintained in a spreadsheet program or delimited text files where there is an inconsistent structure if a column is missing in one spreadsheet but present in another.
- An API response in JSON without adherence to specific protocols for maintaining consistency. There is an inconsistent structure if it returns an error message or multiple file types, or if the JSON files have optional attributes.
- Text files with consistent patterns (for example, some searchable PDFs).

Unstructured data are neither parsable nor consistently organized, which makes them the most difficult to exchange. Because they contain few consistent patterns, there are few or no ways to reliably access the data. Typically, preparation of data with these forms must be done manually. As stated in the Defining Data Categories and Collection Methods Fact Sheet, some unstructured data can be transcribed in various ways into structured data. However, expert interpretation is often

needed to find patterns and structure. Examples of unstructured data include:

- reports
- field notes
- images such as scanned PDFs or photos
- videos
- data with inconsistent formats

2.4 Develop a Strategy

Although there are many benefits and long-term cost-savings associated with the migration of data, the process can be expensive, particularly if a high percentage of the data to be migrated is unstructured. Once the process shifts from a higher-level inventory of available data to working with the data in more detail, the level of effort required to migrate the data may exceed available budget.

For certain data sets, it may be possible to go back to the original data provider and request an electronic data deliverable (EDD) that will directly upload to the target EDMS. One example of this is requesting that a laboratory send EDDs for historic analytical data still available in its laboratory information management system. The Historical Data Migration Case Study: Filling Minnesota's Superfund Groundwater Data Accessibility Gap describes a data migration where this approach was used successfully. When it is not possible to receive an EDD from the original data provider, it may be possible to copy large chunks of data from the source files (for example, high-quality optical character recognition (OCR) PDFs, spreadsheets, delimited text files) to an EDD. The Electronic Data Deliverables and Data Exchange Fact Sheet contains more information on the use of EDDs.

If the source data are structured, such as a spreadsheet or delimited file, it may be possible to automate the transformation of data from a source EDMS format to a target EDMS format using spreadsheet formulas or software applications that can process large data sets while recording the necessary transformations. Section 3 below provides considerations for automation vs. manual migration. It should also be noted that it is much easier to go from structured to unstructured data than vice versa.

An additional option if staffing is limited is to hire a third party to migrate the data; however, this approach may carry risks if those performing the migration are unfamiliar with the data set, as context may be lost.

2.4.1 Compartmentalization

One of the most effective mechanisms to manage a data migration budget is to assess how data can be compartmentalized into clearly defined subsets during the initial planning process. Examples of data subsets include:

- **Distinct data format:** These data subsets may be further compartmentalized by focusing on one laboratory/data provider at a time in cases where there are different structures/patterns within the data format. Means by which migration steps can be categorized by data format include:
    - data in specific lab EDD format(s)
    - data in less structured digital files
    - data extracted from a PDF or other document file
- **Value of the data to project goals:** Although this approach prioritizes project needs and allows for clear cut-off points, prioritized data sets that meet different project priorities may not be formatted similarly. Means by which migration steps can be categorized by project priority include:
    - data from a specific time period
    - one sample medium at a time
    - one operable unit/property at a time

Even when a predominantly manual approach becomes necessary, compartmentalization can be a key to success for the following reasons:

- An accurate and complete inventory will help the project team decide if any of the data are not needed once it is understood the process will be labor-intensive.
- Each subcategory of data to be migrated can follow a uniform set of instructions that increases efficiency and aids in documenting the process.

By first assessing if the data requiring a manual approach can be lumped into categories, it may be possible to identify portions that can be semi-automated, particularly if the manual approach was selected based on staffing instead of data format.

2.5 Documentation

When automation is not possible, it is even more essential that the process be well-documented by those performing the migration, because it will be the only record of the approach followed should questions arise in the future. Documentation is also essential when using automated approaches, but for manual processes, it will not be possible to refer to lines of code to retrace the steps taken. One possible approach for documenting a migration is to add a note to each item in the inventory created the data audit describing the source and process used to transform the data for upload to the new EDMS. If work is saved in a script or spreadsheet file, a link to that file may also be added to the inventory.

If the decision to tackle this process manually is driven by staffing considerations, there are no-code and low-code options (for example, Microsoft's Power Query or use of formulas in a spreadsheet program) to limit the number of entries typed manually.

Documentation is beneficial because if a migration is continued in the future, there will not be confusion regarding what has and has not been migrated.

3 AUTOMATED VS. MANUAL MIGRATION

When deciding how automated a data migration process should be, "the more, the better" seems like a reasonable target for several reasons:

- Automated processes are faster and typically cheaper, although experienced vs. untrained labor costs may influence this consideration. The remaining points may negate the "cheap labor" advantage of some largely manual workflows.
- An automated process implies reproducibility, which has multiple benefits:
    - Transparency, consistency, and accuracy.
    - Errors will be systematic, making them more likely to be caught and more efficient to fix. When a manual approach is used, searching for random mistakes due to human error requires double checking every data point.
    - Discovery of additional source data than can be processed with the same automation is common.
- Time spent gaining experience in automation will make staff more productive in the future.

Given the advantages of automated over manual processes, it is important to know that things that are often not thought of as automatable are readily automatable with appropriate technologies. A macro or script can automatically iterate over rows in a spreadsheet. However, scripts can also iterate over nested folders of files, so manual setup for each file to be processed is rarely needed. Similarly, PDF files are machine-readable in an automated way with the right technologies (for example, OCR, natural language processing). In addition, compartmentalizing a migration into smaller, clearly defined tasks can aid in identifying opportunities to automate each task.

Although it is preferable to automate as much of the migration workflow as possible for the reasons listed above, there are reasons why portions of the process may need to be performed manually. These reasons include:

- The need to handle source files that have been poorly scanned or are limited to hand-written information (for example, paper field notes) for one of the following reasons:
    - A structured or semi-structured version is not available.
    - There are reasons to question the accuracy of data organized into a semi-structured or structured format that make it worthwhile to start from an original source. (For example, someone who was not fully aware of the data context may have oversimplified it for the sake of organizational efficiency).
    - Manual entry of individual values is highly undesirable as it allows for the introduction of scattered errors that will be nearly impossible to identify without a comprehensive quality control (QC) of each individual record. For files on which OCR is not possible, manual entry is the only option, and values digitized this way should be thoroughly QC'd. Even for documents where OCR is possible, thorough QC is required to catch OCR errors such as "S" instead of "5" and "r n" instead of "m."
- Lack of personnel within the organization who have the requisite skill set. This lack of personnel may fall into one

of two categories:
    - The organization does not have any individuals on its staff capable of performing automated approaches.
    - The organization has only a limited number of individuals versed in automated approaches and a corresponding need for the process to be accessible to personnel unfamiliar with more advanced automation tools to limit scheduling bottlenecks.
- The amount of data is small enough that it will take the same amount or less time to handle it manually, and it is not in a commonly encountered format where it may be worthwhile to develop the automated process.

Automation should not be abandoned just because the source data contain more than one structure or layout. For example, handling variations in spreadsheet structure to process 200 spreadsheets with several years of observation data can be an onerous task when performed manually but take minutes when automated. A saved automated process can be used again if additional spreadsheets are identified.

4 MIGRATION QUALITY CONTROL

Verification of the data migration is critical to the long-term successful usability of the EDMS. The available options are limited by the source and final data structure and the process used to transform and migrate data. At a minimum, an expert who understands the meaning of the data must be involved in the migration.

The source and final data structure may have general features that can be compared, including the number of records (that is, rows), the number of fields (that is, columns), data types, and data values. Automating QC comparisons between source and final data may be an option. One method to build in QC is to use an intermediate or test database during migration where QC is completed before migration to a production database. While not necessary, it can help streamline which data have been evaluated for quality and are ready for use and which are still under review.

4.1 General Checks

Depending on the structure of the initial and final formats, it may not be possible or easy to measure or estimate the corresponding features. For example, if the initial format is in a cross-tab format and the final format is in a long-data format, one may be able to estimate the expected number of columns in the final format by counting the number of row and column headers in the initial format. Conversely, it may not be possible to estimate the number of rows in the final format because some of the body cells in the initial format were filtered out (for example, blank cells). If components of the data migration were known to be of high quality, it may be possible to transform the migrated data from the new system to the original format for an efficient check that a systematic error was not introduced during the migration process.

Plotting data can also be an effective way of spotting issues. As an example, Figure 2 shows a hypothetical plotting of temperature data by month. In this example, plotting temperature data allows visual inspection of outliers. This can lead to further inspection of the data which would identify that some temperatures in May were erroneously recorded in degrees Celsius rather than the specified unit of degrees Fahrenheit. This same error would not be identified as easily with a simple bounds check.

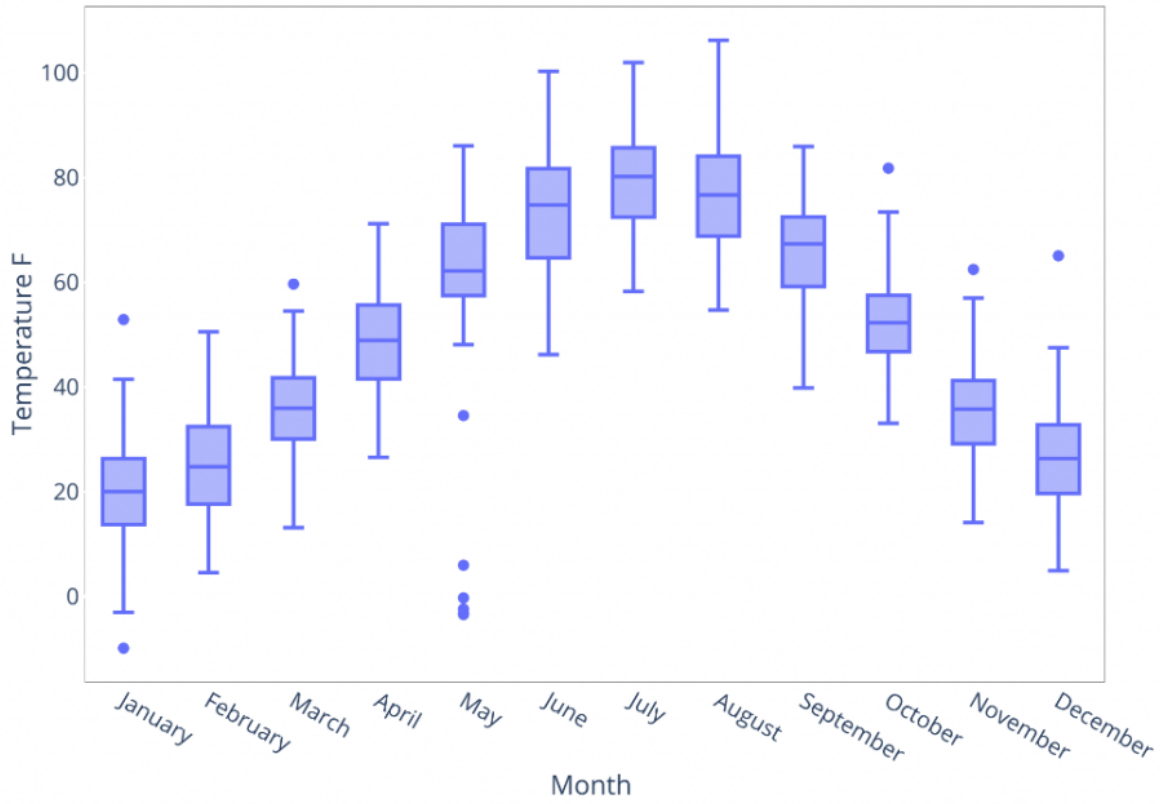

Figure 2. An example of a plot of data revealing an issue of unit inconsistency.

4.2 Data Types

The final data types should generally be determined before data migration and be verified after.

Table 1 describes some common data type mismatches.

Table 1. Common issues in data type format

Expected Data Type	Actual Data Type	Cause
NumericExample: 1.5	StringExample: 1.5*	Non-numeric characters included with numbers
DateExample: 10/26/2021	Numeric/IntegerExample: 44495	Dates stored as numeric or integer but not converted
Factor/Valid ValueExample: Monday	StringExample: Mon	Strings did not follow expected format
IntegerExample: 10	NumericExample: 10.0000001	Integers were stored as numeric; field was calculated and not rounded
BooleanExample: TRUE	StringExample: Yes	Strings didn't follow expected format
Text intervalExample: 4-5 for a sample depth	Date	Some applications will automatically format certain strings as numeric if they are not clearly stored as strings

The final data values should also be verified after migration. Verifying values after a data migration is inherently a manual process because, if it could be automated, that means that another trustworthy method of data migration already exists and, if it did, there would be no need to develop another approach to migrate the data. Broad data-type verification checks are listed in Table 2.

Table 2. Broad data-type verification methods

Data Type	Descriptive Statistics	Visualization
Numeric (for example: integer, double/float, dates, times)	Create a summary table of min, max, range, mean, etc. If factors are present in the data set, consider creating separate summary tables for combinations of factors and numeric data.	Graph values as, for example, histograms, box plots, and time series. For coordinates, plot the coordinates on a base map. If there are multiple numeric values, consider plotting their interactions on, for example, a scatter plot.
Factors/Valid Values	Create a summary table of counts. If there are multiple factors, check the interaction between the factors. For example, check the count for all days of the week and month; maybe all Mondays in October are missing.	Graph counts and interactions in heat maps.
Strings	Check for special characters by creating a summary table of character counts. Create a count of each unique word. Check the length of each string.	Create a word cloud. Create a histogram of string lengths.

In addition to the QC approaches summarized in the table above, checking sample/result comments on historical data can be extremely informative and give hints where other data gaps exist. Any unexpected values should be investigated further.

Specific verification checks include spot checks, which compare a single, original value to the corresponding, single, final value. Spot checks are typically labor-intensive and are measured by the percent of the original data to be checked, such as a 10% check or a 100% check. Spot checks are essentially a random sample, and they may be improved with standard sampling methodologies such as stratification (for example, by initial page, worksheet, column, row, etc.).

4.3 EDDs

Copying large chunks of data from the source files (for example, high-quality OCR'd PDFs, spreadsheets, delimited text files) to the EDD format selected for transmittal of the data can lead to issues if one is not careful to align information properly in the new file, although it will reduce random errors in the data set.

As with any communication, it is possible for some data to be lost in translation. Even if the data are accurately transformed and subsequently migrated to an EDD for transmission to the final EDMS, the EDD used to migrate the data may have additional, specific issues associated with its file format.

For example, CSV files assume that commas are used to delimit separate fields so any unescaped commas in the data may cause parsing issues when loading the data. Other text files may require specific character encodings, such as ASCII or UTF8, so unexpected or special characters may not be accepted. Some file formats automatically—and possible incorrectly—interpret data, such as Excel interpreting some strings as dates or removing leading zeros. There are too many examples to comprehensively document, but one should be aware that there are additional issues to consider when choosing and using an EDD file type for data migration.

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.