The goal of this tutorial is to teach the user how to uniquely catalogue a data set and set up a basic workbook structure, which allows the user to easily manipulate the database output or an initial laboratory electronic data deliverable (EDD).

This tutorial is intended for both beginners to environmental data analysis and practitioners looking to supplement current tools and methodologies.

1. Introduction

For this initial tutorial, we assume the user is provided a table produced by a database in a format similar to an electronic data deliverable (EDD) from the analytical laboratory. Assume this output is provided as a flat file or in a format where the columns are defined categories, and each row represents a unique result. A companion Microsoft Excel workbook is provided in sections below to aid this tutorial.

2. Unique Data Entries

Within a subset of the environmental data analysis world, the primary data set will be in the style of "X samples in Y media for A, B, C chemical analyses." This working format will be assumed for the remainder of this tutorial (although the principles outlined can be adapted as needed). Using this overall data format, every unique result within our EDD can be classified using a subset of information. This information can be summarized using two categories: (1) analyte-specific information and (2) sample-specific information. Examples are included below:

Analyte Information	**Sample Information**
• Analyte or Chemical Name • Chemical Abstracts Service Registry Number (CASRN) • Analytical Method • Chemical Class • Analytical Fraction (e.g., total or dissolved) • Surrogate Information	• Sample Location • Sample Identifier (ID) • Date • Depth • Medium • Sample Code (for example, normal [N] or field duplicate [FD]) • Exposure Area

3. Cataloging the Data set

To help aid this tutorial, an Excel workbook has been created as a companion document.

Excel workbook

The workbook contains many worksheets, but the user should currently concentrate on three worksheets (WS #1a) "Data" – this is the laboratory EDD or database output, (WS #1b) "Sample Info" – currently blank but will be filled, and (WS #1c) "Analyte Info" – currently blank but will be filled. The example database output has been trimmed down and simplified to allow for ease of learning. Values have been removed to concentrate the user on unique identification of entries. The same principles used to catalogue basic data sets scale up with "relative ease" for more detailed data.

This process is rather simple but requires an elementary working knowledge of how to manipulate data within Microsoft Excel. Given how often this step is overlooked, I would like to emphasize the **importance** of this task—the initial way you catalogue/organize data will impact every single step within your workflow from start to finish.

Step 0: Review important assumptions and notes below:

- Colors – There are five color associations used within this tutorial.
    - Green is used to denote active references to the "Sample Info" worksheet.
    - Blue is used to denote active references to the "Analyte Info" worksheet.
    - Orange is used to indicate summary/QC statistics.
    - Gray is used to indicate our static database output.
    - Pink is used to indicate "helper" or supporting columns that contain formulas to assist our workbook.

- Units – To avoid overcomplicating this tutorial, the data set assumed uniform units for each media type. This is often not the case.
    - (Optional Read) Assigning a preferred datatype as a column in “Analyte Info” and then adding a conversion factor (CF) to automatically generate in “Data” as a pink column would be a preferred route. The CF would then be used to generate new results and analytical limits.

- “Extraneous steps” below may be intended to illustrate best QC practices—please read through each step and subparts
    - They may result in an error—this will be explained and rectified in further steps

Step 1: Find the unique identifier for each row or result (WS #2a)

Conceptual Overview—Steps 1 and 2

Every entry within our “data set” should have unique identification. Depending on the source of the data, the overall integrity may be suspect. Historical data sets may contain duplicate entries, or the approach used to reference the data set may not be comprehensive enough. These steps walk through the initial cataloguing of our data set and initial quality control steps.

- We need to find a combination of data categories that exists only once within our data set.
- To aid further steps, we would create two columns next to our existing data file—one called “Sample_MATCH” and another called “Analyte_MATCH”:
    - Left or right does not matter, try not to break up columns within the EDD because uploading new information later will require extra steps—chance for manipulation errors.

- Add a third column called “Unique_MATCH” that combines “Sample_MATCH” and “Analyte_MATCH”. This combination will be our way to uniquely identify each result or row within our data set.
- A fourth column called “Unique?” will be used to count the entries in the “Unique_MATCH” column and confirm only one entry exists.
    - The formula to use under the “Unique?” column is outlined below.
    - COUNTIF(S$3:S$31622,S3)=1, where “Column S” refers to the “Unique_MATCH”.
    - TRUE will indicate the result is only counted once.

- The beginning of this tutorial, WS #2a will assume that “Client Sample ID” and “CASRN” will provide us a unique match for all results (see Step 2).
    - Note that the vertical pipes symbol (“ | ”) is used to separate combined fields in these instructions. You can use whatever symbology you’d like, but vertical pipes will be used globally throughout this tutorial.

Step 2: Resolve Duplicate Matches (WS #2b)

- FALSE has been returned using our initial criteria—that’s okay. At this point one needs to review “Sample_MATCH” and “Analyte_MATCH” for the rows returning FALSE (use filters on the header row to make this easier).
    - Generally, by filtering to only show FALSE, the duplicate entries should be apparent. Sometimes historical data sets can contain duplicate or triplicate results. However, in the case of a new laboratory EDD, this is less likely. For example, one cause could be not including a “Sample Code” in the “Sample_MATCH” tag or “Analytical Fraction” in “Analyte_MATCH”. If the data set includes matrix spike or matrix spike duplicates, this could yield three results under the same Sample ID.

- A review of FALSE entries reviews the following discrepancies that will need to be corrected (WS #2c).
    - Sample Type—need to specify N or FD in “Sample_MATCH”
    - Analytical Fraction—need to specify T (total phase) or D (dissolved phase) in “Analyte_MATCH”
    - Test Method/Chemical Class—need to specify either field in “Analyte_MATCH”
        - Naphthalene is present under multiple analyses—represented both as a volatile organic compound (VOC) and polyaromatic hydrocarbon (PAH).
        - Generally, a simplified “Chemical Class” field needs to be linked by Test Method—for the purposes of this tutorial, the methods have already been categorized.

    - The same set of data categories must be used for all results in the “Sample_MATCH” and “Analyte_MATCH” columns, respectively.

- Oh no—we are still returning FALSE entries. Filter “Unique?” by FALSE and “Unique_MATCH” by any singular entry.
    - Duplicate results are plaguing our data set. We need to use an interim step to rectify the situation. It would be a good idea to follow up with the source of this data (e.g., database or laboratory) to resolve the duplicate entries as well.

- See WS #2d, an extra column has been added solely to resolve duplicates.
    - Note – the “$” used to limit the top of our vertical range has been removed. This will allow the second entry (duplicate) not to count the first and consider itself unique.
    - We will then filter the column “Duplicate?” by yes and delete these entire rows (see WS #2e).

- Our data set (WS #2e) is now resolved and uniquely matched.

Step 3: Extract sample-specific and analyte-specific information

Conceptual Overview – Step 3

Interim steps are often required to build a working tool around a data set; these interim steps often do not get captured within the final working product. This step is intended to explicitly walk through extracting information using Microsoft Excel. Other software or scripting tools can be used to aid overall workbook development as well.

- Now that our initial data set has been QC’d and we have uniquely identified every entry, we can extract information to aid further table development.
- Select all data in our worksheet (WS #2e) prior to selecting “PivotTable” under the “Insert” ribbon. Choose the option for the PivotTable to appear on a new worksheet (WS #3a).
- Prior to manipulating this PivotTable, let’s review our goal here—we want to populate the fields within our Sample Info (WS #1b) and Analyte Info (WS #1c).
- To populate the Sample Info worksheet, follow these steps (WS #3b):
    - From WS #1b, copy the headers from column C:G into Row 3 of WS #3b.
    - Under PivotTable Fields, click on “Sample MATCH”, it should appear under the “Rows” under the “Drag fields between areas below:” dialogue.
    - Click on the PivotTable to reopen the fields, select “Client Sample ID”, “Matrix”, “Sampling Date”, “Sample Type”, and “Sample Depth” so that they are all checked on and appearing underneath our unique identifiers.
    - Use C4:G4 to link the cells to the appropriate entry within the PivotTable.
    - Select cell range C4:G9, then use the autofill feature in the bottom right of the current selection and drag the selection all the way to the bottom of our PivotTable.
    - Copy range C4:G682 and “Paste as Values” into range I4:M682.
    - Select I4:M682, under the “Home” ribbon select “Find & Select” → “Go to Special…” → select “Blanks” → right click and “Delete” and “Shift Cells Up”.
    - These reduced entries can now be pasted into “Sample Info” worksheet (WS #3c).

        - Use “Custom Sort” to sort by “Medium”, “Sample Name”, “Depth”, “Sample Code”.
            - At this point, the use of leading zeros in sample naming convention becomes apparent. The manual manipulation of rows has been performed to simplify the companion workbook.

- To populate the Analyte Info worksheet, follow these steps (WS #3d):
    - From WS #1c, copy the headers from column C:G into Row 3 of WS #3d.
    - Under PivotTable Fields, click on “Analyte_MATCH”, it should appear under the “Rows” under the “Drag fields between areas below:” dialogue.
    - Click on the PivotTable to reopen the fields, select “Analyte”, “CASRN”, “Analytical Fraction”, “Test Method”, and “Chemical Class” so that they are all checked on and appearing underneath our unique identifiers.
    - Use C4:G4 to link the cells to the appropriate entry within the PivotTable.
    - Select cell range C4:G9, then use the autofill feature in the bottom right of the current selection and drag the selection all the way to the bottom of our PivotTable.
    - Copy range C4:G682 and “Paste as Values” into range I4:M988.
    - Select I4:M988, under the “Home” ribbon select “Find & Select” → “Go to Special…” → select “Blanks” → right click and “Delete” and “Shift Cells Up”.
    - These reduced entries can now be pasted into “Analyte Info” worksheet (WS #3e).
    - Use “Custom Sort” to sort by “Chem Class”, “Fraction”, “Analyte Name”.

Step 4: Assign unique numbering within “Sample Info” and “Analyte Info”

Conceptual Overview – Steps 4 and 5

These steps illustrate how to connect the static worksheets (used for consolidated editing) back to the main data set. Implementation of some basic summary counts can help capture any final discrepancies. Establishing these connections ensures any edits will automatically tie into other tables/tools referencing our main data set.

- Reapply the “Sample_MATCH” and “Analyte_MATCH” from WS #2e within both WS #4a and #4b, respectively.
- In column B of each worksheet, order the entries. For the “Sample Order” in WS #4a, note the number shift used between “Medium”—this will be used as a failsafe when creating sample summary tables.
    - After ordering each column, note that each column was referenced using the Name Manager feature. “SO” refers to “Sample Order” and “AO” refers to “Analyte Order”.
    - Also note the rows added to help lock the reference in place and make it apparent to insert rows within that locked range.
    - Finally, note that our “Sample_MATCH” and “Analyte_MATCH” columns have been referenced as “SO_MATCH” and “AO_MATCH”, respectively.

- Establish QC counts that cross-reference the “Sample Info” and “Analyte Info” worksheets.
    - Note the orange columns added to each worksheet.
    - Counting each sample by chemical class and each analyte class by sample medium helps provide a final cross-check on our data cataloguing.

Step 5: Index the unique numbers back to each result within the EDD

- Within WS #5a add these helper rows identified in Step 4 to our overall data set. This will help in naming functions and locking references in place.
- Note some additional changes to this worksheet. “Unique_Match” has been changed to “SO | AO” and two new columns “SO” and “AO” have been added. The “Duplicate?” column has been removed.

- Within WS #5b note how the INDEX/MATCH function was used to bring in our unique sample-specific and analyte-specific #s applied.
    - The decision to use numbers will be more apparent during the creation of the sample screening table. Additionally, it is much easier and quicker to search for numbers than the unique ID tags originally derived from the data set.
    - Additionally, note the reference, “SO_AO”, applied to our column “SO | AO”. Name references have also been applied to our “Result”, “Limit of Quantitation”, “Limit of Detection”, and “Interpreted Qualifier” columns.
    - References to “SO” and “AO” have been named to indicate their secondary origin (that is, the “Data” worksheet; primary reference is “Sample Info” and “Analyte Info” worksheets initially).

- Reconfirm unique result with “Data” using indexed number combination
- Pat yourself on the back—the hard part is over.

4. Create Sample Screening Table

At this point, creating a sample screening table should be very straightforward. For the purposes of this tutorial, project action limits or other kinds of screening levels will be omitted from the table. Adding screening levels can be easily done by first adding the levels to the “Analyte Info” worksheet and then using INDEX/MATCH to populate the levels into the screening table. The steps below will continue the numbering from previous sections to retain references within the companion Excel workbook.

Step 6: Create unformatted structure for screening table

- Within WS #6a, you will notice the skeleton of a sample summary table with no formatting. The idea is to concentrate solely on the formulas being used.
- Formulas have been added in the highlighted column and row headers—note the formulas used and active references to the other worksheets.
- WS#6b shows the expanded formulas to all other cells and performs a QC summary count in the row beneath the table—since only soil is being presented in this table, the total count is only for soil.
    - The filter in column B can be used to hide rows that do not populate for the table (i.e., dissolved metals for surface soil).

- WS#6c presents the fully formatted table for soil—try creating a table on your own for groundwater following the same steps.

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.