

# Quality Assessment and Relational Organization of the WVDEP NPDES Nutrient Dataset

a

## TECHNICAL MEMORANDUM

for the

West Virginia Department of Environmental Protection

Division of Water and Waste Management

prepared by

Interstate Commission on the Potomac River Basin

May 08, 2013



### **ICPRB Technical Memorandum: PRC 13-1**

This report can be downloaded from the Publications tab of the Commission's website, [www.potomacriver.org](http://www.potomacriver.org). To receive hard copies of the report, please write:

Interstate Commission on the Potomac River Basin  
51 Monroe St., PE-08  
Rockville, MD 20850  
or call 301-984-1908

### **Disclaimer**

The opinions expressed in this report are those of the authors and should not be construed as representing the opinions or policies of the U. S. Government, the U. S. Environmental Protection Agency, the several states, or the signatories or Commissioners to the Interstate Commission on the Potomac River Basin. No official endorsement should be inferred.

### **Suggested citation for this report**

Griggs, A. 2013. Quality Assessment and Relational Organization of the WVDEP NPDES Nutrient Dataset: a Technical Memorandum. Report prepared by Interstate Commission on the Potomac River Basin for the West Virginia Department of Environmental Protection, Water Quality Standards Program. PRC 13-01.

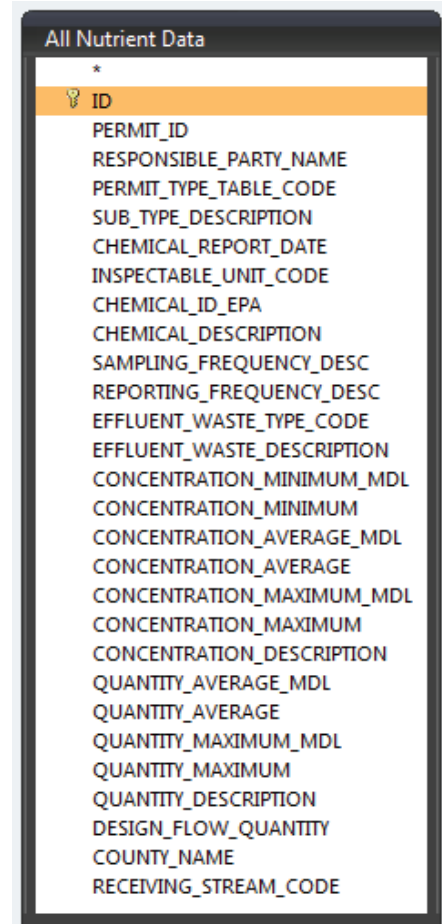
## **ICPRB hand-off March 21st, 2013**

The data arrived as a single .dbf database file with the following individual tables:

- “All Nutrient Data” – a single spreadsheet containing all the data tasked with database development and analysis
- “Chem ID”
- “Removed because not direct discharge outlets”
- “Removed by Effluent Description”

The “Chem ID” table was a lookup table for the chemical numeric identifications reported in the “All Nutrient Data” table. Narrative descriptions of the chemical parameters, however, were also contained in the “All Nutrient Data” table. The two tables “Removed because not direct discharge outlets” and “Removed by Effluent Description” were compiled by WV DEP staff (see above) and the records separated from the “All Nutrient Data” table due to the reasons named in the table titles. In each case, the discharges in question should not be direct discharges to waterbodies and therefore excluded from analyses for waste load allocations.

Only the “All Nutrient Data” table was considered for this project. The remaining data could be easily incorporated into the final structure, following the methods outlined below. The “All Nutrient Data” table consisted of 28 fields and contained 204,296 records individual records (See Figure 1). There were 23 nominal, classification-type data fields that described attributes of the record, and five numerical data fields reporting measured concentrations and quantities.



ID
PERMIT_ID
RESPONSIBLE_PARTY_NAME
PERMIT_TYPE_TABLE_CODE
SUB_TYPE_DESCRIPTION
CHEMICAL_REPORT_DATE
INSPECTABLE_UNIT_CODE
CHEMICAL_ID_EPA
CHEMICAL_DESCRIPTION
SAMPLING_FREQUENCY_DESC
REPORTING_FREQUENCY_DESC
EFFLUENT_WASTE_TYPE_CODE
EFFLUENT_WASTE_DESCRIPTION
CONCENTRATION_MINIMUM_MDL
CONCENTRATION_MINIMUM
CONCENTRATION_AVERAGE_MDL
CONCENTRATION_AVERAGE
CONCENTRATION_MAXIMUM_MDL
CONCENTRATION_MAXIMUM
CONCENTRATION_DESCRIPTION
QUANTITY_AVERAGE_MDL
QUANTITY_AVERAGE
QUANTITY_MAXIMUM_MDL
QUANTITY_MAXIMUM
QUANTITY_DESCRIPTION
DESIGN_FLOW_QUANTITY
COUNTY_NAME
RECEIVING_STREAM_CODE

Figure 1 Structure of the "All Nutrient Data" table as received

### **Investigation of Nominal Data**

The investigation began with identification, categorization, and enumeration of the (non-water quality) nominal data types and data. Examining this information allowed for better understanding of the data fields and informed the design of the relational database structure. Relationship tests were performed to identify fields and properties that assisted in the assignment of unique IDs. A

spreadsheet of tables is provided that details the properties and values of each of the nominal data fields investigated and summarized below:

- Permit IDs (PERMIT\_ID)
  - There are 1,498 unique permits in the dataset
- Permit holders (RESPONSIBLE\_PARTY\_NAME)
  - There are 947 unique permit holders in the “All Nutrient Data” table
  - There is no other unique data associated with the permit holder
  - This information is specific to a unique Permit ID
- Permit types (PERMIT\_TYPE\_TABLE\_CODE)
  - There are only two permit types: INDUST and SEWAGE
  - This information is specific to a unique Permit ID
- Permit Sub-types (SUB\_TYPE\_DESCRIPTION)
  - There are 16 unique permit\_sub\_types
  - This would be an ideal classification for comparing facility types
  - This information is specific to a unique Permit ID
- Report Date (CHEMICAL\_REPORT\_DATE)
  - The dates are end-of-month reporting dates from January 2002 through December 2012
- Inspectable Unit Code (INSPECTABLE\_UNIT\_CODE)
  - There are multiple inspectable unit codes for each permit
  - Under each Permit ID , the Inspectable Unit Code identifies the location a sample was taken and represents the best unique identifier for assessment
- Chemical ID (CHEMICAL\_ID\_EPA & CHEMICAL\_DESCRIPTION)
  - There are nine parameters (water chemistry and flow) measured in the dataset
  - The field is specific to individual records
- Sampling Frequency (SAMPLING\_FREQUENCY\_DESC)
  - There are 66 unique sampling frequencies
  - Sampling frequency varies according to the parameter within a unique inspectable unit
  - The frequency of sampling frequencies reported for each water quality parameter is captured in the query “CRSTB\_WQPARAM X SAMP\_FREQ”
  - Differences in sampling frequencies will complicate comparative or trend analyses
- Reporting Frequency (REPORTING\_FREQUENCY\_DESC)
  - There are six unique reporting frequencies
  - Reporting frequency varies according to the parameter within a unique inspectable unit
  - The frequency of reporting frequencies reported for each water quality parameter is captured in the query “CRSTB\_WQPARAM X REPORT\_FREQ”
  - Differences in reporting frequencies will complicate comparative or trend analyses

- Effluent Waste Type (EFFLUENT\_WASTE\_TYPE\_CODE & EFFLUENT\_WASTE\_DESCRIPTION)
  - There are 33 effluent waste types
  - Many assignments are combinations of other individual types
  - This information is generally specific to Permit/Inspectable Units with some exceptions/errors.
- Design Flow (DESIGN\_FLOW\_QUANTITY)
  - A numerical value that should be specific to an Inspectable Unit
  - Numerous discrepancies and possible reporting errors present
- County (COUNTY\_NAME)
  - There are 55 unique county assignments which correctly correspond to West Virginia's 55 counties
  - This information is generally specific to unique Permit/Inspectable Units
- Receiving Stream Code (RECEIVING\_STREAM\_CODE)
  - There are 585 unique Receiving Stream Code assignments
  - This information is generally specific to unique Permit/Inspectable Units

***Creating the Relational Database***

Storing data in a relational database format helps a user to better understand, compare, manipulate, and analyze data. If data can be stored in a related format, there is a much better grasp of the quality and connectedness of the data. Organizing data into a relational database also helps to find and address errors in the database.

The first step toward relating the information of this dataset was to identify the information that is unique to a Permit ID.

- TBL\_PERMITS - *Identify and assemble a unique permit ID table*
  - RESPONSIBLE\_PARTY\_NAME, PERMIT\_TYPE\_CODE, and SUB\_TYPE\_DESCRIPTION were all unique to PERMIT\_ID

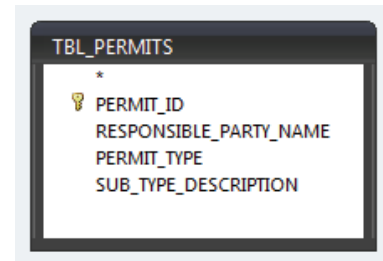
The table below (Table 1) shows the number of unique combinations of Permit ID, Responsible Party Name, Permit Type, and Permit Sub-type code within the database. In this case, the addition of each

Permit ID	Permit ID Responsible Party	Permit ID Responsible Party Permit Type	Permit ID Responsible Party Permit Type Sub-type Code
1,498	1,498	1,498	1,498

**Table 1 Number of unique Permit IDs under nested data classifications.**

new classification did not increase the number of unique assignments, indicating that the information was specific to the Permit ID and there were no mistakes in the database.

A relational table for permits was created. The 1,498 unique assignments identified by PERMIT\_ID, RESPONSIBLE\_PARTY\_NAME, PERMIT\_TYPE\_TABLE\_CODE, and SUB\_TYPE\_DESCRIPTION data were extracted from the original “All Nutrient Data” table and placed in a new table named “TBL\_PERMITS” (Figure 2).



**Figure 2 The design and fields of TBL\_PERMITS**

Now that all permit data was captured and confirmed, an identifier was needed for the unique physical units being monitored under each permit. The Inspectable Unit Code “INSPECTABLE\_UNIT\_CODE” (a numeric assignment between 01 and 316) was found to be that unique identifier. However, the Inspectable Unit Code were not themselves unique, and had to be nested within unique Permit IDs. The remaining nominal data types were evaluated for their relation to the Permit ID/Inspectable Unit assignment in order to build a relational table of Inspectable Unit information.

- TBL\_INSPECTABLE\_UNITS – *Identify and assemble an Inspectable Unit ID table*
  - The following data were found to relate (somewhat) to Permit/Inspectable Units: Design Flow Quantity, County, Receiving Stream Code, and Effluent Waste Type
  - All four data types had inconsistencies, indicating that they were not entirely specific to Permit/Inspectable Units and were investigated to identify and determine the causes thereof

In order to test the specificity of the four data fields to unique Permit/Inspectable Units, the list of unique Permit/Inspectable Units was generated using a select query with the “Group by” function. Adding each individual data-field to the query added a certain number of additional records that had multiple assignments for a unique Permit/Inspectable Unit (See table 2). These additional records

Permit ID	Permit ID	Permit ID	Permit ID	Permit ID
Inspectable Unit	Inspectable Unit	Inspectable Unit	Inspectable Unit	Inspectable Unit
	County	County	County	County
		Receiving Stream	Receiving Stream	Receiving Stream
			Design Flow	Effluent Waste Type
2,673	2,739	2,754	3,130	3,124

**Table 2 Number of unique Inspectable Unit assignments under nested data classifications.**

were identified using a “Find Duplicates” function on the previously generated table. Of the four data

types associated with unique Permit/Inspectable Units, County assignments and Receiving Stream Codes were most specific (less than 100 having multiple assignments). In these cases, discrepancies seemed to indicate actual distinct units being monitored rather than reporting errors when examined more closely. In most cases, when a unique Permit/Inspectable Unit ID had multiple County or Receiving Stream Code assignments, it also differed among the water quality parameters sampled, sampling and reporting frequencies required, or Design Flow. This seemed to indicate an actual unique unit. Design Flow and Effluent Waste Type, however, had non-unique assignments to Permit/Inspectable Units largely due to mistakes in reporting. Many of the non-unique Design Flows were issues of significant digits and rounding – errors that will have to be addressed if design flows are to be incorporated into analyses. Non-unique Effluent Waste Type assignments are most likely the result of reporting errors, and the ~370 errors will require correction if Effluent Waste Type is to be included in the relational database, or used as a classification assignment in analyses. Queries identifying Design Flow and Effluent Waste Type records in question are contained in the provided Quality Assurance Testing database.

TBL_INSPECTABLE_UNITS	
*	
🔑	DISCHARGE_ID
	PERMIT_ID
	INSPECTABLE_UNIT_CODE
	COUNTY_NAME
	RECEIVING_STREAM_CODE

**Figure 3 The design and data-fields of TBL\_INSPECTABLE\_UNITS**

TBL\_INSPECTABLE\_UNITS was constructed by querying out unique Permit ID/Inspectable Unit/County/Receiving Stream Code units and assigning a unique numerical ID (DISCHARGE\_ID) which served as the table key. There are 2,754 unique Discharge IDs in the table.

The next step was to identify unique reporting events associated with each unique discharge.

- TBL\_REPORT\_EVENTS – *Identify and assemble unique reporting events for each discharge*
  - The only field associated with events was the CHEMICAL\_REPORT\_DATE
  - All other remaining fields were specific to individual water chemistry parameters within an event

The same unique identifiers used to build the TBL\_PERMITS and TBL\_INSPECTABLE\_UNITS were used to identify report dates associated with those unique Permits and Inspectable Units. Those dates, and their respective Discharge IDs were queried out to a new table and a unique ID (EVENT\_ID) added which served as the table key.

TBL_EVENTS	
*	
🔑	EVENT_ID
	DISCHARGE_ID
	CHEMICAL_REPORT_DATE

**Figure 4 The design and data-fields of TBL\_EVENTS**

- TBL\_WQFLOW – *Assemble a table of all water chemistry and flow records, and their associated qualifiers*
  - All remaining fields not specific to Permits, Discharge Units, or Events were migrated to TBL\_WQFLOW

The remaining fields from the original “All Nutrient Data” table were specific to individual water chemistry parameters, within a sampling event. For example, a certain discharge unit may be sampled for TN, TP, pH, and flow, however, each parameter will have its own sampling and reporting frequency. Additionally, detection limit flags, units, etc. also are specific to the parameter being measured during a single event. For this reason, the water chemistry data was unable to be transposed when building the relational database. Additionally, since multiple values (minimum, average, and maximum) were possibly reported for each parameter, data could not be easily transposed without leaving values behind. Numerous manipulations were attempted, but in the end, the data was brought into the relational structure in its original format. The number of individual records remained at the original 204,296 initially contained in “All Nutrient Data”.

### ***Quality Assessment of the Water Quality and Flow Data***

The data that was migrated to TBL\_WQFLOW underwent a series of Quality Assurance (QA) tests in order to increase the quality and reliability of the dataset. This was performed prior to establishing the relationships in the database structure. The following is a list of the types of issues found during the QA checks under each data field:

- CONC\_MIN\_MDL, CONC\_AVG\_MDL, CONC\_MAX\_MDL
  - Over 100 different values reported as Minimum Detection Limits
  - Non-MDL values flagged for detection limits
- CONC\_MIN, CONC\_AVG, CONC\_MAX
  - Suspect values
  - Decimal point errors
  - Minimum, average, or maximum reported appears as a single measurement
  - Average reported is higher than MAX
  - Null values
  - Zero values (flagged as DLs or not)
- CONC\_DESC, QUANT\_DESC
  - Incorrect unit assignments for WQ parameter sampled

ID
EVENT_ID
CHEMICAL_ID_EPA
CHEMICAL_DESCRIPTION
SAMPLING_FREQUENCY_DESC
REPORTING_FREQUENCY_DESC
CONCENTRATION_MINIMUM_MDL
CONCENTRATION_MINIMUM
CONCENTRATION_AVERAGE_MDL
CONCENTRATION_AVERAGE
CONCENTRATION_MAXIMUM_MDL
CONCENTRATION_MAXIMUM
CONCENTRATION_DESCRIPTION
QUANTITY_AVERAGE_MDL
QUANTITY_AVERAGE
QUANTITY_MAXIMUM_MDL
QUANTITY_MAXIMUM
QUANTITY_DESCRIPTION
DB_FLAG
SING_MEAS

**Figure 5 The design and data-fields of TBL\_WQFLOW**



### Data Flags in the WQ FLOW Table

After reviewing the data in TBL\_WQFLOW and finding a variety of errors, a list of various flag types were assembled that denoted the types of errors found in the records. The following flags were created in order to capture data errors:

- U Units assigned to record are incorrect for parameter measured
- V Value is suspect
- D Seemingly incorrect decimal placement
- DL Value was flagged in the original dataset as being above or below a detection limit
- Z Zero values reported – may be null
- N Null values reported – no measurement for record
- M Multiple error types

TBL\_WQFLOW was exported to Microsoft Excel™ in order to more quickly sort the data and identify errors. The type and number of data filters applied helped to find certain error types. Each record of a certain error type was flagged with the appropriate flag code. For example, in order to identify records with null values, the filters for the five quantitative data fields were filtered for “blanks” and the DB\_FLAG field was updated to “N”. If new error types were found, than a new flag type was added to the DB\_FLAGS table and the records updated accordingly. The final updated TBL\_WQFLOW table was then brought back into the Access database labeled as TBL\_WQFLOW2. A series of consistency checks were applied to ensure the integrity of the data, and then the new table replaced the old as TBL\_WQFLOW.

### Detection Limits

There were numerous issues encountered involving detection limits in this dataset. The types and number of errors precluded the ability to address all issues under the time allowed on this project. For the moment, all values indicating a detection limit was exceeded were maintained and a flag of “DL” assigned in the DB\_FLAG field. Reporting errors / discrepancies are common in the five Minimum Detection Limit (MDL) fields and occur somewhat randomly. The types of errors vary, but include: non-detection limit values recorded with flags assigned, zero values, multiple detection limit values within parameters, and detection limit values with no flag assigned (not captured by the DB\_FLAG assignment).

### Single Measurement Assignment

An additional data field was added to capture what appeared to be single measurement reports in the database. In the dataset, measurements are reported in the Concentration Minimum, Concentration Average, or Concentration Maximum fields, indicating that multiple measurements are being

reported. However, for many records, the reported Minimum, Average, or Maximum records had equal values, and the sampling and reporting frequencies were the same. Any records with this criteria were flagged with an “S” under the field “SING\_MEAS”, which was added to TBL\_WQFLOW. In total, there were 109, 892 records flagged for being single measurements.

Summary of Data Quality Flagging

Overall there were 21,035 records flagged in the DB\_FLAG field, the majority of these were detection limit flags and null values. The type and number of each assignment are in Table 3. 183,261 records were flagged with “N” indicating no flag assigned.

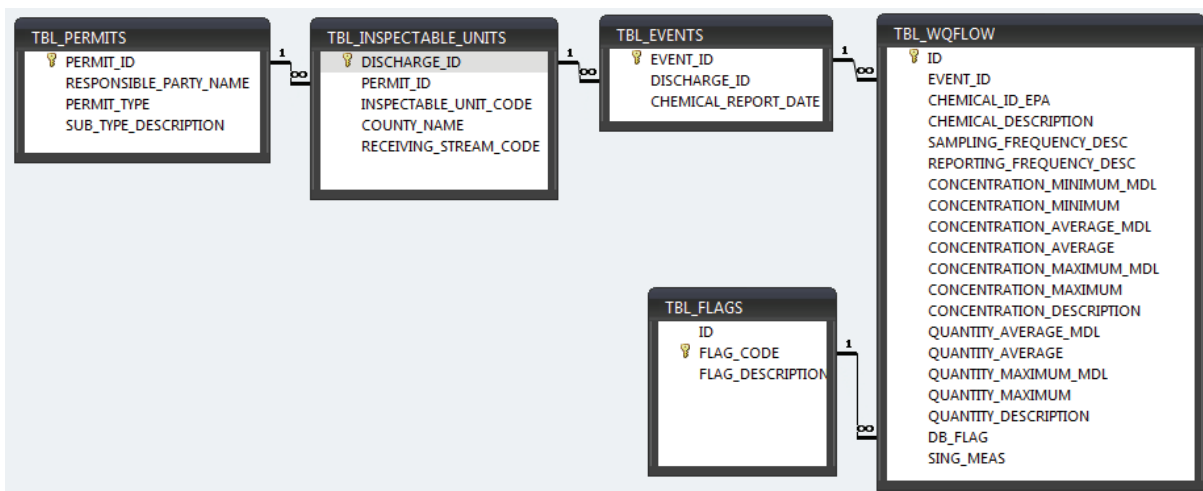
It is likely that more than a few records still require flagging for repair or exclusion before analysis datasets are exported from the final database. This was a best first-cut effort at identifying records that could prove problematic for analyses.

Flag Type	#
Zero Value	162
Incorrect Unit	770
Null Value	7,298
Value Suspect	119
Detection Limits	12,672
Multiple	14

**Table 3 The type and number of database flags assigned under the DB\_FLAG field.**

**Relating the Database**

The five created tables were related to each other using the unique identifiers and keys that were built into each table. The relationships ensure the consistency of the information and help when building queries for specific analysis datasets. Referential integrity was applied to be sure that no records existed in any table without the related information being present in the parent table. In Figure 5 below, the relationships are the connecting lines between tables. On each, “1” indicates a unique set of values located in the parent table, and “∞” indicates multiple records matching a single value.



**Figure 6 The relational structure of the WVDEP NPDES Nutrient database**