Cleaner Data for a Better National Emission Inventory

Sally Dombrowski, Douglas Solomon and Rene Carrier
U. S. Environmental Protection Agency
OAQPS/EMAD/EIG - D205-01
Durham, NC 27711
dombrowski.sally@epa.gov

Abstract

Fixing data can be resource intensive for the U. S. Environmental Protection Agency (EPA), as well as the state/local and tribal agencies (S/L/T). The EPA receives data with referential integrity problems, missing lat/lons, incorrect unit codes, out of range emissions and stack parameters, which all must be corrected or defaulted prior to handing over the inventory to modelers and policy makers. Because we check for these errors in batch mode, S/L/Ts can receive on average three to four different error reports from EPA over the course of inventory development.

Once they have submitted their inventory to EPA, the Consolidated Emission Reporting Rule requirements are considered complete. With the current process, S/L/Ts must re-engage with the inventory to answer requests for corrections. To add to the problem, S/L/T data may not be kept in NIF format and therefore these requests add an additional burden of trying to locate a facility within their own system.

The solution lies in giving as much feedback to the S/L/Ts at one time and as early as possible. EPA believes that developing a pass/fail system, implemented at EPA's Central Data Exchange (CDX), is a potential solution. The pass/fail system would not allow data to be passed on to EPA that did not meet certain minimum requirements. Agencies would receive instant reports back and would be able to make corrections immediately. EPA would receive cleaner data which would limit or may even eliminate re-engagement of the S/L/Ts.

NEI Submissions 1996 - 2002

During the development of the 1996 National Emission Trends (NET) inventory, any type of data was received in various forms, from electronic to paper. Difficulties were encountered in conducting quality control (QC) of these data files as each file contained different format types and lacked consistency.

Several improvements to the National Emission Inventory (NEI-formerly the NET) process were developed and implemented for the 1999 cycle. First, submitters were required to use a new standardized format called the National Input Format (NIF). Second, submission through the Central Data Exchange (CDX) was implemented which allowed submitters to use a uniform system for submission. Finally, a basic quality control tool was developed which electronically checked data files. The tool checked for basic format errors in NIF files, as well as content checks, such as proper use of unit codes, stack parameters and latitudes and longitudes. Since the tool was not delivered to agencies until late in the year, widespread

use was lacking. While the data received was better than in 1996, use of a new standard format and the timing of the QC tool caused errors such as referential integrity and incorrect code use.

The 2002 NEI submissions showed marked improvement. This was the second year that the NIF was being used and there was more widespread use of the QC tool. Unfortunately, not all data were corrected that were identified by the tool. Large amounts of resources were needed to QC the data, contact S/L/Ts for corrections, and default data where data could not be obtained. With looming budget constraints, a new way of doing business needed to be developed for the 2005 NEI effort.

Current Process is Flawed

The submission/data correction process is also problematic for S/L/T agencies. While a lot of agencies are performing the data validation checks prior to transferring the data files to EPA, in some cases, basic format issues are still left uncorrected. EPA is learning of this after receiving the data and has had to attempt corrections with the submitting agencies in order to use the data. This situation introduces redundancy into the process, is an inefficient use of resources, and causes delay in the data development cycle.

During the past year, EPA has been conducting planning calls with S/L/T agencies to discuss changes which will be made during the 2005 inventory cycle. The most frequent complaint from these agencies was the frequency in which they were contacted by phone or e-mail with regard to problems in their inventories. Re-engagement for S/L/Ts is difficult and resource intensive.

A New Approach

Over the last couple of NEI reporting cycles, there has generally been more data submitted by the S/L/Ts – on time, and with less formatting errors. The agencies are implementing the NEI data collection formats with much more success and many are using the data format and quality checking tools prior to sending their files on to the EPA. During that same time, EPA has continued to evolve its data quality checks as well as look for efficiencies in processing all data we are now receiving. The time has come to build on this success with the next challenge.

For 2005 NEI, EPA is proposing to only accept data files for the NEI that pass a standard set of quality control checks. Standardizing a quality control practice will clarify the EPA's data expectations as well as help automate and facilitate the practice. It is proposed that these data validations be performed at EPA's Central Data Exchange (CDX) as the data are received. If the data do not pass the required standards, the data and an error report will be sent back to the data submitter for correction and re-submission. If a submission passes the data standards, the dataset will be forwarded to EPA for processing. Resources would be expended only on files that are fully successful in meeting the basic quality standards.

EPA believes this approach will benefit both the S/L/Ts submitting data to the NEI and EPA. S/L/Ts will benefit because they will know, at the time they submit, if there are problems with their data submission.

Once a submission is successful, S/L/T agencies are done. They will not receive an email from EPA a month or two later asking them to fix basic format or content problems with their data. It will benefit EPA because data received will be cleaner and ready to incorporate into the NEI. The EPA step of checking the basic format and content will be eliminated.

Pass/Fail System

XML users will not see much difference in their current system. There will be additional requirements added to the schema to reflect these new requirements. Upon submission to CDX, the schema will be validated checked for format and content errors. Any questionable entries will be placed into a report, and the file and report will be returned to the submitter. The submitter will then need to correct any errors in the report and resubmit the data file through CDX until the file passes all requirements.

NIF users will be required to use the Basic Format/Content Checker (BFCC) prior to submission to CDX. The BFCC is being retooled to produce two reports. One report will list all errors which "must" be corrected prior to submission to CDX. The second report will be a list of items which are questionable and could be defaulted by EPA, if not corrected. A data submitter can consult the National Emission Inventory QA and Augmentation Report (http://www.epa.gov/ttn/chief/net/2002inventory.html#info) , and make a decision as to whether a default value is acceptable. After all entries on the "must" report have been corrected and a subsequent running of the BFCC produces a clean report, the tool will attach an electronic signature. Files containing this electronic signature will be passed through CDX. Files without the signature will be returned to the data submitter.

The pass/fail system will check for correct codes, referential integrity and absolute ranges, such as no more than 7 days in a week. Appendix A describes the specific proposed standards, and the information is organized by source type file and would apply to both criteria and HAP process-level reporting, and to both of the two acceptable data transfer formats – NEI Input Form (NIF) or NEI XML Schema. The standards focus on a subset of data elements in each source file format that needs to be implemented correctly and the type of check to assure the correct practice. Pending the outcome of changes for V4.0 of the NIF and NEI XML schema, there may be other standards that should be considered and developed in the future.

Benefits

While a pass/fail system puts the burden of QC on the S/L/Ts, they will save resources by not having to re-engage with the data repeatedly over the course of the inventory development. EPA would be assured of data, free of format errors, and be able to apply resources in looking for things such as changes in trends or emission outliers. Declines in resources at both the S/L/T and EPA levels force us to rethink how we do business and how best to get more for less.

Appendix A - NEI Collection
POINT SOURCE FILE
Proposed Data Element Checks for Successful Submission of Data Files

Data Submission Group (Record)	Data Element Name	Required Data	Must Use Valid Code Value	Primary Key Field Required in all related records	Required Value Range	Comment
Transmittal	RECORD TYPE	●	●			
	STATE AND COUNTY FIPS CODE	●	●	●		
	INVENTORY YEAR	●				
	CONTACT PERSON NAME	●				
	CONTACT PHONE NUMBER	●				
	ELECTRONIC ADDRESS TEXT	●				
	FORMAT VERSION	(●)				Required reporting is conditional. Only implement if multiple versions in play.
	TRIBAL CODE	●	●	●		
Site	RECORD TYPE	●	●			
	STATE AND COUNTY FIPS CODE	●	●	●		
	STATE FACILITY IDENTIFIER	●		●		
	FACILITY CATEGORY		●			
	NAICS PRIMARY	●	●			
	FACILITY NAME	●				
	LOCATION ADDRESS	●				Must be physical address. Check zip code consistency with county FIPS.
	CITY	●				
	STATE	●				
	ZIPCODE	●				Check zip consistency with FIP cnty and surrounding cntys.
	TRIBAL CODE	●	●	●		
Emission Unit	RECORD TYPE	●	●			
	STATE AND COUNTY FIPS CODE	●	●	●		

Data Submission Group (Record)	Data Element Name	Required Data	Must Use Valid Code Value	Primary Key Field Required in all related records	Required Value Range	Comment
	STATE FACILITY IDENTIFIER	●		●		
	EMISSION UNIT ID	●		●		
	NAICS UNIT LEVEL		●			
	DESIGN CAPACITY UNIT NUMERATOR		●			
	DESIGN CAPACITY UNIT DENOMINATOR		●			
	TRIBAL CODE	●	●	●		
Emission Release Point	RECORD TYPE	●	●			
	STATE AND COUNTY FIPS CODE	●	●	●		
	STATE FACILITY IDENTIFIER	●		●		
	EMISSION RELEASE POINT ID	●		●		
	EMISSION RELEASE POINT TYPE	●	●			
	X COORDINATE	●				Check if coordinates outside FIP cnty and X Coordinate > 0.
	Y COORDINATE	●				Check if coordinates outside FIP cnty and Y Coordinate < 0.
	XY COORDINATE TYPE	●	●			
	UTM Zone	(●)				Required reporting if XY COORDINATE TYPE=UTM.
	HORIZONTAL COLLECTION METHOD CODE	●	●			Measurement Accuracy Determination (MAD) Codes for XY
	HORIZONTAL REFERENCE DATUM CODE		●			MAD Codes for XY
	REFERENCE POINT CODE		●			MAD Codes for XY
	COORDINATE DATA SOURCE CODE		●			MAD Codes for XY
	TRIBAL CODE	●	●	●		
Emission Process	RECORD TYPE	●	●			
	STATE AND COUNTY FIPS CODE	●	●	●		

Data Submission Group (Record)	Data Element Name	Required Data	Must Use Valid Code Value	Primary Key Field Required in all related records	Required Value Range	Comment
	STATE FACILITY IDENTIFIER	●		●		
	EMISSION UNIT ID	●		●		
	EMISSION RELEASE POINT ID	●				
	PROCESS ID	●		●		
	SCC	(●)	●			Required reporting for POLLUTANT CODE = criteria; Conditional reporting for POLLUTANT CODE = HAP.
	PROCESS MACT CODE		●			
	WINTER THROUGHPUT PCT				●	Out of range if value <0 or >100, or if sum of the four season throughput pcts < 100 or >100.
	SPRING THROUGHPUT PCT				●	Out of range if value <0 or >100, or if sum of the four season throughput pcts < 100 or >100.
	SUMMER THROUGHPUT PCT				●	Out of range if value <0 or >100, or if sum of the four season throughput pcts < 100 or >100.
	FALL THROUGHPUT PCT				●	Out of range if value <0 or >100, or if sum of the four season throughput pcts < 100 or >100.
	ANNUAL AVG DAYS PER WEEK				●	Out of range if value <0 or >7.
	ANNUAL AVG WEEKS PER YEAR				●	Out of range if value <0 or >52.
	ANNUAL AVG HOURS PER DAY				●	Out of range if value <0 or >24.
	ANNUAL AVG HOURS PER YEAR				●	Out of range if value <0 or >8760 and Value does not equal (days/ week * weeks/yr * hours/day).
	PROCESS MACT COMPLIANCE STATUS		●			
	TRIBAL CODE	●	●	●		
Control Equipment	RECORD TYPE	●	●			CE record may be absent. If CE rec is present these checks apply.

Data Submission Group (Record)	Data Element Name	Required Data	Must Use Valid Code Value	Primary Key Field Required in all related records	Required Value Range	Comment
	STATE AND COUNTY FIPS CODE	●	●	●		
	STATE FACILITY IDENTIFIER	●		●		
	EMISSION UNIT ID	●		●		
	PROCESS ID	●		●		
	POLLUTANT CODE	●	●	●		
	TOTAL CAPTURE CONTROL EFFICIENCY	●			●	Out of range if value <0 or >100.
	PRIMARY DEVICE TYPE CODE	●	●			
	SECONDARY DEVICE TYPE CODE		●			
	THIRD CONTROL DEVICE TYPE CODE		●			
	FOURTH CONTROL DEVICE TYPE CODE		●			
	TRIBAL CODE	●	●	●		
Emission Period	RECORD TYPE	●	●			
	STATE AND COUNTY FIPS CODE	●	●	●		
	STATE FACILITY IDENTIFIER	●				
	EMISSION UNIT ID	●		●		
	PROCESS ID	●		●		
	START DATE	●		●	●	Must be valid date. Must be before End Date.
	END DATE	●		●	●	Must be valid date.
	THROUGHPUT UNIT NUMERATOR	(●)	●			Required reporting if ACTUAL THROUGHPUT value is provided.
	MATERIAL		●			
	PERIOD DAYS PER WEEK				●	Out of range if value <0 or >7.
	PERIOD WEEKS PER PERIOD					Out of range if value <0 or > X (no. of months in period as specified by Start Date & End Date).
	PERIOD HOURS PER DAY				●	Out of range if value <0 or >24.
	PERIOD HOURS PER PERIOD				●	Out of range if value does not equal (days/ week *

Data Submission Group (Record)	Data Element Name	Required Data	Must Use Valid Code Value	Primary Key Field Required in all related records	Required Value Range	Comment
						weeks/yr * hours/day).
	TRIBAL CODE	●	●	●		
Emission	RECORD TYPE	●	●			
	STATE AND COUNTY FIPS CODE	●	●	●		
	STATE FACILITY IDENTIFIER	●		●		
	EMISSION UNIT ID	●		●		
	PROCESS ID	●		●		
	POLLUTANT CODE	●	●	●		
	EMISSION RELEASE POINT ID	●		●		
	START DATE	●		●	●	Must be valid date. Must be before End Date.
	END DATE	●		●	●	Must be valid date.
	EMISSION NUMERIC VALUE	●				
	EMISSION UNIT NUMERATOR	●	●			
	EMISSION TYPE	●	●	●		
	FACTOR UNIT NUMERATOR		●			
	FACTOR UNIT DENOMINATOR		●			
	MATERIAL		●			
	HAP EMISSIONS PERFORMANCE LEVEL	(●)	●			Required reporting if POLLUTANT CODE = (HAP codes).
	CONTROL STATUS	●	●			
	EMISSION DATA LEVEL		●			
	TRIBAL CODE	●	●	●		

AREA & NONROAD MOBILE SOURCES
Proposed Data Element Checks for Successful Submission of Data Files

Data Submission Group (Record)	Data Element Name	Required Data	Must Use Valid Code Value	Primary Key Field Required in all related records	Required Value Range	Comment
Transmittal	RECORD TYPE	●	●			
	STATE AND COUNTY FIPS CODE	●	●	●		
	INVENTORY YEAR	●				
	CONTACT PERSON NAME	●				
	CONTACT PHONE NUMBER	●				
	ELECTRONIC ADDRESS TEXT	●				
	FORMAT VERSION	(●)				Required reporting is conditional. Only implement if multiple versions in play.
	TRIBAL CODE	●	●	●		
Emission Process	RECORD TYPE	●	●			
	STATE AND COUNTY FIPS CODE	●	●	●		
	SCC	●	●	●		
	PROCESS MACT CODE		●			
	NAICS		●			
	WINTER THROUGHPUT PCT				●	Out of range if value <0 or >100, or if sum of the four season throughput pcts < 100 or >100.
	SPRING THROUGHPUT PCT				●	Out of range if value <0 or >100, or if sum of the four season throughput pcts < 100 or >100.
	SUMMER THROUGHPUT PCT				●	Out of range if value <0 or >100, or if sum of the four season throughput pcts < 100 or >100.
	FALL THROUGHPUT PCT				●	Out of range if value <0 or >100, or if sum of the four season throughput pcts < 100 or >100.
	ANNUAL AVG DAYS PER WEEK				●	Out of range if value <0 or >7.

Data Submission Group (Record)	Data Element Name	Required Data	Must Use Valid Code Value	Primary Key Field Required in all related records	Required Value Range	Comment
	ANNUAL AVG WEEKS PER YEAR				●	Out of range if value <0 or >52.
	ANNUAL AVG HOURS PER DAY				●	Out of range if value <0 or >24.
	ANNUAL AVG HOURS PER YEAR				●	Out of range if value <0 or >8760 and Value does not equal (days/ week * weeks/yr * hours/day).
	PROCESS MACT COMPLIANCE STATUS		●			
	TRIBAL CODE	●	●	●		
Emission Period	RECORD TYPE	●	●			
	STATE AND COUNTY FIPS CODE	●	●	●		
	SCC	●	●	●		
	START DATE	●		●	●	Must be valid date. Must be before End Date.
	END DATE	●		●	●	Must be valid date.
	THROUGHPUT UNIT NUMERATOR	(●)	●			Required reporting if ACTUAL THROUGHPUT value is provided.
	MATERIAL		●			
	PERIOD DAYS PER WEEK				●	Out of range if value <0 or >7.
	PERIOD WEEKS PER PERIOD					Out of range if value <0 or > X (no. of months in period as specified by Start Date & End Date).
	PERIOD HOURS PER DAY				●	Out of range if value <0 or >24.
	PERIOD HOURS PER PERIOD				●	Out of range if value does not equal (days/ week * weeks/yr * hours/day).
	TRIBAL CODE	●	●	●		
Control Equipment	RECORD TYPE	●	●			CE record may be absent. If CE rec is present these checks apply.
	STATE AND COUNTY FIPS CODE	●	●	●		
	SCC	●	●	●		
	POLLUTANT CODE	●	●	●		

Data Submission Group (Record)	Data Element Name	Required Data	Must Use Valid Code Value	Primary Key Field Required in all related records	Required Value Range	Comment
	TOTAL CAPTURE CONTROL EFFICIENCY	●			●	Out of range if value <0 or >100.
	PRIMARY DEVICE TYPE CODE	●	●			
	SECONDARY DEVICE TYPE CODE		●			
	TRIBAL CODE	●	●	●		
Emission	RECORD TYPE	●	●			
	STATE AND COUNTY FIPS CODE	●	●	●		
	SCC	●	●	●		
	POLLUTANT CODE	●	●	●		
	START DATE	●		●	●	Must be valid date. Must be before End Date.
	END DATE	●		●	●	Must be valid date.
	EMISSION NUMERIC VALUE	●				
	EMISSION UNIT NUMERATOR	●	●			
	EMISSION TYPE	●	●	●		
	FACTOR UNIT NUMERATOR		●			
	FACTOR UNIT DENOMINATOR		●			
	MATERIAL		●			
	TRIBAL CODE	●	●	●		

ONROAD MOBILE FILES
Proposed Data Element Checks for Successful Submission of Data File

Data Submission Group (Record)	Data Element Name	Required Data	Must Use Valid Code Value	Primary Key Field Required in all related records	Required Value Range	Comment
Transmittal	RECORD TYPE	●	●			
	STATE AND COUNTY FIPS CODE	●	●	●		
	INVENTORY YEAR	●				
	CONTACT PERSON NAME	●				
	CONTACT PHONE NUMBER	●				
	ELECTRONIC ADDRESS TEXT	●				
	FORMAT VERSION	(●)				Required reporting is conditional. Only implement if multiple versions in play.
	TRIBAL CODE	●	●	●		
Emission Period	RECORD TYPE	●	●			
	STATE AND COUNTY FIPS CODE	●	●	●		
	SCC	●	●	●		
	START DATE	●		●	●	Must be valid date. Must be before End Date.
	END DATE	●		●	●	Must be valid date.
	THROUGHPUT UNIT NUMERATOR	(●)	●			Required reporting if ACTUAL THROUGHPUT value is provided.
	TRIBAL CODE	●	●	●		
Emission	RECORD TYPE	●	●			
	STATE AND COUNTY FIPS CODE	●	●	●		
	SCC	●	●	●		
	START DATE	●		●	●	Must be valid date. Must be before End Date.
	END DATE	●		●	●	Must be valid date.
	POLLUTANT CODE	●	●	●		
	EMISSION NUMERIC VALUE	●				
	EMISSION UNIT NUMERATOR	●	●			
	EMISSION TYPE	●	●	●		
	TRIBAL CODE	●	●	●		