QA/QC – An Integral Step in the Development of the 1999 National Emission Inventory for HAPs

Paper #42283

Anne Pope
Emission Factor and Inventory Group, U.S. Environmental Protection Agency, D205-01, Research Triangle Park, NC 27711

Darcy Wilson and Stephanie Finn
Eastern Research Group, Inc., 1600 Perimeter Park, Morrisville, NC 27560

Justin Oh,
School of Engineering/Computer Sciences
North Carolina State University, Raleigh, NC 27606

ABSTRACT

Requirements of the Clean Air Act (CAA) and Government Performance Results Act (GPRA) have established the need for a comprehensive hazardous air pollutant (HAP) emissions inventory that can be used to track progress by the U.S. Environmental Protection Agency (EPA) over time in reducing HAPs in ambient air. To estimate risk and HAP emission reductions, the EPA compiles the National Emission Inventory (NEI) as a model-ready emissions inventory. The EPA previously compiled a baseline 1990 and 1996 NEI, and completed the development of the draft 1999 NEI, Version 2.0. The EPA is currently preparing the draft 1999 NEI, Version 3.0, which will be available for public comment in October 2002.

The NEI contains estimates of facility-specific HAP emissions and source-specific parameters needed for modeling, such as location and facility characteristics (stack height, exit velocity, etc.). Complete source category coverage is needed, and the NEI contains estimates of emissions from stationary point and non-point and mobile source categories. Point source categories include major and area sources as defined in section 112 of the CAA. Non-point source categories include area sources and other stationary sources that may be more appropriately addressed by other programs rather than through regulations developed under sections 112 or 129 in the CAA.

The data sources in the point source NEI are state and local agency and tribal data, industry data, data gathered by the EPA during the development of Maximum Achievable Control Technology (MACT) standards, and Toxic Release Inventory (TRI) data. Because of these multiple sources, the compilation of the NEI for HAPs requires many steps. Key activities include:

- submittal of HAP inventory data by state and local agencies, and tribes;
- blending/merging of data from multiple data sources;
- augmentation of blended data for missing data elements;
- quality assurance/quality control (QA/QC) of the data;
- preparation of draft NEI for external review;

- incorporation of external review comments; and
- preparation of final NEI.

The EPA conducted a variety of internal activities to QC the data including the development of an automated QC tool to identify potential errors with data integrity, code values, and range checks; use of GIS tools; and content analysis of the draft inventory by pollutant/source category and geographic coverage. During the external review of the draft inventory, state and local agencies, tribes, and industry provided external QA of the data.

This paper discusses the 1999 NEI for HAPs development steps of blending/merging of data from different sources, data augmentation, QA/QC, and resolution of errors identified in QA/QC.

INTRODUCTION

The Emission Factor and Inventory Group (EFIG) of the U.S. Environmental Protection Agency (EPA) prepares the National Emission Inventory (NEI) for criteria pollutants and hazardous air pollutants (HAPs) every three years. The EFIG has completed the development of the Version 2.0 draft 1999 NEI, and solicited state and local agency, industry, and EPA revisions. The EFIG also implemented an in-depth quality assurance/quality control (QA/QC) program to identify outliers, facilities located incorrectly, and duplicate facilities and emission estimates. The EFIG is currently compiling Version 3.0 draft 1999 NEI for HAPs and incorporating revisions and new data provided by reviewers and resolving errors identified in the QA/QC of Version 2.0 draft 1999 NEI for HAPs. Version 3.0 draft 1999 NEI for HAPs will be available for review in October 2002 and a final 1999 NEI for HAPs will be available in June 2003.

The point source NEI contains estimates of unit-, process-, or facility-specific emissions and their source-specific parameters necessary for modeling such as location and facility characteristics (stack height, exit velocity, etc.). The data sources in the NEI are state and local agency and tribal data, industry data, data gathered by the EPA during the development of Maximum Achievable Control Technology (MACT) standards, and Toxic Release Inventory (TRI) data. Data from the 1996 NEI were included if 1999 state and local agency data were not available for facilities not included in the 1999 TRI and 1999 MACT databases.

The EFIG conducted a variety of internal activities to QC the data provided by state and local agencies. These included:

- Automated QC format tool to identify potential errors with data integrity, code values, and range checks[1] - the QC format tool is available for use and evaluation by state and local agencies at: www.epa.gov/ttn/chief/nif/
- Geographic Information System (GIS) tools to verify facility locations
- QC Content Analysis to identify potential errors with emissions estimates - pollutant-, source category-, and facility-level emission estimates were reviewed to identify outliers and duplicate emissions and sites

After the errors identified by the automated QC format and GIS tools were investigated, the

EFIG followed specific guidance on augmenting data for missing fields of data.[2] The EFIG then performed content QC to identify potential errors with emission estimates. This paper discusses the 1999 NEI for HAPs development steps of blending/merging of data from different sources, data augmentation, QA/QC, and resolution of errors identified in QA/QC.

.

INVENTORY AND QA/QC PROGRAM GOALS

Inventory Uses and Goals

Before undertaking any QA/QC effort, it is important to clearly identify the intended use of the inventory and the goals of the QA/QC program. The ultimate use of the NEI is to provide information needed to protect human health and the environment. Data are used for air dispersion modeling, regional strategy development, regulatory development, air toxics risk assessment, and tracking trends in emissions over time. For example, emissions data from the NEI for HAPs are used in EPA's National Scale Air Toxics Assessment (NSATA) program to estimate population exposures and potential health effects.

Thus, the quality and completeness of the NEI data are of utmost importance. To this end, state and local agency data are critical, because most agencies have emissions inventory and/or air quality permitting programs in place, and update their inventories on a regular basis. As noted in the point source NEI documentation, state and local agencies provided a large amount of data for the first draft of the 1999 NEI.[3]

An additional internal EFIG goal is to stay on schedule. Development of the NEI is a multi-year process with several state and local agency review phases. The EFIG established an ambitious timetable for all participants because of EFIG's responsibility to continuously improve the NEI - to update the data as often as possible, within the resources available each year, and in a manner that makes it ready to process for air quality modeling and risk assessment purposes. It is possible, unfortunately, that these three overriding goals (quality, completeness, and schedule) may conflict with one another in the inventory development process.

Goals of QA/QC Program

The primary goal of the QA/QC program is tied to the data blending/merging of emissions estimates from state and local agencies with EPA MACT and TRI data. While the process used to compile the emissions data from multiple sources continues to be refined by the EFIG, it is inevitable that the first draft of the inventory will include duplicate facility, site, and emissions data. In fact, the approach used to compile the inventory errs on the side of including duplicate facilities and sites if there is any question at all that they are in fact duplicates. Reviewers would not necessarily know if a facility or site had already been removed in error. Another goal includes identifying and correcting erroneous emissions data. For the most part, the errors detected in this phase are outliers with very high emissions estimates. The last goal is to verify the coordinate data in the NEI. Coordinate data are key in identifying duplicate facilities, and when the data are used in air quality modeling it is important to site an emissions source correctly. Other QA/QC efforts that will be completed prior to the release of the final NEI focus on the review and augmentation of the stack parameters.

IMPLEMENTATION OF THE QA/QC PROGRAM
In Conjunction with Inventory Development

To a large extent, the inventory development and QA/QC programs were performed in concert with one another, as data blending/merging and data correction procedures are interrelated. The NEI QA/QC process was initiated immediately after state and local agency files were provided to EFIG. An automated QA program was developed and used to check each file for format and data field errors. Format checks are based on the minimum data requirements for file acceptance by EFIG. Data field checks are related to the codes, numeric data ranges, and locational data in the file. The EFIG accepted data with data field errors as these could be corrected with minimal effort.

Duplicate records were then removed, along with records that had null and zero emissions values. Referential integrity violations, invalid codes, and erroneous locational data were then corrected (or added) if possible.

These files were then used in the data blending/merging process, as facilities included in the state and local agency files were compared with those in EPA's MACT and TRI databases. Duplicate facilities were identified based on the state, county, facility name, address, and location (latitude/longitude), and common IDs. An automated facility-matching program was run to identify common facilities. The program is designed to first attempt to pair up facilities by state, county, and facility ID (if provided). For unmatched facilities, an algorithm was then applied that strips out punctuation and leading/trailing spaces, drops insignificant punctuation (e.g.,: _ - * "), standardizes corporate tags, compares the facility names on a case-insensitive basis, and identifies similar-sounding facility names in each county with exact locational data, then similar coordinates. "Candidate" pairs were then reviewed manually.

Difference between Site and Facility ID

To better understand the matching and blend/merge processes, it is important to distinguish between the terms "site" and "facility" as defined by the NEI for HAPs. In the NEI for HAPs, there can be multiple "sites" associated with the same NTI Unique Facility ID. (The NTI Unique Facility ID is currently stored in the Federal Facility ID field in the Site table.) Each of these sites will have a unique record in the Site table, with a unique site ID. There are two reasons for this one-to-many relationship between facilities and sites:

- Multiple data sources have supplied data to the NEI for the same facility; or
- One source supplied multiple site records for co-located facilities.

For example in the first case, a state may have submitted a set of records for a facility with site ID AL001. This site ID is part of the primary key in all of the remaining tables, Emission Unit, Emission Process, etc. The NIF, Version 2.0 contains more information on the data structure of the NEI for HAPs.[4] The EPA may have provided MACT data for the same facility under site ID EM234. Although these data are for the same facility, its emissions are for different processes at that facility and do not duplicate the emissions data submitted by the state. Rather than attempt to change the site ID in all tables to be consistent with one ID or the other, a common NTI

Unique Facility ID was assigned to the two different site IDs. Not only is it easier to make this assignment than change the site IDs in the remaining tables, it preserves the original site IDs. This helps reviewers track their data in the review process, aids users in tracing the origin of data, and helps EPA compare data from the same sites from year to year.

The records in Site table would appear as follows.

State FIPS	County FIPs	Site ID	NTI Unique ID	Facility Name
01	001	EM234	NTIAL001	AAAPaperMill
01	001	AL001	NTIAL001	AAAPaperMill

In the second case, one data source may have submitted data for closely located but distinctly separate sources of emissions under separate site IDs. This is a situation similar to the one discussed above. For example, Randolph Airforce Base submitted data under several site IDs. Each of these sites correspond to a different emission process:

NTI Unique ID	Site ID	Facility Name	SCC	Process Description
NTI11234	TX0113947	Randolph Air Force Base	10200602	Boiler
NTI11234	TX0113950	Randolph Air Force Base	20400101	IC Engine
NTI11234	TX0112953	Randolph Air Force Base	40400498	Working Losses
NTI11234	TX0113961	Randolph Air Force Base	40400270	Standing Losses

MACT Code Assignment Process and Blending/Merging of Data

After the NTI Unique Facility IDs were assigned to facilities common to the different data sources, MACT codes were assigned. This was an important step in the blending/merging process. These codes were assigned at either the site or process level (but not both) based on:

- emissions data provided by the MACT engineer,
- a facility list provided by the MACT engineer, or
- the Standard Industrial Classification (SIC) code and the Source Classification Code (SCC) associated with the site and its process records.

For example, if a site had a SIC code of 4922, it was assigned a MACT code of 0504 (Natural Gas Transmission and Storage). All of the processes associated with this site would be associated with the same MACT and no further examination of SCCs was performed. If, however, the SIC code field was blank or did not have a match, the SCC was used to assign the MACT code to the process level. Thus, some processes at a site might be associated with a MACT while others might not be. In all cases, any one process, and hence any emissions record, can be tied to one and only one MACT category. If data were supplied by EPA for a particular MACT category, the appropriate MACT code was assigned, and the SIC codes and SCCs were NOT used to default the MACT code. If neither a SIC code or SCC is available, and the site has not been flagged by EPA, the MACT code field remains blank.

In the merging of the different data sets, where data were supplied for the same facility, MACT category, and pollutant from two or more data sources, one data source was chosen. The records from the other data source were not brought into the draft inventory. In this manner, duplicate

emissions were eliminated. In choosing which records to keep, the order of hierarchy for the most part was as follows: state and local agency (S) data were preferred over MACT data (M), which were preferred over TRI data (T), which were preferred over 1996 base year NEI data (N). An exception to this approach was given to municipal waste combustor (MWC) ESD data and mercury ESD data for coal-fired utilities.

For example, three data sources have records for the same facility, MACT category, and pollutant.

State	County	NTI Unique ID	Pollutant Code	MACT Code	Data Source	Keep?	Choices?
01	001	NTI67	300	0713	State	Keep	(S_M_T)
01	001	NTI67	300	0713	MACT	Delete	(S_M_T)
01	001	NTI67	300	0713	TRI	Delete	(S_M_T)

In this case, state data were retained.

The MACT Code assignment process was not perfect. Incorrectly assigned MACT codes and non-assigned MACT codes may have led to retention of duplicate emission records. In the example above, if the MACT codes were assigned differently, all three records would have been retained.

State	County	NTI Unique ID	Pollutant Code	MACT Code	Data Source	Keep?	Choices?
01	001	NTI67	300	0713	State	Keep	(S)
01	001	NTI67	300	0101	MACT	Keep	(M)
01	001	NTI67	300	Null	TRI	Keep	(T)

In this second case, MACT codes were wrongly assigned because of missing or erroneous SIC codes or SCCs. As a result, all three data points were retained and emissions are duplicated.

Following Preparation of First NEI Draft

After the first draft of the NEI was compiled from state and local agency and EPA data, it was made available for review from October 1, 2001 – February 1, 2002. Reviewers were asked to identify duplicate facilities, sites, and emission records, identify facilities not operating in 1999, and to provide emissions data for missing facilities. State files in NIF Version 2.0, documentation, and summary files were posted on the ftp site at: ftp://ftp.epa.gov/EmisInventory/draftnei99ver2/haps/ .

Data Summary Reports

A series of reports were created to summarize the data in the draft 1999 and facilitate its review. These reports were posted on the ftp site and are as follows:

1) 96-99 Site List
2) 99 NTI County Emissions Summary
3) 99 NTI Facility Emissions Summary
4) 99 NTI Data Source Summary

The first report, "96-99 Site List," provided a list of sites found in the 1996 base year NEI, the 1999 draft, and sites common to both the 1996 and 1999 inventories. By sorting this list by state, county, and facility name, reviewers could evaluate the sites listed in each county and detect potential duplicates. If a site was found in the 1996 version but is not in the 1999 draft, reviewers could verify that the facility closed. The "County Emissions Summary" report provided a snapshot of the emissions for each HAP in each county (with emissions divided into stationary area and major subtotals), while the "Facility Emissions Summary" provided a detailed list of HAP emissions per facility. These summary tables enabled reviewers to target states, counties, and facilities for detailed evaluation. Lastly, the "Data Source Summary" table provided summary emissions data for each facility/HAP/MACT combination where more than one data source (state or local agency, TRI, ESD) was available. The report indicates which source was selected for inclusion in the draft and allows a quick comparison of data from the different sources.

The EFIG received 1999 HAP stationary source inventories from 48 agencies located in 39 states in June 2001 and received revisions for stationary source inventories from 34 agencies located in 28 states in February 2002. As of March 1, 2002, agencies and tribes in 40 states have either provided 1999 HAP inventory data or revisions to the EFIG. Revisions/and or inventories were also received from MACT engineers for 92 of the 134 MACT source categories and from 9 facilities or industrial trade associations by February 2002.

Evaluation of Locational Data

The EFIG continued the QA/QC program by first re-evaluating the locational data for each facility. Correct locations are critical to identifying duplicate facilities and sites. To facilitate this evaluation, coordinates provided in Universal TransMercator units (UTMs) were first converted to latitude/longitude in degrees.

After all coordinates were converted, obvious errors in latitude and longitude were corrected, as in the example described below.

All of the X and Y coordinates associated with different emission release points (ERPs) at one facility had coordinates in decimal degrees except for the last pair.

NTI Unique Facility ID	ERP ID	Site ID	X Coordinate	Y Coordinate	XY Coordinate Type
NTI34249	1	25300	-91.38	38.18	LATLON
NTI34249	2	25300	-91.38	38.18	LATLON
NTI34249	3	25300	-91.38	38.18	LATLON
NTI34249	4	25300	-91.38	38.18	LATLON
NTI34249	5	25300	-9138	3818	LATLON

As it appears as if a decimal was simply omitted, this last pair was corrected from:
(-9138, 3818) to (-91.38, 38.18).

Next, we examined all of the coordinates associated with one facility to see if these coordinates might be too far apart and therefore indicate incorrect assignment of the NTI Unique Facility ID. We calculated the standard deviation for all latitudes and longitudes associated with one facility and then compiled a list of all facilities in which the standard deviation for either coordinate was >.02 degrees. These records were set aside for manual review. Conversely, we also created a set of facilities that had identical latitude/longitudes, but different facility IDs to find possible duplicate facilities.

The last stage in the locational data review process was to use a Geographic Information System (GIS) program to overlay the corrected latitude/longitude pairs with the reported county's boundaries. This step included the GIS plotting of each latitude/longitude value and comparing it to the physical boundaries of the county to which the value is associated. If the plotted release point was within five (5) kilometers of an outside boundary of the county, the point was assumed to be valid.

If the plotting comparison indicated that the recorded release point lies farther than 5 km from the county, we attempted to find a valid latitude/longitude using geocoding software[5] or the Facility Registry System's (FRS) database of EPA plant information.[6]

As latitudes/longitudes are critical to finding duplicate facilities from different sources, the tests described above are a key process in blending/merging the data. Unfortunately, the schedule did not allow complete evaluation of the latitude/longitudes prior to the compilation of the draft. During the external review period for the first draft, we were able to complete the activities discussed above and re-evaluate the blend/merge process with more accurate and complete locational data.

OTHER EFIG QA/QC ACTIVITIES

The EFIG developed a series of additional internal QA/QC reports to target outliers, duplicate facilities, and duplicate emissions. The first approach was to evaluate significant changes between the 1996 and 1999 data, and/or extreme variation within the 1999 data. This included comparing 1996 HAP emission estimates to 1999 HAP estimates for each facility, total emissions for each state between 1996 and 1999, and total emissions for each MACT category between 1996 and 1999. These big picture summaries highlighted source categories, states, and facilities with potential problems. The next set of QA/QC reports specifically highlighted individual facilities, and included identifying the top emitters for each HAP nationwide, ranking each facility based on its emissions of each HAP on a national basis, and listing the top emitters for each HAP/MACT combination nationwide.

Detecting Duplicate Facilities and Sites

Duplicate Facilities – Different Facility IDs

Two facilities may have different EFIG-assigned NTI Unique Facility IDs, but upon closer examination they appear to be the same facility. The starting point for this assessment was development of a file that listed facilities with different IDs but identical latitudes and longitudes. Another file that was useful in uncovering duplicates was one that listed the top emitter of each

HAP. This file listed the facility (-ies) which had the highest emissions for each of the 188 individual HAPs. If two different facilities were identified as top emitters for the same HAP, and their emissions were identical; they were examined more closely.

State	Count	NTI Facility	Facility	SCC	HAP	Emissions	Data Source
06	059	NTI22068	FOAMEX	30800801	Methyl Chloroform	446.22	M(M)
06	059	NTICA059210	FOAMEX	30800801	Methyl Chloroform	446.22	S(S)

These facilities were top emitters for methyl chloroform (1,1,1-trichloroethane). This review reveals that these are actually duplicate facilities with identical emissions data.

If the two facilities were located in the same state and county, and had a "similar" name, they were assumed to be the same plant and one was marked for deletion. The second step is to try to determine what prevented identification of these two facilities as being the same—are the coordinate and address data missing or incorrect for one or both facilities? Are the facility names recorded in a way that only a manual review can identify them as the same facility?

Two facilities may also have different EFIG-assigned NTI Unique Facility IDs, identical or similar emissions estimates, completely different facility names, and yet still be the same facility. In these cases, the facilities were flagged for further review to determine if the plant changed ownership and one data source was not updated, or if different subsidiary names were used in the different data sets.

Duplicate Sites – Same Facility IDs

The "top emitters" file was also a useful tool for discovering site duplicates. In many cases the "top emitter" had data provided by multiple data sources. In these cases, it is possible, that one or more sources were undetected duplicates. For example:

Top Emitter	HAP	Emissions	Data Sources
NTI002	Benzene	150 TPY	S, M, T

Upon further examination, three sites falling under this facility have benzene emissions:

NTI Unique ID	Site ID	HAP	Emissions	Data Source
NTI002	AL123	Benzene	36 TPY	S
NTI002	T$124	Benzene	36 TPY	T
NTI002	EM23	Benzene	78 TPY	M

It appears as though, during the blend/merge process these emissions were all retained because they were not identified as being associated with the same MACT category, SIC Code, or SCC. This facility was flagged as a good candidate for manual review. Furthermore, the MACT code assignment was reviewed.

In reviewing duplicate sites, one should not assume all of the emissions are duplicated as well. Two sites may share the same NTI Unique Facility ID, but only some of their pollutant emissions are duplicated. If all of the pollutants were duplicated, one site was marked for complete

deletion. If only some of the pollutant emissions are duplicated as in the following example, only the duplicated emissions records (lead) were marked for deletion.

- Site ID 001 has emissions records for benzene, acetaldehyde and lead; and
- Site ID 002 has emissions records for lead, polycyclic organic matter (POM), and mercury.

As discussed previously, in choosing which records to keep, the order of hierarchy was: state data (S) were preferred over MACT data (M), which were preferred over TRI data (T), which were preferred over 1996 base year NEI data (N). The records selected to be retained and those marked for deletion were then reviewed manually for verification.

Outliers

This type of error is probably the most difficult to spot – what appears to be a high emissions value may in fact be acceptable for a particular facility or source category. To aid in detecting these errors, the emissions data were compared to the range of values in the NEI and the percent contribution to total emissions.

For example, one indication of a potential error was that the value makes up almost 100% of the emissions in the nation for that HAP regardless of the source category, and overall there were a large number of facilities emitting the HAP (thus all the other facilities had *very* small emissions in comparison). Another indicator may be a large increase over 1996 total emissions for the particular pollutant, with much of this increase coming from one facility. A high standard deviation may also be indicative of an outlier, although it is best to view that data point within the context of its MACT category (not shown here)

The following "Top Emitters" table shows an example of the facilities with the maximum emissions for each HAP nationwide regardless of the source category (Table 1). The table includes a count of facilities emitting each HAP, the total 1999 and 1996 nationwide emissions for each HAP, the percent of emissions attributable to the top emitter, the maximum and minimum emission values for each HAP, and the range and standard deviation of the 1999 emission estimates, and the top emitting facility names and IDs.

A summary table with the list of facilities that appear multiple times as top emitters for different HAPs also helped identify sites with outliers. These high values may be due to a series of outliers or duplicated emissions records. The high emissions may also be correct for that facility and category. Thus, these summary data need to be closely reviewed before any records are marked for deletion. Table 2 has example facilities that were the top emitters nationwide for more than one HAP.

Incorrect/Missing MACT Codes

As described above, default MACT codes were assigned based on a site's SIC code and/or the process SCC. However, if no SIC codes or SCCs were provided for the sites and processes associated with a facility, then no default MACT code could be assigned. Furthermore, there are

Table 1. Example of Top-Emitting Facilities

HAP	Mercury	Hexamethylphosphoramide
# Facilities	24660	3
99 Emissions (tpy)	5.35E+06	1.00E+01
96 Emissions (tpy)	1.24E+02	1.43E+02
Max Site (%)	100	99.9
Maximum Emissions	5.35E+06	1.00E+01
Minimum Emissions	1.18E-12	6.56E-10
Standard Deviation	3.4E+04	5.77E+00
Range	5.35E+06	1.00E+01
Maximum Emitter	Tranamerican Waste Central Landfill Inc	Los Alamos National Laboratory
NTI Facility ID	NTIMS1091185	NTINM0001

Table 2. Example of Facilities with Multiple Top-Emitted HAPs

# Top Emitted HAPs	**NTI Unique Facility ID**	**Facility Name**	**State**
8	NTI44507	GOODYEAR TIRE & RUBBER CO	NE
8	NTI16369	HOLNAM INC	MO
4	NTI13363	SIMPSON TACOMA KRAFT CO.	WA
3	NTINM0001	Los Alamos National Laboratory	NM
3	NTIMI26163E9	WAYNE DISPOSAL INC.	MI
3	NTI16788	TENNESSEE EASTMAN DIV.	TN
3	NTIWY5603701	PACIFICORP_JIM BRIDGER	WY
3	NTI43309	SHELL	WV
3	NTI4352	HELENA CHEMICAL COMPANY	AR
3	NTIWV0009	BAYER CORPORATION	WV
2	NTI11352	CARPENTER CO.	KY

cases in which the SCC and SIC code match to different default MACT codes, and in choosing one, the more correct default code could have been lost. As a result, during the data merging process, duplicate emission records would have been retained where different sources were tagged with different MACT codes.

Finding incorrectly assigned or missing MACT codes is a byproduct of examining individual records when looking for duplicate emissions and outliers. Additionally, comparison of overall MACT category sums between 1996 and 1999 revealed potential gross assignment errors. For example, for pulp and paper there may be three MACT codes:

- 1626: Pulp & Paper Production (the original MACT category was replaced with more specific codes below)
- 1626-1: Pulp & Paper Production - Non-Combustion
- 1626-2: Pulp & Paper Production - Combustion (Kraft, Soda, Sulfite, & Semi-Chemical)

MACT code #1626 is an outdated, non-specific code assigned to Pulp & Paper facilities on a SIC code basis. In many cases, SCCs were not available, so emissions for a pulp and paper facility were flagged with this general code. Records from ESD were flagged with the more specific 1626-1 and 1626-2 MACT codes. If a facility in a state data set was flagged with 1626 and the ESD facility was flagged with 1626-1, all of the emissions records associated with both sources have been retained in the draft NEI During the QA/QC process, these duplicate facilities were identified, and many of the facilities flagged with MACT code 1626 were removed from the inventory. Better assignment of MACT codes would have eliminated this duplication.

RESOLUTION

The results of the QA/QC steps outlined above were compiled into several "holding" tables for ease of review and use during the revision process. Each of these tables is discussed below**.**

Duplicate Facilities – Different Facility IDs

These facilities were entered into a "Duplicate Sites/Facilities" table. The site IDs and current NTI Unique Facility IDs were entered into separate records. One NTI Unique Facility ID was then chosen from the pair (or grouping) and assigned to both facilities as the "Revised NTI Unique Facility ID." If one of the NTI Unique Facility IDs did not have a two-digit state code imbedded in it, it means this code was assigned to the facility in the 1996 NEI, and it should be retained as the Revised NTI Unique Facility ID if at all possible. For example:

Site ID	Name	Current NTI Unique Facility ID	Revised NTI Unique Facility ID
4445	Bob's Big Boy	NTI1234	NTI1234
CA0002	Robert's Big Boy	NTICA2314	NTI1234

Duplicate Sites – Same Facility IDs

If one entire site needed to be deleted from the database because all of its information duplicates the information provided for another site (i.e., all pollutants, emissions data, etc.), one site was marked for deletion and the other for retention. When two sites did not have completely duplicate pollutant records, both sites were retained, with only the duplicate pollutant records marked for deletion from one of the sites. These changes were entered into the "Duplicate Sites/Facilities" table and/or the "Duplicate Pollutants" table.

Outliers

Once potential outlier records were identified, they were copied into the "Outlier table." These records were reviewed manually, and EFIG contacted the state and local agency or EPA reviewers to alert them to these questionable values.

Incorrect/Missing MACT Codes

If sites were identified which appear to have duplicate emissions, it may have been necessary not only to mark duplicate emissions for deletion, but also to reassign or fill in a MACT code. If one of the sites had an ESD-based MACT code, this code was assigned to all sites/processes. If both

sites had default MACT codes, it may be difficult to choose one MACT code over another. Further review was necessary to choose the correct code or add missing codes. MACT lead engineers in ESD reviewed the assignment of MACT codes during the period of external review of the draft NEI and helped to resolve many of these issues. The old and new MACT codes (along with relevant IDs) were stored in the "MACT changes" table.

CONCLUSIONS, NEXT STEPS

Although QA/QC procedures were utilized during compilation of the draft point source 1999 NEI, there remained duplicate sites, facilities, and emissions in the draft. There are also incorrect and incomplete latitude/longitudes. This paper addresses the reasons for and the steps that have been taken to correct these problems.

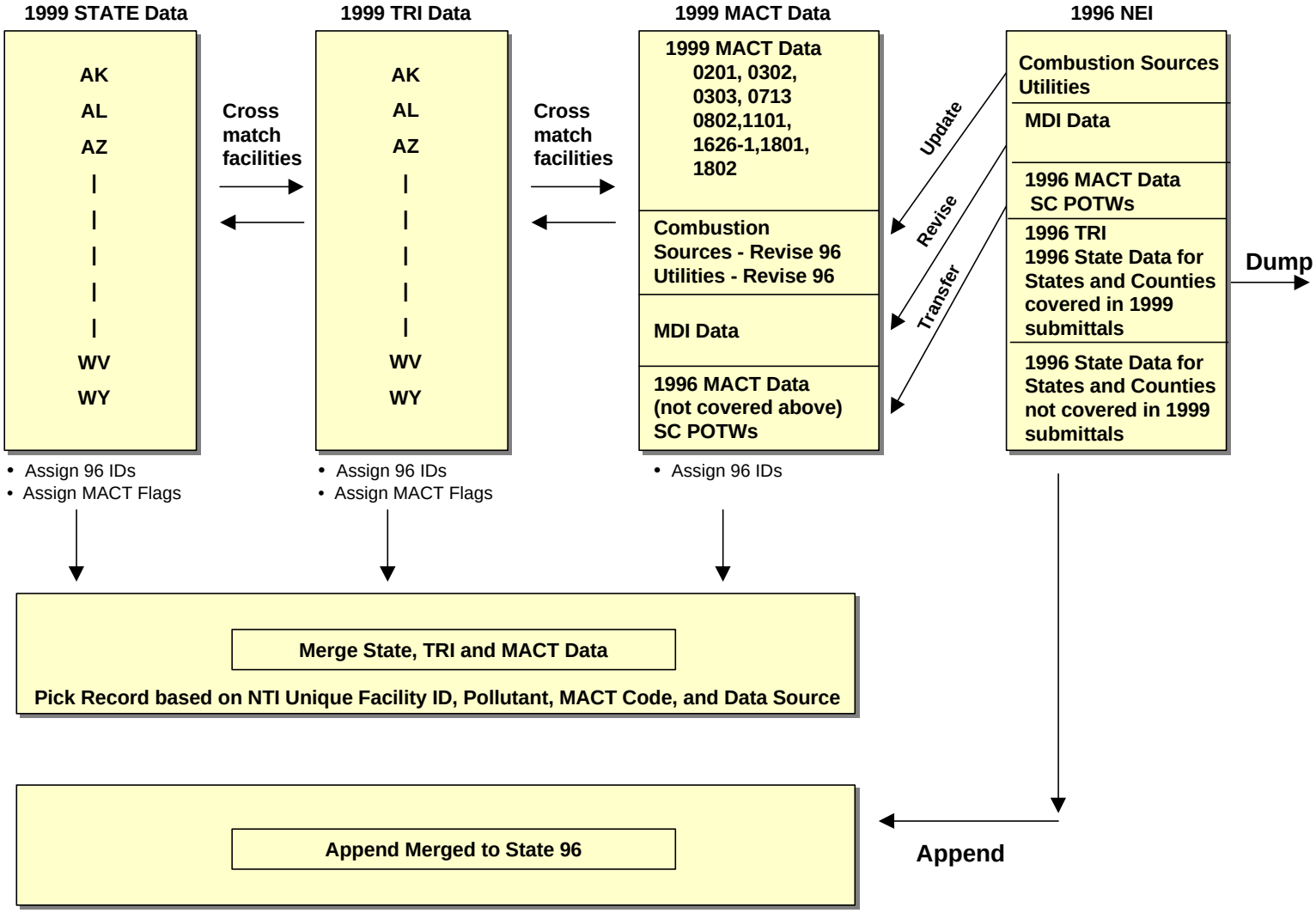
The NEI is constructed from state and local agency and tribal data, industry data, data gathered by the EPA during the development of MACT standards, and TRI data. These data sources were supplemented with data from the 1996 NEI for HAPs. Duplicate facilities, sites, and pollutants were not always detected during the blend/merge process due to: 1) the different ID and naming conventions among the different data sources, 2) variability in the completeness of SIC codes and SCCs, and 3) incomplete and incorrect locational data (addresses and latitude/longitudes). To a certain extent, the duplication of sites, facilities, and emissions is intentional in the draft inventory. Where facilities could not be conclusively matched, both had to be retained to enable external reviewers to make the final judgment. But the draft inventory, and the entire blend/merge and inventory review process, would be much more efficient and targeted if the more complete and consistent facility IDs and names, SIC codes and SCCs, and locational data, were provided for the first iteration.

The blend/merge process is dependent upon facilities being correctly matched across data sources, and MACT codes being assigned properly. In order to assign MACT defaults, SIC codes and SCCs need to be complete and correct. Additionally, the MACT default dictionaries need to be as comprehensive as possible. States can assist in future iterations by providing complete identification information (address, latitude/longitudes, TRI IDs, NTI IDs) and by completing the SIC code and SCC fields. Furthermore, IDs should be maintained consistently from year to year.

In compiling the initial draft of the 1999 NEI for HAPs, many tasks were being completed simultaneously: state data were run through the QA/QC program, MACT data were formatted, TRI data were extracted and formatted, and data gaps were detected and filled with 1996 data (see Figure 1). Also, latitude/ longitudes were checked, facilities from the different data sources were matched, and MACT codes assigned. Although due care was given to each step in the process and to the many individual components of the NEI, it is difficult, if not impossible, to assess the whole through the sum of its parts. Errors in the process itself (e.g., the order of steps) were not readily apparent until global QA/QC was undertaken.

The QA/QC steps discussed here would have been more effective, if the original data sets had

Figure 1. 1999 Draft Point Source NEI (Toxics): Blend/Merge Process



more complete facility IDs, SIC codes, and SCCs, and locational data. In addition, these QA/QC steps should have been performed prior to the release of the initial draft NEI. If MACT code assignment problems had been ironed out and latitude/longitudes thoroughly checked in the compiled database, more duplicate records could have been confidently removed from the draft.

The Data Quality Objective for NEI inventory release includes a schedule component. If schedule is given the highest priority, some aspects of quality and completeness will invariably suffer. Taking an extra month to perform the QA/QC checks outlined in this document would have produced a cleaner draft NEI that would have been easier to review and amend. After performing the QA/QC discussed in this paper and incorporating revisions on the initial draft of the 1999 NEI for HAPs, the EFIG will produce a Version 3.0 draft NEI for HAPs for review in October 2002. This next draft will be greatly improved as a result of the external and internal QA/QC performed by EPA, state and local agencies, tribes, industry and the public.

REFERENCES

1. More information on EFIG's QA/QC National Emission Inventory Input Format tool is available at http://www.epa.gov/ttn/chief/nif/index.html#qa.
2. National Emission Inventory QA and Augmentation Memo, www.epa.gov/ttn/chief/emch/invent/index.html
3. Documentation for the Draft 1999 Base Year Point Source National Emission Inventory for Hazardous Air Pollutants. ftp://ftp.epa.gov/EmisInventory/draftnei99ver2/haps/documentation/.
4. NEI Input Format (NIF) Version 2.0. www.epa.gov/ttn/chief.nif/index.html
5. More information on the Tele Atlas North America "EZ-Locate" geocoding software is located at http://www.geocode.com/.
6. More information on the EPA Facility Registry System (FRS) is located at http://www.epa.gov/enviro/html/facility.html.

KEY WORDS

National Emission Inventory, National Toxics Inventory, air toxics, hazardous air pollutants, quality assurance/quality control, NEI, NTI, QA/QC