A Demonstration of the Quality Assurance (QA) software specifically developed for the National Emission Inventory (NEI).

Rhonda L. Thompson
U.S. Environmental Protection Agency, MD-14, RTP, NC 27711
thompson.rhonda@epa.gov

ABSTRACT

The goal of the Emission Factor and Inventory Group (EFIG) is to develop a high quality national inventory that can be used for a variety of purposes. EFIG has developed the National Emission Inventory (NEI) relational database for both criteria and toxic pollutants. Use of these relational standards minimizes duplication of data and provides flexibility to support different functional requirements of the database over time.

Inventory data developed and supplied by the States are critical data needed to develop a high quality national inventory. Regional offices are crucial to the validation of the inventory data provided by the States. The quality assurance of the inventory data is an iterative process performed by the State and Regional offices with a final QA check by EFIG before loading into the NEI.

To facilitate this evolution, automated QA checks will utilize resources and streamline the entire QA process. A personal computer (PC) software tool is needed to automate QA. Development of this PC software will also provide consistency for criteria and toxic pollutants, and help to establish Regional and State office roles in the QA process. Microsoft (MS) Access was chosen as it is the software that reads the National Input Format (NIF).

The purpose of this paper is to demonstrate the use of the QA software. The particular quality checks and results of the checks will be shown through examples.

INTRODUCTION

The NEI relational database was developed by EFIG to take advantage of relational standards. The NEI requires a specific format for the data to follow. This will provide consistency in the minimum standards to which all input data must adhere. In order for the States and Regions to assure that their data follows the minimum input standards, a list of checks for those standards is provided in our NEI Preparation Plan. The NEI Preparation Plan and formats are located on our EFIG CHIEF website at http://www.epa.gov/ttn/chief. Under Emission Inventories select National Emission Inventory Data and Submitting Data to EPA.

The list of checks have been divided into two types, format and content. Format checks are the minimum required for EFIG to accept the State data. Content checks are provided for the user as possible errors. These are data that the user may want to check and verify as valid before submitting the data to EFIG.

The software has automated the checks where it was possible to do so. There are still other QA practices, such as comparing present to past inventories, that this version of the software will not be able to

perform. This software was intended to QA one State submitted database for one source at a time, for either version 1.2 or 2.0 of the NIF. The software will be posted to the EFIG CHIEF website and available to download. Please keep in mind that errors will continue to be discovered. The executable (.exe) file will be updated periodically to correct errors. Only the .exe will need to be reloaded once the software has been installed.

GETTING DATA READY FOR USE WITH THE SOFTWARE

EFIG has stated that we will accept the NIF in two different file types; ascii or the MS Access database (.mdb) file. For point sources this is either a database with 8 tables or 8 ascii files. The software reads the MS Access database. We are aware that States may have software that generates their inventory in an ascii file type that follows the NIF. We have an import mask, which is part of the empty MS Access file (shell) we have out on the CHIEF website, that imports an ascii file into the associated MS Access database table.

Using the import mask to change ascii files into tables in a MS Access database will also perform some QA checks. It assures that the length and data type of each field is correct. When you import the data you can see with which fields the data lines up. A list of table errors will be generated if your ascii data does not conform to the specified field lengths and data types.

This is done by opening the empty MS Access file (shell). Under File, Get External Data, select Import. In the Import window, find the ascii (.txt) file that you wish to import. Once you click Import, the Import Wizard appears. We want fixed width format, not delimited. Click Advanced to find the Import specifications. When you click Specs... there is one for each table for each of the source types. Highlight the one you want and click Open to return to the specification window to see the specs you chose. Click OK to return to the import text window. Once you click next, you will see the break lines between fields to see if things line up right. Click next and tell it you want to store your data in a new table. Click next and don't change anything as it should be correct. Click next again and choose no primary key. Click next and use the naming convention tblSourceTypeXX as the table into which you import data. For example, tblPointTR or tblAreaEP. Then Finish. Do this for all of the tables in the specific source database. After importing all of the tables into the empty MS Access file (shell), create a .mdb for the specific source and import the tables from the shell into this specific source .mdb. You can then delete them form the empty MS Access file (shell) so that when you use it to import again, the tables will not get confused.

LOADING THE SOFTWARE

Using the install file that will be on the CHIEF website, click Start, Run, and the Setup.exe file to install the NEI software. A user's guide explaining the minimum computer hardware properties necessary to run the software will also be on the website. This guide will include a list of errors encountered during installation and running the software and their solutions.

RUNNING THE SOFTWARE

Currently the NIF and QA software have two versions, 1.2 and 2.0. The first computer screen to appear when running the software identifies which version you are using and asks the user to select from which

source type is the inventory database. Currently the choices are Point, Area and Non-Road Mobile, and On-Road Mobile. It also asks the user to supply the location of the database (.mdb) file that it will QA and the location of the codes database. The codes database contains all of the acceptable codes for the coded fields and is located on the CHIEF website. The other options are to exit or click to the next screen.

The processing status screen appears to show that the software is creating the temporary database c:\tempDB.mdb to store error tables. Then the QA/QC screen appears showing which source type you chose. It also allows you to go back to the previous screen to change source type or databases, to exit, or to go ahead and click the QA/QC button. The processing status screen has the message to "Press the QA/QC button to begin...".

When you go ahead and click the QA/QC button, the screen goes away and the processing status window says that the software first checks for the existence of the tables and fields that it expects to see. If there are table and/or field names that are not correct, a window will appear telling the user that the program detected a table or field with an incorrect name and that a list of problems will be displayed and was saved in c:\DBProbs.txt. The Database Problems window lists which table and/or field names that it expects and does not see. The only options for this window are to exit or press OK. At this point, the table and field names must be corrected for the software to run so it stops. At this time, the user needs to click exit at the QA/QC window and go make corrections to the database, or back if they accidentally chose an earlier uncorrected version of the database on which they have fixed the names.

If all of the table and field names are correct, the processing status window shows on which table the software is checking for errors. It begins with the TR table and goes through to EM. Next it checks referential integrity by going through each of the table relationship checks.

When finished, a window appears announcing that QA/QC is complete and that error tables will next appear for each table (8 for point) and each relationship (9 for point). Even if there are no errors, Error tables appear with titles like "Point Source Transmittal Table Error Report" and "Point Source Transmittal and Site Referential Integrity Report". The primary key fields and an error field are the column headers for the error report. The PK's should make it easy to identify exactly which record the error is in as they are what make each record unique. This should help to make corrections easier. You can save and/or print these .html screens. Do recall that tempDB.mdb has also been created to contain the error tables. You can save tempDB.mdb with another name before you run the software again which will overwrite it.

FORMAT CHECKS

EFIG has determined that some errors preclude the data from being processed and have termed these as format errors. EFIG cannot process data with format errors therefore these type of errors are cause for rejection of the data. EFIG will not accept data with format errors.

There are four types of format errors. The first is incorrect table names. The table name must be of the form tblsourcetypeXX. All table names begin with tbl. The source type is either point, area (includes on-road mobile), or (non-road) mobile. And XX is one of the following record types: TR, SI, EU, ER, EP, CE, AC, or EM.

The next set of format errors all deal with the field properties. The field name must **EXACTLY** follow what is in our .mdb shell or the MS Access field names located to the far right of the excel files on the CHIEF

website. The MS Access database name for the field containing record type should be strRecordType. The str denotes that the data type is text. The MS Access database name for the field containing the start date of the inventory should be lngInventoryStartDate. The lng denotes that the data type is a number that is a long integer. The software is not smart. It expects strRecordType, and will not recognize Record Type, strRecordTypes, strRecord Type, or strRecord_Type. It must be strRecordType. Please follow the field naming convention exactly if you are not sending an ascii file and are submitting an MS Access .mdb file.

The software looks for all of the correct table and field names as the first step of the program. If it can't find all of the correct names, the file c:\DBProbs.txt is created to list all of the correct table and field names that it could not find in the user supplied database and the Database Problems window lists them. For example, if strRecord Type was the field name used, the window will say that it cannot find strRecordType. The file DBProbs.txt is useful if there were so many incorrect names that they could not all be displayed in the window.

The field type must match what is in our .mdb shell or the MS Access field names located to the far right of the excel files on the CHIEF website also. The text data types are all text but the number types are more specific. A long integer is not an integer nor is it a single precision decimal. Even though you have the correct name lngInventoryStartDate and have the field type as a number; if you choose integer, single, or anything other than long integer, you will receive a run-time error that the item cannot be found in the the collection corresponding to the requested ordinal. Please follow the field types in our .mdb shell or the MS Access field types located to the far right of the excel files on the CHIEF website if you are not sending an ascii file and are submitting an MS Access .mdb file.

And lastly for the field properties, the length of each field must be exact. The length for each field as well as the begin and end position for each field is specified in the excel files on the CHIEF website. The field size for the text fields is located in the design view of the tables in the .mdb shell on the CHIEF website. In the import specifications for each table the field name, data type, the start position, and the width for each field are displayed once a certain table is selected.

The next format check that must be correct before we can process the data is on mandatory fields. **All** mandatory fields must be filled in for the data to be processed. Many of the mandatory fields are primary keys. Primary keys are the key fields that relate the tables. The primary keys are in bold and the mandatory fields have an M by the field name on the excel files on the CHIEF website. The key fields are necessary to relate the distinct tables in order to query information from different tables. The most common mandatory field left blank is stack id. The most common reason is that the stack is a vent or a fugitive emission release point. We need the release point because that is where the locational information is stored or, more specifically, the latitude and longitude coordinates. The id is only a local unique identifier so putting in anything like A, B, ... or sequentially numbering them 1, 2, ... is necessary.

The last format check is on the referential integrity (relationship between tables) of the relational database. One violation of referential integrity occurs when there are duplicate records in a table. The primary keys (PK)s are the unique identifiers for a record in a database. The most simple example is the transmittal file (TR). The PK's are State and County so there should be one record for each County in the State. If there are 2 records for the same County, even though there is different contact information, they are duplicate records with respect to the PK's. The software cannot currently check for duplicate records. If there are duplicates, MS Access will not let the user assign the PK's which must be done in order to establish the relationships between tables and enforce referential integrity. An error will appear telling the user that duplicate records have been detected and will not let the user assign the PK's. The user can then detect the duplicates by using the find duplicates query in the database based on the table PK's and delete them.

The next violation of referential integrity occurs when there are widow or orphan records. Take the TR and Site (SI) files for the most simple example. Let's say the TR file has a record for Counties A, B, and C. Let's also say that the SI file has Plant records for only Counties A and B. The TR record for County C is a widow. But what if the SI file had Plant records for Counties A, B, C, and D? Then the SI Plant records for County D are all orphans. The relationships that the software is checking for are all one to many. For example, for every one record in table TR, there is at least one or more matching records in the SI file and will be denoted as TR-SI. The relationships are TR-SI, SI-EU, SI-ER, EU-EP, EP-AC, AC-EM and ER-EM. The CE table has conditional relationships. The relationships EP-CE and CE-EM only need to be checked where CE records exist.

Take care when trying to correct records for widows and orphans. In the above example where the TR record for County C is a widow, you should check the other tables for County C. If the other tables have records for County C, you will need to add a record for County C to the SI table. If the other tables do not have records for County C, you will need to delete the record for County C from the TR table. In the example where the SI Plant records for County D are all orphans, you should check the other tables for County D. If the other tables have records for County D, you will need to add a record for County D to the TR table. If the other tables do not have records for County D, you will need to delete the records for County D from the SI table. This can get complicated for the tables with many PK's so be careful to keep the relationships of ALL the tables in mind before adding or deleting records to a table to make one relationship work.

Recall that EFIG cannot process data with format errors. The four types of format errors are incorrect table names, field properties (name, type, and length), blank mandatory fields, and referential integrity. EFIG will work with the submitter to correct all format errors in an iterative communication process which is detailed in the NEI Preparation Plan on the CHIEF website.

CONTENT CHECKS

EFIG will accept data with content errors. The software does not distinguish between format and content errors. Table and field names are checked first and the software will not continue unless they are correct. After Table and field names are checked and verified as correct, the software produces error reports for each table and then for each of the relationships between tables. The format errors of blank mandatory fields and content checks are in the error reports for each table. The content checks are for acceptable codes, normal numeric ranges, and locational data. Content errors in data are pointed out to the user as not normal and possible errors that they may want to go back to the data to verify.

The most common unacceptable codes are entered into the material and factor fields in the EM table. The most common unacceptable code is UNK or unknown. The acceptable codes for material IO are I, O, and E. I for used, O for produced, and E for existing. Do not put USED in the Material IO field as it is not the code, I is the code. Another common mistake is with the xy coordinate type code. The code for latitude/longitude is latlon not latlong. Another common mistake is with the number of digits in the County FIPS code. There should be three digits. The County FIPS code for the first County is "001" not just "1". These are easy to fix with a find and replace. However, since they are pointed out in the table by table error reports, the reports are much easier to read if they have fewer errors.

The software also checks for "normal" numeric ranges. The most common out of range value entered is 0. The minimum value in all of the range checks is > 0. Sometimes a 0 is a real value. The out of range

checks are provided to tell the user that, although these data values may be real, they are out of the normal range and should be checked and verified before submitted. Another possibility is that the values have not been entered in the specified units. With most numeric values, the NIF allows for different units to be entered. On the stack parameters, however, units were specified. If the user entered the data in incorrect units, these range checks may help point out that a correction before submitting the data to EPA is necessary. If meters / second squared instead of feet / second were entered, the range check may help identify it. Also, if numeric values are in one set of units but the Units field for the numeric value has a different set of units then the range check may help identify this mistake as well.

The last of the content checks is on locational data. Currently the software checks to see if the latitude and longitude coordinates submitted fall within the State boundaries. The State boundaries were created by drawing a box around the State providing a maximum and minimum for each coordinate. For the next version of the software we hope to either draw boxes around the County or add some GIS capability to include a map to show where the coordinates fall.

Recall that EFIG will accept data with content errors. The content checks are for acceptable codes, normal numeric ranges, and locational data. Content errors in data are pointed out to the user as not normal and possible errors that they may want to go back to the data to verify.

SUMMARY

The NEI relational database was developed by EFIG to develop a high quality national inventory. The NEI requires a specific format for the data to follow. Format checks are the minimum required for EFIG to accept the State data. Content checks are provided for the user as possible errors.

Although EFIG has stated that we will accept the NIF in ascii file types, the QA software reads the MS Access database or .mdb file. We have an import mask, which is part of the empty MS Access file we have out on the CHIEF website, that imports an ascii file into the associated MS Access database table.

EFIG cannot process data with format errors. The four types of format errors are incorrect table names, field properties (name, type, and length), blank mandatory fields, and referential integrity. EFIG will work with the submitter to correct all format errors in an iterative communication process which is detailed in the NEI Preparation Plan on the CHIEF website.

EFIG will accept data with content errors. The content checks are for acceptable codes, normal numeric ranges, and locational data. Content errors in data are pointed out to the user as not normal and possible errors that they may want to go back to the data to verify.

REFERENCES

http://www.epa.gov/ttn/chief

KEYWORDS

National Emission Inventory (NEI)

Quality Assurance (QA)
National Input Format (NIF)
NEI Preparation Plan