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

Rhonda L. Thompson
U.S. Environmental Protection Agency, D205-01, 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 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. QA 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 was developed in Visual Basic to automate the QA of a Microsoft (MS) Access database and provide consistency for criteria and toxic pollutants. As this software was designed specifically to QA the NEI, familiarity with its input format is expected.

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. The new features of this version of the software and their benefits will be covered.

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, the NEI Input Format (NIF) is 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. A working knowledge of the NIF is essential to understand the QA of that NIF.

The most notable new feature of the software is that QA checks have been separated into format and content because it does not make sense to check content when the format is not correct. 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 latest version of the software allows the user to choose whether to QA the data for format, the minimum standards required to put the data in the database, or the more resource intensive content or reasonableness checks. When checking for content, the format is also checked as the format must be correct in order for content checks to even be performed.

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 version 2.0 of the NIF. This software does not change the input database. 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 or converter programs that generate their inventories in an ascii file format that follows the NIF. We have an updated 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. The shell and mask have been updated to no longer put in default values that are out of range for the QA software as well as to correct a few typos in the field names.

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 (PK). 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 from 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. In order for the software to be compatible with Windows NT, we now have the capability to specify the path for the error table database that is produced.

RUNNING THE SOFTWARE

Currently the NIF and QA software have one acceptable version which is 2.0. The first computer screen to appear when running the software identifies which version you are using (old version is 1.2) and asks the user to select from which source type is the inventory database. Currently the choices are Point (default), Area and Non-Road Mobile, and On-Road Mobile. On this same screen, the user is asked to specify whether or not the more resource intensive content checks are to be included by checking a box. It explains that format checks are always made as content can not be checked until the format is correct. It is recommended that the user run the format only checks first so that all of the format errors can be corrected before running the content checks. This way there will only be content errors reported when content checks are requested since the format will be clean.

Also on this opening screen, the user is asked to supply the location path of the database (.mdb) file that it will QA, the codes table database, and the error table database. The codes table database contains all of the acceptable codes for the coded fields and is located on the CHIEF website. The error table database is where the format and content error messages are stored for each record by table and table relationships so that the user can browse it to make corrections at their leisure. The other options are to exit or click to the next screen.

The “Preparing to QA/QC” status screen is the next screen that appears. The software then creates the error table database PtempDB2.mdb for point (Atemp and Mtemp for area and mobile) to store the format and content error messages where the user specified that it should go. One of the new features we added is that the error table database is no longer temporary in that is not deleted when the software completes the QA but when another error table database is requested to be written to the same place.

Then the QA/QC screen appears showing the path to which file you chose to QA, which source type you chose, and whether it was format only or both format and content. It also allows you to go back to the previous screen to change your selections, 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 names and field names or data types that are not correct, a window will appear telling the user that the program detected a table or field with an incorrect name or an invalid data type and that a list of problems was saved in c:\DBProbs.txt. After c:\DBProbs.txt is created, the QA/QC screen appears again with the only option being to exit so that you can go correct the table and field names and data types so the software can run. One of the new features is that now the data type of the fields is checked up front. This will hopefully eliminate some of the error messages in the error table database because these things must be fixed for the software to even run. Also when content checks are requested, the software will have had those format errors taken out so that it can make the content check. Remember that content can not be checked without format being correct first.

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 Transmittal (TR) table and goes through to Emission (EM). Next it checks referential integrity by going through each of the table relationship checks. One of the great new features we added to each of the table processing status windows is a record counter. If there is something that is just taking a long time to process like the content checks and you have a HUGE database, you will at least be able to know whether the machine has locked up or if it is really working. You can also tell at which point the software bombs if something goes wrong but most useful is that you can better estimate how long it is going to take something to process.

When finished, a window appears announcing that QA/QC is complete and asks if the user wants to display the error tables. The ability to not display the error tables and only put them in the error table database is another new feature of the software. There is an error table for each table (8 for point) and each relationship (9 for point). Even if there are no errors, error tables are created with titles like “Point TR” and “Point TR_SI” for the relationships. The Primary Key (PK) fields and an error field are the column headers for the error report. The PK’s will make it easy to identify the exact record for which the error occurred as they are what make each record unique. This should help to make corrections easier. If you do decide to display the error reports, you can save and/or print the .html screens. If not, the error table database has been created to contain the error tables. If you do not want to display the error reports, click no and the last screen to appear tells the user that the QA/QC is finished and allows the user to exit the program. Recall that it is recommended that the user run the format only checks first so that all of the format errors can be corrected before running the content checks. This way there will only be content errors reported when content checks are requested since the format will be clean.

FORMAT CHECKS

EFIG has determined that some errors preclude the data from being processed and have termed these as format errors. When the removal of format errors has not been completed the next steps of data processing - content checking, data augmentation, and merging with other databases, can not take place. EFIG cannot process data with format errors therefore these type of errors are cause for rejection of the data. Recall that it is recommended that the user run the format only checks first so that all of the format errors can be corrected before running the content checks. 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, PE, or EM.

The next set of format errors all deal with the field properties name and data type. 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 enough to read the user’s mind. 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 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 an error. 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.

The software looks for all of the correct table names and field names and data types as the first step of the program. If it can't find all of the correct names and data types, the file c:\DBProbs.txt is created to list all of the correct table and field names as well as data types by table that it could not find in the user supplied database. DBProbs.txt lists tables it couldn't find first, then lists field and data types that it couldn't find by table. For example, if strRecord Type was the field name used, the file will have a line like: TBLPOINTTR Field not found: strRecordType. Having a file like DBProbs.txt is useful so that it can be printed out and used to correct the user supplied database.

If the tables and fields have the correct names and data types, the software begins running the rest of the format checks. The last check for the field properties is that 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, 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 for mandatory fields. **All** mandatory fields must be filled in for the data to be processed. Many of the mandatory fields are Primary Keys (PK)s. PK's are the key fields that relate the tables. The PK's 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 emission release point id. The most common reason is that the release point 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 PK's are the unique identifiers for a record in a database. The software lists the PK's for the duplicate records and the number of duplicates with respect to the PK's on the error report for that table in the error table 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 State FIPS, County FIPS, and number of duplicates for that County would be listed on the Transmittal error table in the error table database.

Another 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-PE, PE-EM and ER-EM. The CE table has conditional relationships. The relationships EP-CE (orphans) and CE-EM (widows) 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.

CONTENT CHECKS

EFIG will accept data with content errors. Perhaps the most useful of all the enhancements, is that it now has the ability to distinguish format from the more resource intensive content checks. The first computer screen to appear is where the user can check a box to include the content checks and is reminded that format checks are always made because content errors often can not even be checked for when the format is incorrect. If the format data type is wrong for example, forcing content checks on those fields will automatically yield errors. You can not make a numeric content check if text had been mistakenly put in the field. If numeric data has been put in a text coded field, content checks for the text code do not even make sense to try. Recall that it is recommended that the user run the format only checks first so that all of the format errors can be corrected before running the content checks. This way there will only be content errors reported when content checks are requested since the format will be clean.

The content checks are for acceptable codes, normal numeric ranges, and locational data. The codes table database contains a table listing acceptable codes for each of the coded fields in the NIF. 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. After QA/QC is finished, the software produces error reports for each table and then for each of the relationships between tables in the error table database. The format errors not already checked (blank mandatory fields, invalid field lengths, and duplicates) and content checks are in the error reports for each table.

The most common unacceptable codes are entered into the material, material I/O, and unit fields in the PE and EM tables. 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. If you are using a converter program to get the data from your database into the NIF, you need to be sure that your data is in the format expected of the converter program. For example, when your data is not in the standard units expected of the converter program, the nonstandard units will be replaced with UNK or unknown in all unit coded fields.

Another common code mistake is with the xy coordinate type code. The code for latitude/longitude is latlon not latlong. Also most common is a mistake 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 which is another reason that format needs to be checked and corrected before content is checked. The code table database contains acceptable NIF codes.

The software also checks for “normal” numeric ranges for release point parameters, annual emission values, EP and PE numeric fields, and other temporal numeric fields. The "normal" or expected release point parameter ranges are on the code table database. These are different from the data augmentation procedures in that EFIG augments data with ranges specific to SCC and the QA software checks for very broad National ranges. The "normal" maximum annual emission values are also on the code table database. By annual, the emission type code should be 30 for the entire period, the dates must be annual, and the units must be in tons. These "normal" ranges will be based on percentiles from previous inventories and are conservative in that we want to point out everything suspicious even though it may be real. The temporal field checks on the EP and PE tables and other temporal fields are hard coded since they are obvious. These are for fields with the number of days per week ≤ 7, weeks per year ≤ 52, hours per day ≤ 24, hours per year ≤ 8760 for a non-leap year and 8784 for the leap year, seasonal throughput sum ≤ 100 percent, months between 1 and 12, hours between 0 and 24, and minutes between 0 and 59.

The most common out of range value entered is 0. The minimum value in most 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 possible cause of out of range parameters 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 emission release point 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, for example, meters / second squared instead of feet / second squared 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. If the format for the locational fields is correct, the software will content check these fields. But if the format is wrong, for example, if the coordinate type says UTM yet no UTM zone is present or if the type says LATLONG and the UTM zone is filled, then the software will be confused and cannot make the coordinate content checks.

First 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 thereby providing a maximum and minimum for each coordinate. These maximum and minimum latitude and longitude coordinates are located on the State FIPS code table in the code table database. If the coordinates are within the State boundary, the software then compares them to maximum and minimum latitude and longitude coordinates from boxes drawn around the County which are located in the codes table database.

Recall that EFIG will accept data with content errors. The content checks are for acceptable codes, normal numeric ranges, and locational data and are checked after the format has been verified. Content errors in data are pointed out to the user as not normal and are possible errors that they may want to go back to the data to verify. Recall that it is recommended that the user run the format only checks first so that all of the format errors can be corrected before running the content checks. This way there will only be content errors reported when content checks are requested since the format will be clean.

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 and are checked after the format has been verified.

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.

EFIG will accept data with content errors. The content checks are for acceptable codes, normal numeric ranges, and locational data and are checked AFTER the format has been verified. 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.

It is recommended that the user run the format only checks first so that all of the format errors can be corrected before running the content checks. This way there will only be content errors reported when content checks are requested since the format will be clean.

REFERENCES

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

KEYWORDS

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