Data Miner Version 1.0 Validation Report

Executive Summary

This validation report details the efforts EPA made to ensure Data Miner Version 1.0 operates according to its design. U.S. EPA developed Data Miner under the Regional Air Impact Modeling Initiative (RAIMI) to meet the functional need of accessing large regulatory air emissions databases, such as the Texas Point Source Database (PSDB), to extract information to support air and risk modeling, risk attribution analysis, and solutions tracking.

This validation effort tested the core capabilities of Data Miner. These core capabilities, namely the capabilities to link multiple tables and conduct queries and extractions, constitute the required functionality to meet project objectives. Validation did not test the entire Data Miner feature set, nor did it include regression testing or error handling verification.

The validation used Data Miner's core capabilities to construct a test dataset for each core functional element and compare it to a control dataset built using Microsoft Access. The validation also constructed a view linking tables to produce an emissions inventory containing the data elements necessary to support air and risk modeling, risk attribution analysis, and solutions tracking. Once the tables were linked, the validation effort used Data Miner to generate and execute the query. Testing validated the output Data Miner generated.

While Data Miner will function on any relational database in an InterBase *.GDB format, Data Miner Version 1.0 has predominantly been used with the Texas PSDB and has been validated using a PSDB test data set. Users should consider whether additional validation scope is warranted prior to using Data Miner with a new database.

Functional Capabilities

Air modeling, risk modeling and risk-assessment projects require inspection and analysis of large databases. Because there are numerous tables in emissions databases, each having many fields, it can take significant time to track down particular details embedded in this mass of information.

In addition to the complexity of the information, many applications are unable to handle the massive volumes of data. For example, common desktop software such as Microsoft Excel cannot handle more than 65,600 rows of data. Data Miner overcomes these limitations.

Data Miner is a large client-server database processing system that facilitates the assembly of multi-source emissions inventories for air and risk modeling. With Data Miner, you can:

- Create and edit database table relationships and views for complete access to all emissions attributes maintained in the database
- Link source-specific parameters necessary for air and risk modeling from multiple database tables through the Data Organizer component

- Extract the source-specific data sets by constructing and executing simple or complex data queries in the Query Builder component
- Generate database/spreadsheet tables for input into air and risk modeling components. (Note that the output is not model ready; it will require pre-processing prior to input into models.)

Validation Criteria

Data Miner validation tested the table linkage, query and extraction, and output capabilities.

Validation Approach

The validation approach consisted of the following steps for each of Data Miner's core capabilities:

1. Construct a test dataset.
2. Compare the test dataset to a control dataset built using Microsoft Access.

The Data Miner validation would be considered successful if there were no discrepancies during the comparison. In the event of discrepancies, Data Miner would be validated only if the discrepancy could be attributed to characteristics of the original database, not due to functionality issues associated with Data Miner.

Table Linkage Considerations

The PSDB contains 63 tables. Typically only a small number of these tables must be joined to provide adequate information to conduct air and risk modeling. Other tables may be of interest to support other project objectives, such as attribution analysis or permitting support. The ability to reliably and accurately combine relational data tables is essential to emissions characterization, air modeling, and risk modeling.

Query and Extraction Considerations

Many databases, including the PSDB, contain large information that is not directly or immediately usable to achieve project objectives. Therefore, the capability to select certain data sets over others is necessary. For example, the PSDB contains emissions for the entire state of Texas, but a project may focus on only a single county. The ability to reliably and accurately extract datasets of interest is essential to efficient project and data management.

Execution

The following sections describe the specific steps performed during the Data Miner validation.

Construct Test and Control Data Sets

The validation effort constructed Test and Control data sets using the seven PSDB tables that provide required information for air dispersion and risk modeling, attribution analysis, and solutions management. These tables are:

- AC_ACCOUNT
- CCD_COUNTYCODE
- CE_CURR_EMISS

- CN_CONTAM_NAM
- FC_FACILITY
- PN_POIN_NAME
- PT_POIN

Conduct Validation Comparison

The validation compared a select 30-source test subset, not the entire PSDB. The 30-source test subset was established to include multiple facilities, sources, source types, contaminants, and emission rates and to be representative of the PSDB and information needs typically demanded by Data Miner users. Table 1 presents the 30-source test set.

Build Control Dataset

The validation effort used Data Miner to export the seven individual unjoined PSDB tables in their entirety and import these tables into a Microsoft Access database.
(DM Validation Control.mdb).

Test Table Linkage

The validation effort included the following steps to test the table linkage capabilities of Data Miner:

1. Used the Data Miner Data Organizer Editor to construct a view linking the seven tables to produce an emissions inventory containing the data elements necessary to support modeling, attribution analysis, and solutions tracking. The Data Miner view file '6-20-03 DM Link Test.vef' contains the table list, join conditions, and selected fields used to construct the Linkage Test dataset.
2. Executed the view (without additional query constraints) and exported the results as a Microsoft Access database file (DM Link Test(c).mdb).
3. From the Control and Linkage Test datasets, extracted the records for the 30 test sources.
4. Compared the records for the 30 sources. This information is contained in the Microsoft Excel spreadsheet 'DM Link Compare.xls'. The comparison is a logical test where a positive match returns the value '1', and negative match returns a value of zero.

Test Query and Extraction

The validation effort included the following steps to test the query and extraction capabilities of Data Miner:

1. Used the Data Miner Data Organizer Editor to construct a view linking the seven tables to produce an emissions inventory containing the data elements necessary to support modeling, attribution analysis, and solutions tracking. The Data Miner view file '6-20-03 DM Query Test.vef' contains the table list, join conditions, and selected fields used to construct the Query Test dataset.
2. Used the Data Miner 'Query Builder' component to specify a query that selects all of the following:
   - the 30-source test set
   - stack or flare sources
   - emission rate records greater than 1 ton per year (tpy)
   - emission rate records less than 5 tpy

The text query string used to generate this query is contained in the file '6-23-03 DM Query Test-query string.dat'.

3. Executed the query and exported the results as a Microsoft Access database file (DM Query Test(c).mdb).
4. For the control set, used Microsoft Access to execute a query with the same constraints as above.
5. Compared the records from the Data Miner and Access queries. The results of this comparison are in the Microsoft Excel spreadsheet 'DM Query Compare.xls'. The comparison is a logical test where a positive match returns the value '1', and negative match returns a value of zero.

Conclusions and Considerations

Data Miner Version 1.0 core capabilities have been successfully validated. Specific results follow.

1. Table Linkage: Using the comparison capabilities in Microsoft Excel, the validation effort determined that the test dataset is identical to the control dataset. Both the test and control datasets returned 928 records, and contents of data fields were determined to be identical, as indicated in the attached file (DM Link Compare.xls). (Comparison was done using Microsoft Excel.) Successful completion of this test indicates that the Data Miner table linkage capabilities are repeatable and accurate.

2. Query and Extraction: Test dataset is identical to control dataset. Both the test and control dataset returned 75 records, and contents of data fields were determined to be identical, as indicated in the attached file (DM Link Compare.xls). (Comparison was done using Microsoft Excel.) Successful completion of this test indicates that the Data Miner table linkage capabilities are repeatable and accurate.

Validation Scope Limitations

This validation has been conducted on the Texas PSDB. While Data Miner is designed to accommodate any database in the InterBase GDB format, the application of Data Miner to other InterBase databases has not yet been evaluated. Users should consider whether additional validation effort is warranted prior to using Data Miner with a new database.

Additionally, this validation is limited to the core capabilities of Data Miner. These core capabilities, namely the capabilities to link multiple tables and conduct queries and extractions, constitute the required functionality to meet project objectives. Secondary capabilities of the tool, such as field sorting, have not been validated.

Finally, this validation should not be construed as a validation of the Texas PSDB. Nothing is implied with regard to the quality, completeness, or integrity of the original database.

Table 1 30-Source Test Set				
No.	**UPN[1]**	**Company**	**EPN[2]**	**Type**
1	JE0F01W	AMERIPOL SYNPOL CORP	S-PLANTFLR	FLARE
2	JE0F011	AMERIPOL SYNPOL CORP	F-WWATER	FUGITIVE
3	JE0F00U	AMERIPOL SYNPOL CORP	F-KVGS	FUGITIVE
4	JE0F000	AMERIPOL SYNPOL CORP	S-CBLK	STACK
5	JE0F01K	AMERIPOL SYNPOL CORP	S-NWTRBLST	STACK
6	JE0F01X	AMERIPOL SYNPOL CORP	S-SWTRBLST	STACK
7	JE4D028	DUPONT DOW ELASTOMERS LLC3	NOR-FMPRX	FUGITIVE
8	JE4D021	DUPONT DOW ELASTOMERS LLC	NOR-FAWWT	FUGITIVE
9	JE4D025	DUPONT DOW ELASTOMERS LLC	NOR-FDITCH	FUGITIVE
10	JE4D001	DUPONT DOW ELASTOMERS LLC	HYP-ADS112	STACK
11	JE4D01H	DUPONT DOW ELASTOMERS LLC	NOR-CDR5	STACK
12	JE4D01P	DUPONT DOW ELASTOMERS LLC	NOR-DL36D1	STACK
13	JE1500S	EXXON MOBIL CHEMICALS	EH42	FLARE
14	JE15015	EXXON MOBIL CHEMICALS	EF1	FUGITIVE
15	JE15017	EXXON MOBIL CHEMICALS	EF3	FUGITIVE
16	JE15014	EXXON MOBIL CHEMICALS	EH29C	STACK
17	JE1504P	EXXON MOBIL CHEMICALS	EH9603	STACK
18	JE15036	EXXON MOBIL CHEMICALS	EH55	STACK
19	JE110D3	HUNTSMAN CORPORATION	UER037	FLARE
20	JE1104E	HUNTSMAN CORPORATION	BDFUGS	FUGITIVE
21	JE110ED	HUNTSMAN CORPORATION	JWB1	FUGITIVE
22	JE11027	HUNTSMAN CORPORATION	UW6BB1&2	STACK
23	JE11005	HUNTSMAN CORPORATION	F2HF21	STACK
24	JE11004	HUNTSMAN CORPORATION	F4HF41	STACK
25	JE0S00S	THE GOODYEAR TIRE & RUBBER CO	130FL-Q502	FLARE
26	JE0S009	THE GOODYEAR TIRE & RUBBER CO	130F	FUGITIVE
27	JE0S00J	THE GOODYEAR TIRE & RUBBER CO	370F	FUGITIVE
28	JE0S001	THE GOODYEAR TIRE & RUBBER CO	116S-B101	STACK
29	JE0S002	THE GOODYEAR TIRE & RUBBER CO	116S-B102	STACK
30	JE0S003	THE GOODYEAR TIRE & RUBBER CO	116S-B103	STACK

Notes:

[1] Unique Point Number, generated from PSDB data to assign each source in the PSDB a unique statewide identifier.

[2] Emission Point Number, reported by the account owner and is unique for an account.