U.S. Mine List Data Summary

1/28/2016

Introduction

Use of the U.S. Mine Site Data (MLD) is contingent on the U.S. Mine List Data users reading, understanding and agreeing to the 'Data Use' section of this document (see below).

There are nearly 86K mines listed in this database as of the end of January 2016 for the U.S. and its territories. The MLD contains the most up-to-date information at the point when the information compiled. This product is an enhanced from the original version (see below)

The Excel version of the MLD has some limitations and is provided without charge (for the time being). The commercial version is available as a database and has fees associated with it depending on usage.

This document identifies how the original data has been treated and how it may be applied to your applications. Note that this document will change periodically as new content is added or as provisions of the 'Data Use' change.

Data Sets

There are 2 data MLD data sets: The first is a somewhat limited data set, provided without charge, in an Excel format. The second is a Commercial version provided in a SQL format.

The Commercial MLD is licensed individually and pricing depends on use. Provisioning will be in the cloud or as a file. This set of MLD data is produced monthly, includes lists of period by period changes, identifies new/deleted mines, complete testing results. And an exception log. Also included are all up-to-date reference tables. A feature summary table appears below:

U.S. MINE LIST (MLD)	No-Charge	Commercial
Format	Excel	SQL
Location	Site (1)	Site (2)
Mine List	X	X
Update	TYPE 1	TYPE 2
Period to Period Changes		X
Period to Period Employee information changes		X
Identifies new mine sites		X
Identifies deleted mine sites		X
Test Results	(3)	(4)
Exception Log		X
Distribution	6 months	Monthly
Reference Tables		
Company Type		X
Controller		X

Location Codes		X
Mine Gas		X
Mine Type		X
Operator		X
SIC List (and cross-reference)	X	X

(1) ASA Services site
(2) Cloud distribution: To-Be-Determined
(3) Limited test results (matches data provided)
(4) Complete test results

Data Sources

It should be noted from the outset that MSHA has constructed and maintained a great data set. Anyone who has worked with larger data sets knows how difficult they are to maintain particularly one that is being constantly updated. Any testing, adjustment or extension of the MSHA data to create the MLD should not be considered as some type of deficiency. The fact that so few exceptions exist and that the data set can be extended at all speak to high quality of the core MSHA data.

- MSHA Mine List:
- USPS FIPS County Codes:

Data Use

THE U.S. MINE LIST DATA (MLD) IS PROVIDED WITHOUT ANY EXPRESSED WARRANTY OR GUARANTEE AS TO ITS ACCURACY, COMPLETENESS OR FITNESS FOR ANY SPECIFIC PURPOSE OR APPLICATION WHETHER PERSONAL OR COMMERCIAL. THE VERSION OF THE MLD PROVIDED AS AN EXCEL FILE MAY NOT SOLD, BARTERED, GIVEN AWAY, COMBINED WITH OTHER PRODUCTS, OR OTHERWISE EXCHANGED IN A COMMERCIAL TRANSACTION OR AS PART OF A COMMERCIAL TRANSACTION REGARDLESS OF FORMAT. THE VERSION OF THE MLD PROVIDED AS AN EXCEL FILE MAY BE DIRECTLY EXCHANGED BETWEEN INDIVIDUALS PROVIDING THIS DOCUMENT IS ATTACHED TO THE FILE AND INFORMATION IN THE FILE IDENTIFYING THIS USE POLICY AND ORIGIN MARKS ARE NOT REMOVED OR ALTERED. LIMITATIONS ON COMMERCIAL VERSIONS OF THE MLD ARE SPECIFIED IN THE COMMERCIAL LICENSING AGREEMENTS. ENHANCEMENTS TO THE MLD WHICH ARE PART OF THE DISTRIBUTION ARE BASED ON THE ORIGINAL CONTENT OF THE DATA. SUBSEQUENT ENHANCEMENTS MADE BY ANOTHER PARTY MUST BE CLEARLY IDENTIFIED AS SUCH. DATA USE POLICY FOR THE MLD MAY CHANGE FROM TIME TO TIME AND IS ONLY APPLICABLE TO THE MLD.

Data Treatment

The MLD data is tested based on a profile of the data collected in August of 2011. The Excel version of the MLD is a subset of the commercial version. Only tests pertaining to the Excel MLD are included with the Excel MLD.*

*Note: These tests are associated with reference tables. Only the Commodity (SIC) is included with the Excel MLD.

PROCESS REPORTING

MineList_Score: Scoring for each test (Excel and Commercial versions)

MineList_Tests: Detailed exception log (Commercial version)

MineList: Records are marked where individual field exceptions have been noted. (Commercial Version)

RESOLVING DATA ISSUES

To the degree possible, the original data is preserved. Where it isn't possible, the values are 'plugged' in the MLD. The specific instances where data is 'plugged' are noted in the section on Plugged Data.

Reference tables have been constructed for each of the fields where there appears to be a fixed list of values. This excludes the natural key field (MINE_ID) and other fields such as MINE_NAME, fields with 'free form text (ACCESS_CTRL_NO, DIRECTIONS_TO_MINE and NEAREST _TOWN), numeric fields, and fields where there is a YES-NO type entry.

Using reference tables in conjunction with the MLD can resolve specific issues which are difficult to resolve with the Mine List data itself. For example; in some cases there are multiple controller names for the same controller id in the Mine List data. The Controller reference resolves these cases to a single id-name by using the most current id-name pair in the Mine List data. Therefore, when the Controller reference table is used to make queries, only a single name (the most current) will be used. This approach also has the advantage of preserving the original controller name data. This particular solution is only available in the commercial version of the MLD

Plugged Data

To the degree that it is possible, original data is not altered. There are some items which are altered and this section identifies those changes. Changes are listed in the order of the fields in the data set.

Excel and Commercial Versions:

All Fields (character type): Leading spaces are removed.

MINE_ID: The mine identifier is changed from a character type to an integer type.

CURRENT_MINE_NAME: Names are changed to have mixed case (a number of names are in all upper case).

CURRENT_CONTROLLER_NAME: Names are changed to have mixed case (a number of names are in all upper case).

CURRENT_OPERATOR_NAME: Names are changed to have mixed case (a number of names are in all upper case).

FIPS_CNTY_CD: Select FIPS county codes are not current. To match to the current 2010 FIPS codes, select codes were altered where they could be translated easily (government issued FIPS notes or errors).

#	Original County Name	State	FIPS Code		New County Name	FIPS Code
1	Dade	FL	25		Miami-Dade	86
2	Franklin	AL	959		Franklin	59
3	Yellowstone	MT	113		Yellowstone	111

Note: Exceptions are noted in each load. If these should be corrected in the future, the exceptions will disappear. Should new exceptions appear, they will be identified in the testing.

Note: FIPS Codes are padded with leading zeros to make matching to USPS FIPS codes easier (e.g. FIPS Code '3' becomes '003'. Action occurs in post processing.

LATITUDE-LONGITUDE: Datum, Source and accuracy columns have been added. Datum fields will be filled in as they are identified. Where Latitude-Longitude fields were missing and a latitude was available for the state and county, that value was plugged into the Latitude-Longitude fields. Source and accuracy fields reflect this adjustment.

Commercial Version:

CURRENT_OPERATOR_ID: Where there is a current operator name but the current operator id is missing or is set to *No_LID, the current operator name is added to the operator reference table and is assigned a unique number in a series of negative numbers. This negative number is then used as the current operator id. In future loads, if a controller name receives an id, the reference table will attempt to resolve the name to the new id.

There are cases where the operator name is missing but there is an operator id and the operator id has an associated name in the look up table. In this case, the operator name is plugged from the operator reference table.

FIPS_CNTY_CD: A number of Alaska codes could not be translated and so they were added to the reference and noted as exceptions in the reference table. Note that when the 2015 FIPS table becomes available, this will have to be revisited. These changes can be expected every 5 years with the census.

SIC CODES: Prefix (group) and suffix codes are derived from the SIC Code. These are important for hierarchal grouping so they should be as accurate as possible. A couple of exceptions were noted and adjusted in post processing:

#	SIC Name	SIC Code	Orig.Group	Orig.Suffix		Adj.Group	Adj.Suffix
1	Dimension Granite	141101	1411	00		1411	01
2	Crushed, Broken Basalt	142905	1499	05		1429	05
3							

Testing Approach

It should be noted that the Excel version of the MLD is a subset of the Commercial version of the MLD and so the tests described apply to both versions of the data. Only the SIC refrence data is provided with the Excel MLD.

REFERENCE TABLES

The reference tables are used to both test data and to create data queries. With only a few exceptions, each field in the Mine List table has a limited range of values. If the field has characters, unique values from that field are compiled into a 'reference' table. In select cases (e.g. FIPS codes), a complete data set is pulled in from an external source and is incorporated as a reference table. All reference tables have a 'Mine Reference' (ML) prefix (e.g. ML_Status, ML_Controller and ML_Canvass).

Testing: When a new set of data is retrieved, it is tested before it is added to the database. Part of that validation process is to check the entries for each field against its corresponding reference table and identify exceptions.

Queries: Reference tables often provide information about how to construct queries. For example: If you ask the database to count the number of 'abandoned' mines specifically (CURRENT_MINE_STATUS) you would miss the 'abandoned and sealed' mines. Since the reference table includes both versions, you would be less likely to miss both variants.

There are three basic types of reference tables:

1) Fixed: The list has a fixed length and set of members. Changes (additions, deletions, or changes) are not anticipated for this type of list (e.g. ML_Status, ML_MineType and ML_Canvass)
2) Expandable: These references expand as new data sets are periodically added (e.g. ML_Controller and ML_Operator)
3) Slowly Changing: Some references may change slowly over time. At present, no references have been considered to be in this category type but some Expandable types may be candidates for this treatment in the future (e.g. CONG_DIST_CD).

Reference Table Data Sources: Sources include the MSHA50 Handbook (e.g. ML_SIC_Code and ML_Canvass), the data itself (distinct field values – e.g. ML_MineStatus), and external data such as county latitude-longitude coordinates for counties (ML_Location) from the USPS. It should be noted that some MSHA50 Handbook lists differ from similar lists derived from the actual data.

TESTING SCORES:

The testing results for a specific load are compiled in a 'scoring' table. The tests are run on the data after it has been 'pre-processed'. The tests are applied to the raw data. They don't reflect the results after any post-processing treatments (e.g. missing operator names are loaded from the operator reference if the operator name isn't empty and there is an existing/valid operator id). If there are exceptions the exception count is posted as the score (> 0) for the test. If the score is set to 'P' it indicates the test passed without any exceptions identified.

The exceptions are logged individually so that they can be checked at a later time. In select cases, one entry is made for a number of exceptions (e.g. Outdated FIPS county codes).

Data tests were constructed based on the profiling a complete set of data from August 2011 (82K records). Based on that profile, several exception types were identified. Exceptions generated as the result of testing are counted by field, logged by specific instance (where practical), and the records generating the exception are marked (per the individual test rules). Exception types are as follows:

(T0) UNIQUE: For each imported data set, the MINE_ID is unique for each mine site. This is the only unique identifier in imported data sets. It should be noted that in the destination table for the records, the REF_ID+MINE_LIST are the unique identifier pair for the MineList table.

(T1) MISSING: If a field is NULL or EMPTY (blank) but there is an expectation that a value should be present, it is given this designation (e.g. MINE_STATUS should not be empty).

(T2) UNEXPECTED: This exception type is primarily used when fields in the table are interdependent and an issue arises (e.g. There is an expectation that the FIPS Code and Name Match the FIPS Code and Name in the reference table). See the individual rule notes where this exception type is applied.

(T3) NIL (NOT-IN-LIST): If a field does not match a value in a reference table, an exception is generated. In some cases (e.g. CURRENT_CONTROLLER_ID) the field's data will be added dynamically to the Controller reference table.

(T4) INVALID: If a field is invalid (e.g. Date: 2-30-2016), an exception is generated.

(T5)IDENTITY: When both an Id and corresponding name are in the same table, this can produce issues (e.g. Controller ID and Controller name) There are examples where one Controller ID has 2 or more different names. This is an identity problem (2NF issue) and, if identified, an exception is generated.

(T6)VALUE: If a value (numeric or string) is out of range, an exception is generated.

(T7) CALC: If there is a calculated field (numeric or string) where the calculation is incorrect, an exception is generated.

TEST DRIVEN ACTIONS:

FATAL (F): There are relatively few tests that will cause a load to fail completely. These cases are noted in the chart and test descriptions.

UPDATE (U): Values in the staging table appear to be more recent than matching values in the reference table. These drive dynamic changes in select tables (e.g. Operator ids). Automatic pruning of temporary ids is part of this operation.