

Virginia Tech

VT Exposure Database Guidelines

1 Contents

2	Overview.....	3
3	Requirements	3
4	Exposure Database Metadata Tables.....	4
4.1	VALID_VALUES Table Structure	4
4.2	TABLE_METADATA Table Structure.....	6
4.3	COLUMN_METADATA Table Structure.....	6
4.4	Multiple Schemas	7
5	Demographic Log Structure.....	7
5.1	Master-Detail Terminology	8
5.2	Reduced Demographic Log Table	8
5.3	Demographic Log Translation.....	8

Revision History

Version #	Date	Entered by	Comments
1.0	10/20/2011	Aaron Schroeder	Initial Documentation
2.0	4/27/2012	Austin Mills	Updated Documentation and added required tables
2.1	4/27/2012	Aaron Schroeder	Clarifications and Formatting
2.2	5/11/2012	Austin Mills	Valid Values Table section updated
3.0	5/29/2012	Austin Mills	Added Demographic Log Structure sections. Updated Valid Values Table section for range clarifications. Removed notes regarding Lexicon views mapping to more than one Exposure DB table. Removed foreign key requirement and updated master-detail section.
3.1	6/18/2012	Austin Mills	Added note regarding "VLDS_NULL" value in the Valid Values section. Updated Metadata management descriptions and removed details on Metadata structure after design decision on LMT.
3.2	7/19/2012	Austin Mills	Revised Demographic Log structure for SCHEV and DOE

			and added the updated diagrams. Added VEC Demographic Log structure section and diagram.
3.3	11/5/2012	Austin Mills	Added requirement for the VLDS_ID_MAPPER and VLDS_PERFORMANCE_JOIN tables to be in DATAADAPTER schema.
3.4	2/14/2013	Austin Mills	Added new Exposure Database requirements and design information about Demographic Log Reduction
3.5	4/17/2013	Austin Mills	Added requirement for including primary keys on each detail table. Removed use of VLDS_ID_MAPPER and VLDS_PERFORMANCE_JOIN tables.
3.6	4/23/2013	Austin Mills	Added requirement for only alphanumeric characters in SSN or any other Internal_ID or Match_ID field.

2 Overview

The Exposure Database contains agency data records in tables which they wish to expose to researchers and for joining with other agencies. One responsibility of each agency participating in VLDS will be to create a database to hold the data to be used by the VLDS system. This exposure database should reside behind the agency's existing security measures along with the Data Adapter.

3 Requirements




- REQ-1 The Exposure-DB must exist behind the agency's secured firewall.
- REQ-2 The Exposure-DB should support SQL-92 data types as much as possible.
- REQ-3 Data exposed must be contained in tables, versus views, for compatibility with the Data Adapter.
- REQ-4 DataAdapter must have database privileges allowing for SELECT on all tables.
- REQ-5 Lexicon Views must have a 1-to-1 relationship to the Exposure Database tables.
- REQ-6 Valid Values that can have a Null value must be entered as "VLDS_NULL" in the Valid Values table.
- REQ-7 Demographic log reduction and hashing in the Data Adapter will use a reduced log table and temporary processing tables stored in the DATAADAPTER schema
- REQ-8 Appending new records to agency detail tables should not be done while a data package is being processed that involves those tables.
- REQ-9 Adding new columns to a detail table should not be done before going through the proper governance procedures.
- REQ-10 Exposure database detail tables must have a primary key of VLDS_DETAIL_ID with a data type of BIGINT on SQL Server and NUMBER (20) on Oracle.
- REQ-11 All SSNs or any unique identifier used in the Internal_ID or Match_ID columns in the Demographic Log must only include alpha-numeric and cannot include hyphens or spaces.

4 Exposure Database Metadata Tables

Each participating agency can have the option to create tables inside their exposure database to hold Lexicon Metadata values. The LMT (Lexicon Metadata Tool) is built to manage these tables in the Lexicon, but for convenience the Metadata and Valid Values can be stored in the Exposure Database to be later imported into the LMT. These tables are not used as the permanent location for Metadata and Valid Values. The permanent location is in the Lexicon database. The Exposure Database can be used as the backup location.

4.1 VALID_VALUES Table Structure

This table is used to store Valid Values for a data element in the exposure database. Once imported into the Lexicon using the Lexicon Metadata Tool, it can then be sent to the DRT (Data Request Tool) and Lexicon Report. When researchers are composing a query using the DRT they provide a WHERE clause with filters. The researcher can then choose a value from the Valid Values of a filter element.

Column name	Data type	Description
 TABLE_NAME	VARCHAR(100)	This is the name of a table in the Exposure-DB.
 COLUMN_NAME	VARCHAR(100)	
 VALUE	VARCHAR(500)	Dates must be formatted as 'yyyy-mm-dd'.
DESCRIPTION	VARCHAR(2000)	This will be displayed to the researcher along with the value. It will not be displayed in the query result set.
VALID_USE_BEGIN_DATE	DATE	Required
VALID_USE_END_DATE	DATE	VLDS_NULL allowed if still valid.
LAST_UPDATE	TIMESTAMP (as defined in SQL-92 standard, or closest data type)	For auditing purposes only.

Valid Value Notes:

- Integer and decimal ranges can be specified for column data. A column with an integer or decimal range will have one record with the start of the range and one record with the end of the range. For example, we specify a range of 0..45, 50, 55..100 by splitting the begin and end numbers of the range and

designate them to a single row. The first record will contain the value '0' and the next record will contain the value '45'. For more examples please see the example table below.

- The Description field is a description of a *Value*. Note that COLUMN_METADATA.MS_DESCRIPTION is the proper data element for identifying a *Column's* description to the Lexicon. When a valid value or a range of Valid Values are entered, a thorough description must be entered.
- Dates entered as Valid Values can also be specified as a range. The format used for dates should be 'yyyy-mm-dd'. For example '2012-05-15' is a proper date format. Fields that can contain any kind of free form text should not have Valid Values listed in the table. It should only be explained in the column Metadata table with a detailed description. The same can be applied to strings with a specific format.
- If Valid Values have the ability to contain a "Null" value, it should be specified by using the text string "VLDS_NULL." The Lexicon will expose VLDS_NULL values as database nulls in both the Lexicon Views and web service. This should not be placed in any other column or table except the Valid Values table as a possible value. Exposed data sets should not have this value entered in the tables.

Example VALID_VALUES data:


TABLE_NAME	COLUMN_NAME	VALUE	DESCRIPTION
Building	RoomNum	100	VLDS_RANGE_BEGIN. Beginning of the range of rooms 100-120 in the first floor of the building.
Building	RoomNum	120	VLDS_RANGE_END. Ending room number for the range 100-120.
Building	RoomNum	125	First floor auditorium.
Building	RoomNum	130	VLDS_RANGE_BEGIN. Beginning of the range of rooms 130-150 in the building.
Building	RoomNum	150	VLDS_RANGE_END. Ending room number for the range 130-150.
Building	RoomNum	VLDS_NULL	Null value
Person	Gender	M	Male
Person	Gender	F	Female
Course	Semesters_Offered	01/15/2012	VLDS_RANGE_BEGIN. Start of the spring semester date range.
Course	Semesters_Offered	05/15/2012	VLDS_RANGE_END. End of the spring semester date range. Course not offered in summer.
Course	Semesters_Offered	08/15/2012	VLDS_RANGE_BEGIN. Start of the fall semester date range. Course not

			offered in summer.
Course	Semesters_Offered	12/10/2012	VLDS_RANGE_END. End of the fall semester date range.

4.2 TABLE_METADATA Table Structure

This table is used to store Metadata for an agency exposure database table. It can be stored outside the exposure database. Once imported to the Lexicon using the Lexicon Metadata Tool, it can then be sent to the Data Request Tool and Lexicon Report.



Note that Views in the Lexicon can only have a 1-to-1 relationship to Exposure-DB tables.

Column name	Data type	Description
 TABLE_NAME	VARCHAR(45)	This is the name of a table in the Exposure-DB.
FRIENDLY_NAME	VARCHAR(45)	
MS_DESCRIPTION	VARCHAR(2000)	
CRITICAL_CHANGES	VARCHAR(4000)	
LAST_UPDATE	TIMESTAMP (as defined in SQL-92 standard, or closest data type)	For auditing purposes only.

4.3 COLUMN_METADATA Table Structure

This table is used to store Metadata for columns in an agency exposure database. It can be stored outside the exposure database. Once imported to the Lexicon using the Lexicon Metadata Tool, it can then be sent to the Data Request Tool and Lexicon Report.

Note that Views in the Lexicon can only have a 1-to-1 relationship to Exposure-DB tables.

Column name	Data type	Description
 TABLE_NAME	VARCHAR(45)	This is the name of a table in the Exposure-DB.
 COLUMN_NAME	VARCHAR(45)	
FRIENDLY_NAME	VARCHAR(45)	Friendly name used to describe the column
MS_DESCRIPTION	VARCHAR(2000)	Full description of the column
CRITICAL_CHANGES	VARCHAR(4000)	

Column name	Data type	Description
DATA_DOMAIN	VARCHAR(200)	Comma separated
JOIN_ONLY	CHAR(5)	true/false
VALID_USE_BEGIN_DATE	DATE	Required
VALID_USE_END_DATE	DATE	NULL allowed if still valid.
LAST_UPDATE	TIMESTAMP (as defined in SQL-92 standard, or closest data type)	For auditing purposes only.

4.4 Multiple Schemas

It is possible for the Exposure-DB to contain multiple schemas. In that case, each schema is expected to contain its own VALID_VALUES, VIEW_METADATA, and COLUMN_METADATA table.

5 Demographic Log Structure

Each agency is required to create a demographic log table in their exposure database for their records. Every time an entry is recorded for a person, whether is a duplicate or not, the demographic log criteria should be entered. The Demographic Log is an essential part of the identity resolution and matching process used in the VLDS. The Demographic Log structure for all reporting agencies should be identical and is shown below.

Education Agency Demographic Log

Column Name	Data Type	Description
VLDS_PK	NUMERIC(9)	Primary Key
INTERNAL_ID	VARCHAR(50)	Agency Internal ID used for the agency records (SSN, STI, etc)
FIRST_NAME	VARCHAR(50)	First Name
MIDDLE_NAMES	VARCHAR(50)	Middle Names
LAST_NAME	VARCHAR(50)	Last Name
GENDER	CHAR(1)	Gender. M for male, F for female
DOB_MONTH	CHAR(2)	2 digit birth month
DOB_DAY	CHAR(2)	2 digit birth day
DOB_YEAR	CHAR(4)	4 digit birth year
FIPS_STATE	CHAR(2)	2-digit FIPS/ANSI State or State-Equivalent Code as specified in INCITS 38:2009 (formerly FIPS PUB 5-2). Virginia is 51.
FIPS_COUNTY	CHAR(3)	3-digit FIPS/ANSI County or County-Equivalent Code as specified in INCITS 31:2009 (formerly FIPS PUB 6-4).
ZIP_5	CHAR(5)	

	DATE	Date information recorded. If not available, then date representing unique recording time-period for individual (e.g. school semester start date).
RECORD_DATE		
MATCH_ID_1	VARCHAR(50)	Other ID available to match with other agencies
MATCH_ID_2	VARCHAR(50)	Other ID available to match with other agencies
MATCH_ID_3	VARCHAR(50)	Other ID available to match with other agencies

Note: SSN or any unique identifier must be alpha-numeric only and cannot include any hyphens or spaces in the INTERNAL_ID or MATCH_ID fields.

5.1 Master-Detail Terminology

The master detail structure is not applicable in the exposure database, just used to help describe the relationship between the demographic log and detail tables. In a master detail structure, there are foreign keys for the one and only master record, but the demographic logs may have more than one record so therefore it doesn't apply. The demographic log is loosely defined as a master table when theoretically it is not.

5.2 Reduced Demographic Log Table

Each agency in the VLDS will also contain a Reduced Demographic Log table. This table is a duplicate of the Demographic Log and is used to store records that have been reduced in the Log Reduction process of the DataAdapter. This table will be used in the Identifiers and Performance queries during the execution of a data request. The Reduced Demographic Log should exist in the DATAADAPTER schema of the exposure database.

The Virginia Tech team contains the scripts needed to create these tables and the schema for Oracle and SQL Server database types.

5.3 Demographic Log Translation

The left side of the diagram below in red represents the columns used to identify a person in a specific agency detail table such as Student Record. The orange column names to the right represent the Demographic Log columns needed from that agency to help identify a person. The arrows represent the column translations from the agency tables to the Demographic Log. The INTERNAL_ID column in the detail table should have the column name "INTERNAL_ID" even if it represents a State Test Identifier (STI).

STUDENT RECORD TABLE EXAMPLE

EXPOSURE DATABASE DEMOGRAPHIC LOG

