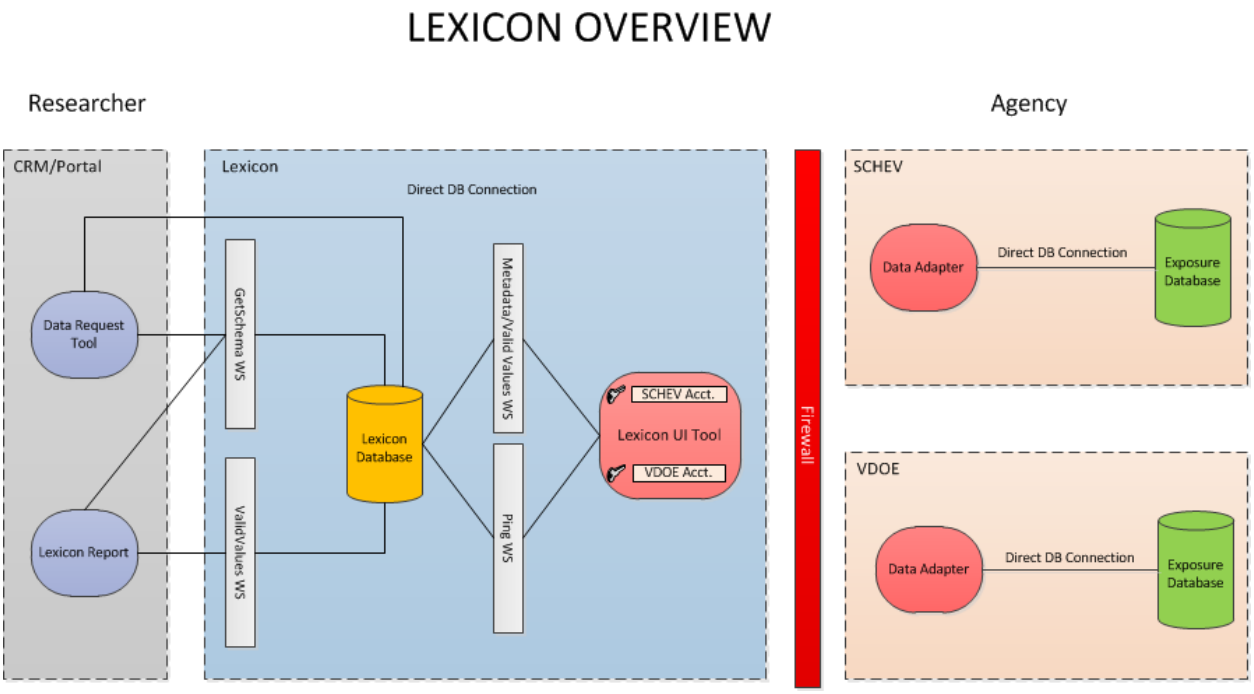
VT-VLDS Lexicon Specifications

Table of Contents

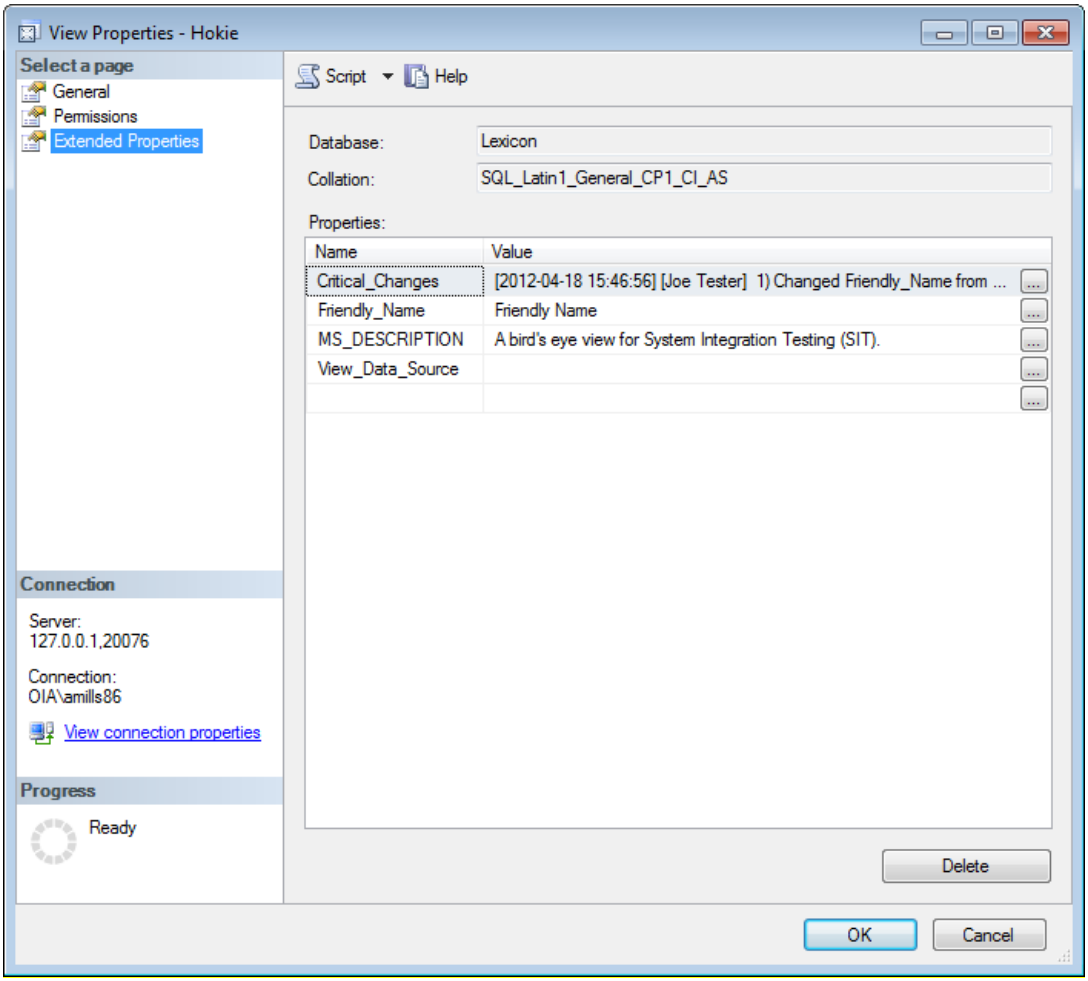
Overview

This document provides specifications for the Lexicon. The Lexicon is hosted outside of the agency firewalls and is used as a Metadata repository for queries to be built against. A Lexicon Metadata Tool exists on top of the Lexicon to provide Metadata and Valid Value management. For more information please refer to the Lexicon Metadata Tool Design document. Below is an overview of the Lexicon and how it integrates with other components.



Extended Properties

The extended properties feature in SQL Server is used to hold the advanced Metadata about a participant's data. For clarity, the information held in the extended properties should be referred to as Lexicon Metadata. An example of extended properties for a view is shown below.



Demographic Log Structure

Each agency is required to create a demographic log table in their exposure database for their records. The agency's demographic log and the associated Metadata will also be stored in the Lexicon. 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.

Master-Detail Terminology

The master detail structure is not applicable in the exposure database or Lexicon, 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.

Lexicon Agency Metadata Tables

Each participating agency will have tables inside the Lexicon database to hold Metadata values. The structure is built around database views which are identical to the agency's exposed tables. For convenience, the Metadata and Valid Values are referring to the agency views in the Lexicon, not their exposed tables. A web interface (Lexicon Metadata Tool) will be built alongside the Lexicon Database for an agency to use in adding/modifying/deleting Metadata and Valid Values.

VIEW_METADATA

The VIEW_METADATA table is used to store Metadata for agency view. The view in the Lexicon is identical to the table in the exposure database. The data can be managed using the LMT manually or with the import tool. The data permanently resides in the Lexicon which in turn provides it to the DRT (Data Request Tool) and Lexicon Report. The structure is listed below.

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

Column name	Data type	Description
VIEW_NAME	VARCHAR(100)	This is the name of a table in the Exposure-DB.
FRIENDLY_NAME	VARCHAR(500)	
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.

COLUMN_METADATA

The COLUMN_METADATA table is used to store metadata for agency exposure database columns in a table. This data can be managed using the LMT manually or with the import tool. This data permanently resides in the Lexicon which in turn provides it to the DRT (Data Request Tool) and Lexicon Report. The structure is listed below.

Note that Views in the Lexicon might contain columns from multiple Exposure-DB tables. Therefore the individual setting up the Metadata for this table will need to know that relationship in order to fill in the VIEW_NAME.

Column name	Data type	Description
TABLE_NAME	VARCHAR(100)	This is the name of a table in the Exposure-DB.
COLUMN_NAME	VARCHAR(100)	
FRIENDLY_NAME	VARCHAR(500)	
MS_DESCRIPTION	VARCHAR(2000)	
CRITICAL_CHANGES	VARCHAR(4000)	
DATA_DOMAIN	VARCHAR(200)	Comma separated
JOIN_ONLY	CHAR(5)	true/false
VALID_USE_BEGIN_DATE	DATE	Required
VALID_USE_END_DATE	DATE	
LAST_UPDATE	TIMESTAMP (as defined in SQL-92 standard, or closest data type)	For auditing purposes only.

VALID_VALUES

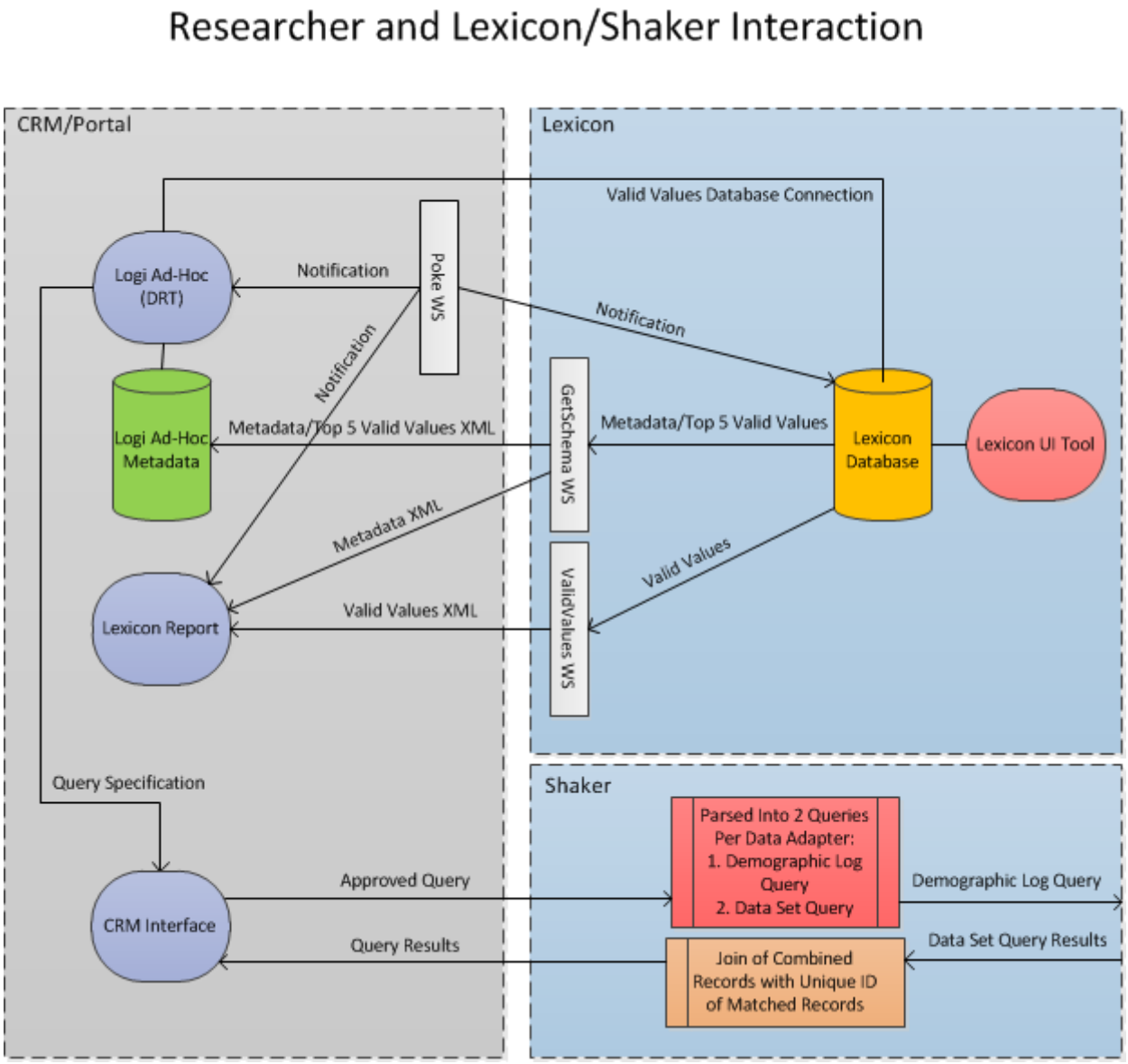
The purpose of this table is to provide Valid Values for specific column in a view. This data is provided to the current DRT (Data Request Tool) by way of a direct database connection and the top 5 Valid Values are sent via the GetSchema web service. When researches are composing a query using the DRT they provide a WHERE clause with filters. The researcher can choose a value from the Valid Values of a filter element. Researchers using the Lexicon Report can also view Valid Values for a column and those are sent using the ValidValues web service. The structure is listed below.

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	
LAST_UPDATE	TIMESTAMP (as defined in SQL-92 standard, or closest data type)	For auditing purposes only.

Integration Point: DRT

Data Request Tool (DRT) is a tool on the portal that allows researchers to select, request, and receive data. The DRT is built by leveraging the Logi Ad-hoc tool. When a change is done to the Lexicon Database, a web service called GetSchema is used to manipulate those changes into XML to be sent to the DRT. The GetSchema web service sends only the database structure and top 5 Valid Values. There is another web service interface between Lexicon database and DRT that will notify the DRT when changes have been made to Lexicon and is referred to as the Poke Web Service. After the Poke web service is initiated, the DRT makes a call to the GetSchema web service to retrieve structure and top 5 Valid Values. There is also a direct connection the DRT has with the Lexicon Database which is used to populate Valid Values when needed. The diagram below helps illustrate the interactions.



Integration Point: Lexicon Report

The Lexicon Report is a tool on the portal used for extracting Metadata as well as Valid Values from the Lexicon Database and submitting a requested user data agreement (RUDA). The Lexicon report makes a web service call to the GetSchema web service on the initial load of the report to populate the database structure from the Lexicon Database and extracts Valid Values through the ValidValues web service. The diagram above helps illustrate the interactions.

Integration Point: Shaker

The Shaker is the point of access for secure, de-identified data extraction from SLDS-connected data sources. The Shaker's general function is to accept an approved query and return a dataset. The query will be broken down into a series of optimized steps, or sub-queries, to retrieve de-identified data from the appropriate data sources in the most efficient manner. In keeping with the intent of the original, query forms (e.g. inner join, left join, equijoin) and specified final output parameters (e.g. counts of non-matching records by demographic categories) will be taken into consideration.

For each query submitted to the Shaker, a random key is generated. Each sub-query in the data retrieval plan will send this random key to the data source to be used in creating a secure one-way hashed key for any applicable records. This list of hashed keys is then used by the Shaker to combine records across multiple data sources, never transferring any identifiable information out of the data source. Any hashed keys used to link records will be removed from the final data set and replaced with yet another random key which cannot be traced back to any original data sources. The resulting combined records are then uploaded to a large file storage system for later access by the user.

Information from the Lexicon concerning data structure and relationships will be used to produce a dynamic sub-query plan for data retrieval that minimizes processing time and workload on the data sources.

Integration Point: Lexicon Metadata Tool (LMT)

The Lexicon Metadata Tool is built in the same environment as the Lexicon Database and allows agencies to make Metadata changes securely. This creates a way for agencies to easily update view or column Metadata as well as Valid Values. The tool will also provide secure access to view Lexicon connections, logs, and statistics for the appropriate admin.

For more information see the Lexicon Metadata Tool Design Document