

# Vocabulary Data Sets Source Documentation

- [Value Set Sources](#)
  - [voc.xml](#)
  - [VSAC](#)
  - [PHIN VADS](#)
  - [Other](#)
- [Loading of Code Systems and Value Sets](#)

Multiple value sets and code systems are used with the Cancer Report Validator. Many of them are made [available](#) from the development team as a gzip compressed tar archive. Some of the value sets, however, must be obtained by the end user. Instructions for obtaining these value sets can be found [here](#). Value sets that are not included in the tar archive are highlighted in yellow.

## Value Set Sources

The value sets were gathered from a variety of sources. A number of them can be found in [voc.xml](#), [VSAC](#), [PHIN VADS](#), or various sources listed in the "Source(s)" column in the following table: [Vocabulary Data Set Sources](#).

### **voc.xml**

For value sets where the "File Name" column specifies voc.xml, the value set can be found [here](#). The voc.xml file was based on the one provided in the [HL 7 CDA® R2 IG: Reporting to Public Health Cancer Registries from Ambulatory Healthcare Providers, R1, DSTU Release 1.1 - US Realm package](#), which can be downloaded from [this page](#). The file was modified to remove value sets that were incomplete or identified as DYNAMIC, according to the Cancer IG. Modifications made to voc.xml include the addition of value sets that were missing and identified as STATIC according to the Cancer IG, and updates to existing value sets to include the correct codes based on sources listed in the "Source(s)" column.

### **VSAC**

A [Unified Medical Language System](#) (UMLS) account is required in order to view and download value sets from VSAC. Value sets that were obtained from VSAC in Excel format have "VSAC" listed in the "Source(s)" column in the table above.

### **PHIN VADS**

A number of complete value sets were downloaded from PHIN VADS as a text document. Although many of these value sets could be downloaded as-is, modifications had to be made to some of the downloaded text documents containing the value sets so that they could be loaded by the validator properly. A description of these modifications can be found in the "Instructions for obtaining value set" column.

### **Other**

A number of value sets could not be downloaded from VSAC or PHIN VADS because the value set did not exist in either system or because the value set present was incomplete. For these value sets, files were created in the PHIN VADS text document format containing at a minimum: codes, display names, value set OID, value set name, code system OID, and code system name. This information was obtained from the sources listed in the "Source(s)" column of the table and any instructions for finding the codes for the value sets can be found in the "Instructions for obtaining value set" column.

## Loading of Code Systems and Value Sets

The Code Validator project loads the ICD-9-CM, ICD-10-CM, ICD-10PCS, LOINC, RXNORM, and SNOMED CT code systems, as well the PHIN VADS and VSAC value sets, into OrientDB databases. Loaders for code systems other than the ones mentioned (e.g. ICD-O-3) are contained by the CRV project.

The files referenced in the [Vocabulary Data Sets Source Documentation](#) for ICD-9-CM, ICD-10-CM, ICD-10PCS do not have any delimiters within the code. As part of the loading of the contents of the file into the OrientDB databases prior to use, periods are added after the third digit (from the left) of a code, for codes with length > 3 digits.