U.S. Department of Veterans Affairs
Standards & Interoperability (S&I)

Blanket Purchase Agreement 54 (VA776-11-BP-0054)
Task Order 2 (VA776-C20249)

August 6, 2013

Task 7D: Report documenting the import of individual terminologies from the Unified Medical Language System (UMLS) into the International Health Terminology Standards Development Organisation's (IHTSDO) Terminology Workbench

Contract: GS35F0009L

Apelon

Apelon Inc

100 Danbury Rd
Suite 202
Ridgefield, CT

p 203.431.2530
www.apelon.com

Table of Contents

1. Overview

This document describes the process that was created to transform individual terminologies from the Unified Medical Language System (UMLS) Rich Release Format (RRF) into a format suitable for the International Health Terminology Standards Development Organisation (IHTSDO) Workbench.

2. Unified Medical Language System

2.1. A brief overview of the UMLS

The Unified Medical Language System is a compendium of many controlled vocabularies in the biomedical sciences, initially created in 1986. It provides a mapping structure among these vocabularies and thus allows one to translate among the various terminology systems; it may also be viewed as a comprehensive thesaurus and ontology of biomedical concepts.

The UMLS was designed and is maintained by the US National Library of Medicine (NLM), and is updated quarterly. The project was initiated in 1986 by Donald A. B. Lindberg, M.D., Director of the Library of Medicine.

The current UMLS release (2013AA) includes over 150 classifications, code sets, thesauri and lists of controlled terms from the biomedical domain. The distribution is nearly 4GB while still compressed.

UMLS distribution includes tooling (MetamorphoSys) which is used to extract and subset the UMLS into the Rich Release Format [1]. The RRF output files are the starting point for loading the UMLS content into the Workbench.

3. Loading UMLS Content into the IHTSDO Workbench

3.1. Obtaining the UMLS content

For other terminologies converted into the Workbench format for the VA, Apelon has hosted a copy of the native source format within the VA Archiva server. However, because of the size of the UMLS and the required MetamorphoSys conversion process, Apelon has not hosted a copy of the UMLS in Archiva.

Developers wishing to convert UMLS content will have to download the UMLS and run MetamorphoSys as a prerequisite to converting the content to the Workbench format.

The UMLS content can be downloaded from:

http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html

3.2. Running MetamorphoSys

[1] http://www.ncbi.nlm.nih.gov/books/NBK9685/

After downloading the UMLS content and extracting MetamorphoSys (mmsys.zip) the next step is to run the UMLS MetamorphoSys to subset the content. This will create the RRF source files, which are required by the Workbench Converter.

The operation of MetamorphoSys will not be covered in depth in this document. Rather, a guide to the required settings for the Workbench converting tooling will be provided. The full documentation for MetamorphoSys tooling can be found on the NLM website:

http://www.nlm.nih.gov/research/umls/implementation_resources/metamorphosys/

After launching MetamorphoSys, select "Install UMLS". After providing the "Output Destination", select the "New Configuration" button. A panel with a number of tabs is displayed.

3.2.1. Output options

- The "Output Format" must be configured to "Rich Release Format" (which is the default setting)
- The "Source Abbreviation Format" can be set to either "versioned" or "unversioned", depending on which style of Source Attributes are desired in the Workbench.
- The "Browser Index Files" are not required
- The "Database Load Scripts" are not required

3.2.2. Source list

Any number of sources can be selected for conversion into the Workbench. The Workbench conversion tool also allows users to further restrict which source files included in a Workbench Data file.

For the best performance, only export the desired terminologies from MetamorphoSys. The currently published Workbench UMLS data file contains:

- ICD-9-CM – International Classification of Diseases, Ninth Revision, Clinical Modification
- ICD-10-CM – International Classification of Diseases, 10th Edition, Clinical Modification
- HCPCS – Healthcare Common Procedure Coding System
- MTHHH – Metathesaurus HCPCS Hierarchical Terms
- CPT – Current Procedural Terminology
- MTHCH – Metathesaurus CPT Hierarchical Terms

The following terminologies are also included in the MetamorphoSys transform. These have special handling – they are not imported into the Workbench directly – but relationships are extracted from these terminologies.

- SNOMED CT – SNOMED Clinical Terms
- SCTUSX – SNOMED Clinical Terms US Extension
- MTH - Metathesaurus

3.2.3. Precedence

The Workbench conversion tool will follow the Precedence specified here when selecting the term description type to use for the Fully Specified Name (FSN) and Preferred Synonym.

3.2.4. Complete subset

Once configured, launch the Subset process within MetamorphoSys. This may take a significant amount of time to run, depending on subset choices and hardware.

3.3. Setting up and running the Workbench UMLS Loader

3.3.1. Getting the transform tools

The Workbench conversion tooling developed to load the RRF output of MetamorphoSys is hosted here:

https://csfe.aceworkspace.net/svn/repos/va-oia-terminology-converters/UMLS/trunk/

The UMLS conversion tooling also depends on a library which is shared with the RxNorm Workbench Data loader (which also processes RRF formatted data files). This shared library is hosted here:

https://csfe.aceworkspace.net/svn/repos/va-oia-terminology-converters/umlsLoaderUtilities/trunk

3.3.2. Using the transform tools

This section documents the process of converting the RRF output from MetamorphoSys into the Workbench format. This document assumes that a working Maven environment on the system. Installation and configuration of the Maven environment for use with the Workbench is not discussed in this document.

3.3.3. Converting the source content

1. Build and install the `umlsLoaderUtilities` project into the local Maven environment. No special configuration is required for this project.

2. Configure the converter by editing the file `UMLS\UMLS-econcept\pom.xml`. The following components of the `buildUMLS` "execution" can be customized:
    a. `<srcDataPath>` - Required – This needs to point to the path where MetamorphoSys was configured to place the output. This path should end with a folder that corresponds to the UMLS version number. For example: `C:\temp\UMLS\output\2013AA\`
    b. `<tmpDBPath>` - Optional – This variable may be used to specify the path where temporary data will be written during the conversion. This is provided because the temporary data can be quite large, and it may be necessary to place it on a different disk drive, depending on your hardware.

c. `<sabFilters>` - Optional – This variable may be used to further subset the content that is converted into the Workbench format. If this variable is omitted, all terminologies that were output by MetamorphoSys will be converted. When this parameter is provided, only terminologies that match the filter will be converted.
   i. Note – the SAB identifiers to use here must match the selection made during the MetamorphoSys process – either "versioned" or "unversioned" SABs.
   ii. Note – The SAB "SRC" should always be included (if a sabFilter is provided) so that Hierarchies can be constructed properly during the conversion.

d. `<additionalRootConcepts>` - Optional – This parameter can be used to specify additional root concepts that should be treated as the roots of a Hierarchy within the Workbench.

3. Additionally, while editing the file `UMLS\UMLS-econcept\pom.xml` the appropriate value for `<version>` should be specified to match the version of the UMLS that was extracted.

4. Execute the command:
   `mvn clean install` from within the `UMLS` folder.
   This command will take a significant amount of time to run, depending on the number of terminologies included in the conversion. For very large conversions, it may be necessary to increase the Java Heap memory settings for the Maven installation.

5. When the conversion process is finished, the resulting file `RRF-MR.jbin` will be located in the folder `UMLS\UMLS-econcept\target\`. This folder contains both a jbin file, and a zip file (which also contains the jbin file).

6. After testing the output file locally in a build of the Workbench, the `UMLS\UMLS-econcept` project can be published to the VA Archiva server by executing a `mvn deploy` command. This will upload the zip file containing the `RRF-MR.jbin` file and some conversion metadata and statistics. The specific deploy command required will depend on the Maven configuration.

4. Including the Output into a Workbench Bundle

The converted UMLS content can be included in the Workbench by building it into the Workbench database. The database project (baseline) pom file will need the following section to declare the dependency and extract the zip file:

```
<plugin>
  <groupId>org.apache.maven.plugins</groupId>
  <artifactId>maven-dependency-plugin</artifactId>
  <executions>
    <execution>
      <id>fetch-UMLS-data</id>
      <phase>generate-sources</phase>
      <goals>
        <goal>unpack</goal>
      </goals>
      <configuration>
        <artifactItems>
          <artifactItem>
            <groupId>gov.va.oia.terminology.converters</groupId>
            <artifactId>UMLS-econcept</artifactId>
            <version>2013AA-loader-1.0</version>
            <type>zip</type>
         </artifactItem>
        </artifactItems>
        <outputDirectory>${project.build.directory}/db</outputDirectory>
      </configuration>
    </execution>
  </executions>
</plugin>
```

Next, use the `load-econcepts-multi` execution, and add in the `RRF-MR.jbin` file to the `conceptsFileNames` section.

```
<plugin>
  <groupId>org.ihtsdo</groupId>
  <artifactId>wb-bdb-mojo</artifactId>
  <executions>
    <execution>
      <id>load-econcepts</id>
      <phase>process-sources</phase>
      <goals>
        <goal>load-econcepts-multi</goal>
      </goals>
      <configuration>
        ...
       <conceptsFileNames>
         <String>eConcepts.jbin</String> <!-- snomed -->
         <String>RRF-MR.jbin</String>
       </conceptsFileNames>
       ...
```

Finally, because a new path is created for the `UMLS` content, configure the Workbench to include the path as an origin for the current path. The FSN of the path concept is `‘UMLS Path’`. One way to configure this path during Workbench assembly is to modify the file:

```
[workbenchAssemblyProject]\database\project-bdb\pom.xml
```

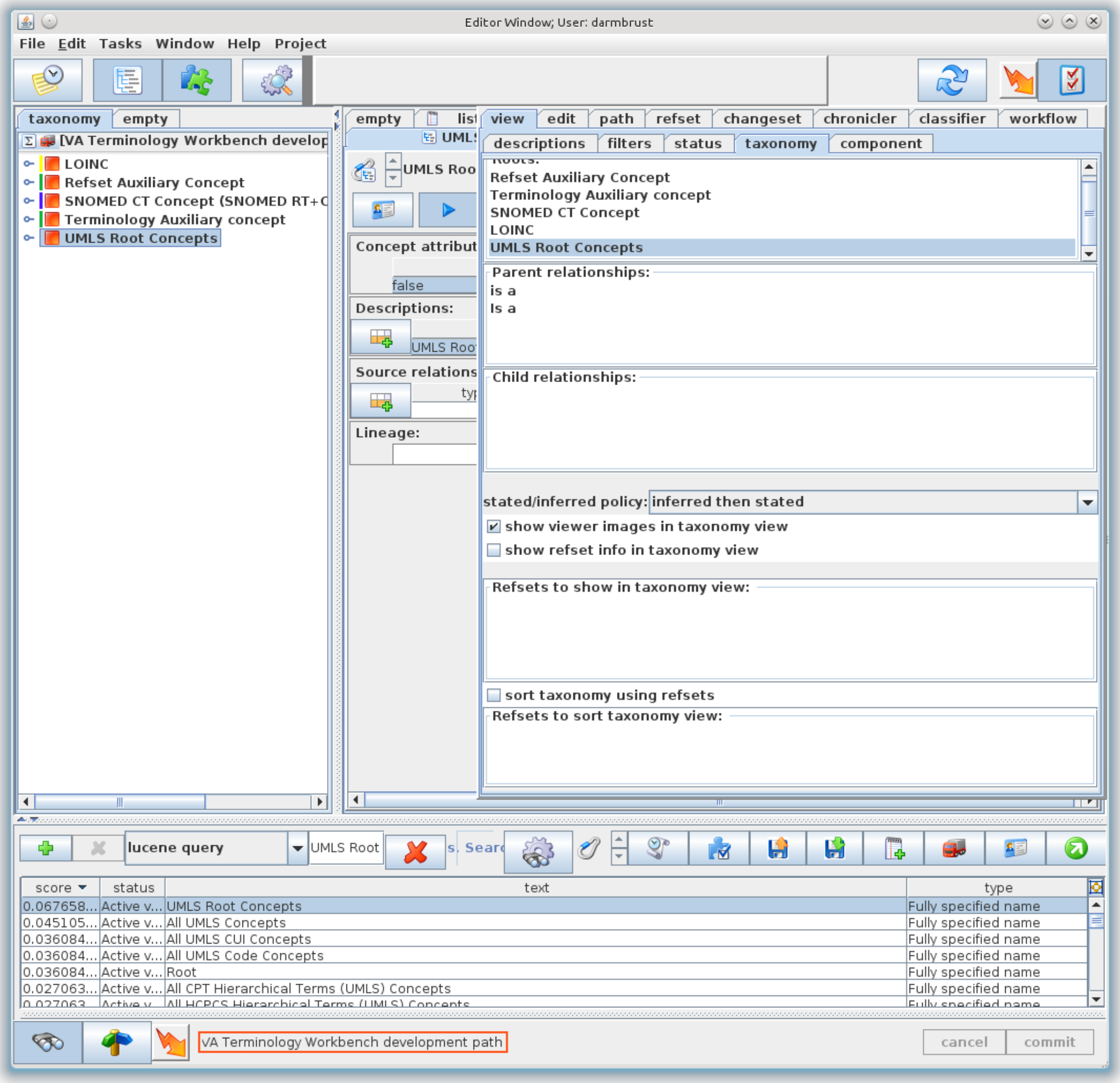
Add the `SimpleUniversalAcePosition` section, as documented below.

```
<profiles>
  …
  <profile> (init-db)
    …
    <build>
      …
      <plugins>
        …
        <plugin> (wb-mojo)
          …
          <executions>
            …
            <execution> (create-baseline-development-path)
              …
              <configuration>
                …
                <origins>
                  …
                  <SimpleUniversalAcePosition>
                    <pathConcept>
                      <pathFsDesc>UMLS Path</pathFsDesc>
                      <description>UMLS Path</description>
                    </pathConcept>
                    <timeStr>latest</timeStr>
                  </SimpleUniversalAcePosition>
                </origins>
```

5. UMLS Terminologies in the Workbench

5.1. The UMLS Taxonomy in the Workbench

In order to browse UMLS terminologies in the Workbench add the root node “UMLS Root Concepts” into the Workbench taxonomy configuration. Search for “UMLS Root Concepts” and drag the result into the taxonomy configuration dialog, as shown below.


Editor Window; User: darmbrust
File Edit Tasks Window Help Project
taxonomy empty
empty list
view edit path refset changeset chronicler classifier workflow
descriptions filters status taxonomy component
Concept attribut
false
Descriptions:
UMLS Roo
Source relations
typ
Lineage:
lucene query
UMLS Root
Searc
cancel commit

[VA Terminology Workbench develop
- LOINC
- Refset Auxiliary Concept
- SNOMED CT Concept (SNOMED RT+C
- Terminology Auxiliary concept
- UMLS Root Concepts

Roots:
Refset Auxiliary Concept
Terminology Auxiliary concept
SNOMED CT Concept
LOINC
UMLS Root Concepts

Parent relationships:
is a
Is a

Child relationships:

stated/inferred policy: inferred then stated

☑ show viewer images in taxonomy view

☐ show refset info in taxonomy view

Refsets to show in taxonomy view:

☐ sort taxonomy using refsets

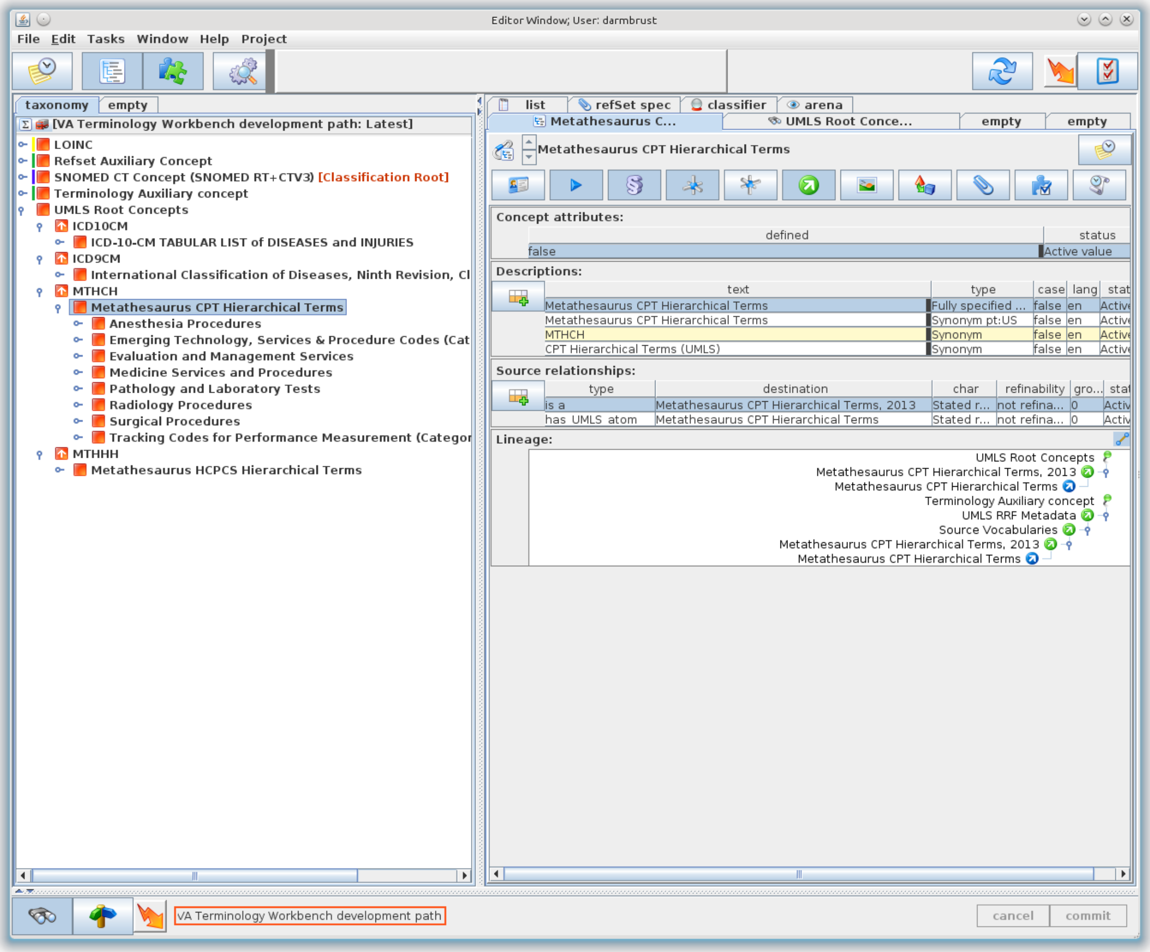
Refsets to sort taxonomy view:

score	status	text	type
0.067658...	Active v...	UMLS Root Concepts	Fully specified name
0.045105...	Active v...	All UMLS Concepts	Fully specified name
0.036084...	Active v...	All UMLS CUI Concepts	Fully specified name
0.036084...	Active v...	All UMLS Code Concepts	Fully specified name
0.036084...	Active v...	Root	Fully specified name
0.027063...	Active v...	All CPT Hierarchical Terms (UMLS) Concepts	Fully specified name
0.027063	Active v	All HCPCS Hierarchical Terms (UMLS) Concepts	Fully specified name

VA Terminology Workbench development path

Figure 1 - Adding UMLS Root Concepts to the Taxonomy View

Each terminology that was imported from the UMLS which has a hierarchy will have a SAB concept located under the “UMLS Root Concepts” node. Under each SAB node will be the hierarchy as provided by the UMLS.



[VA Terminology Workbench development path: Latest]

- LOINC
- Refset Auxiliary Concept
- SNOMED CT Concept (SNOMED RT+CTV3) [Classification Root]
- Terminology Auxiliary concept
- UMLS Root Concepts
  - ICD10CM
    - ICD-10-CM TABULAR LIST of DISEASES and INJURIES
  - ICD9CM
    - International Classification of Diseases, Ninth Revision, Cl
  - MTHCH
    - Metathesaurus CPT Hierarchical Terms
      - Anesthesia Procedures
      - Emerging Technology, Services & Procedure Codes (Cat
      - Evaluation and Management Services
      - Medicine Services and Procedures
      - Pathology and Laboratory Tests
      - Radiology Procedures
      - Surgical Procedures
      - Tracking Codes for Performance Measurement (Categor
  - MTHHH
    - Metathesaurus HCPCS Hierarchical Terms

Metathesaurus CPT Hierarchical Terms

Concept attributes:

defined	status
false	Active value

Descriptions:

text	type	case	lang	stat
Metathesaurus CPT Hierarchical Terms	Fully specified ...	false	en	Active
Metathesaurus CPT Hierarchical Terms	Synonym pt:US	false	en	Active
MTHCH	Synonym	false	en	Active
CPT Hierarchical Terms (UMLS)	Synonym	false	en	Active

Source relationships:

type	destination	char	refinability	gro...	stat
is a	Metathesaurus CPT Hierarchical Terms, 2013	Stated r...	not refina...	0	Activ
has UMLS atom	Metathesaurus CPT Hierarchical Terms	Stated r...	not refina...	0	Activ

Lineage:

UMLS Root Concepts
Metathesaurus CPT Hierarchical Terms, 2013
Metathesaurus CPT Hierarchical Terms
Terminology Auxiliary concept
UMLS RRF Metadata
Source Vocabularies
Metathesaurus CPT Hierarchical Terms, 2013
Metathesaurus CPT Hierarchical Terms

VA Terminology Workbench development path

Figure 2 - Terminology Hierarchies

The content of terminologies imported from the UMLS which do not have a hierarchy can be listed from a corresponding refset concept. See section 5.6.3 Refsets.

5.2. UMLS concepts in the Workbench

The UMLS contains two different types of primary identifiers – Concept Unique Identifiers (CUIs) and AUIs (Atom Unique Identifiers).

CUIs in the UMLS represent a meaning, and are linked to a number of AUI based concepts that have the same meaning.

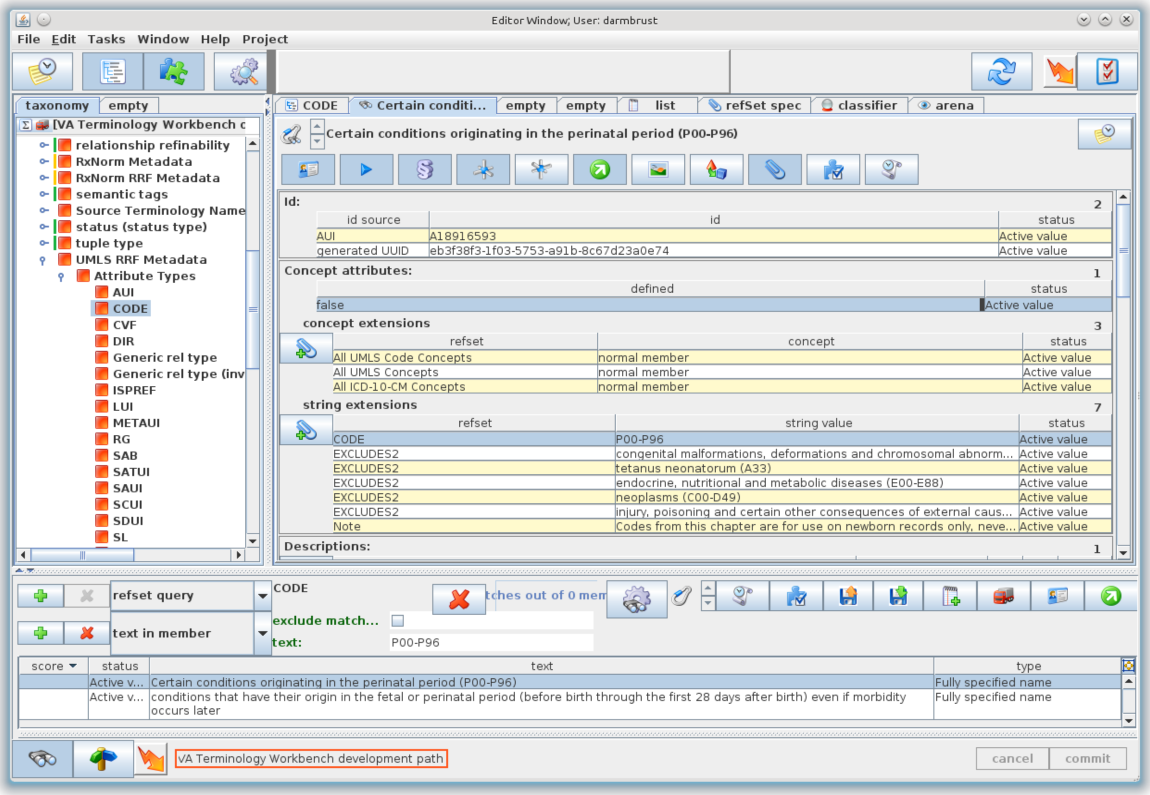
5.2.1. AUIs

In the UMLS, there is one unique AUI per description. Each AUI links to the source terminology code where the description originated. When the AUIs are loaded into the Workbench, the AUIs are grouped into unique AUI based concepts that share the same CUI, SAB, and CODE. The following table illustrates how a set of AUIs is converted into AUI based concepts in the Workbench.

AUI	CUI	SAB	CODE	DESCRIPTION	WB CONCEPT
A0018160	C0000924	MSH	D000059	Accidents	A
A17996753	C0000924	MSH	D000059	Accident	
A4807043	C0000924	SNOMEDCT	158001009	Accidents NOS	B
A16365746	C0000924	SNOMEDCT	218233008	Unspecified accidents	C
A16351099	C0000924	SNOMEDCT	218233008	Unspecified accidents (morphologic abnormality)	

The unique UUID for each AUI based Workbench concept that is created is based on the combination of the CUI, SAB and CODE. Each AUI is added as an additional ID – so any AUI can be looked up directly in the WB by doing a lucene search for the specified AUI. The CODE values in the UMLS do not satisfy the requirements for an ID type in the Workbench, so these are loaded as attributes.

These can be searched by doing a refset query on the “CODE” refset with a “text in member” restriction. See Figure 3 for an example of searching for the code “P00-P96”



Certain conditions originating in the perinatal period (P00-P96)

Id: 2

id source	id	status
AUI	A18916593	Active value
generated UUID	eb3f38f3-1f03-5753-a91b-8c67d23a0e74	Active value

Concept attributes: 1

defined	status
false	Active value

concept extensions 3

refset	concept	status
All UMLS Code Concepts	normal member	Active value
All UMLS Concepts	normal member	Active value
All ICD-10-CM Concepts	normal member	Active value

string extensions 7

refset	string value	status
CODE	P00-P96	Active value
EXCLUDES2	congenital malformations, deformations and chromosomal abnorm...	Active value
EXCLUDES2	tetanus neonatorum (A33)	Active value
EXCLUDES2	endocrine, nutritional and metabolic diseases (E00-E88)	Active value
EXCLUDES2	neoplasms (C00-D49)	Active value
EXCLUDES2	injury, poisoning and certain other consequences of external caus...	Active value
Note	Codes from this chapter are for use on newborn records only, neve...	Active value

Descriptions: 1

refset query: CODE

text in member

exclude match... □

text: P00-P96

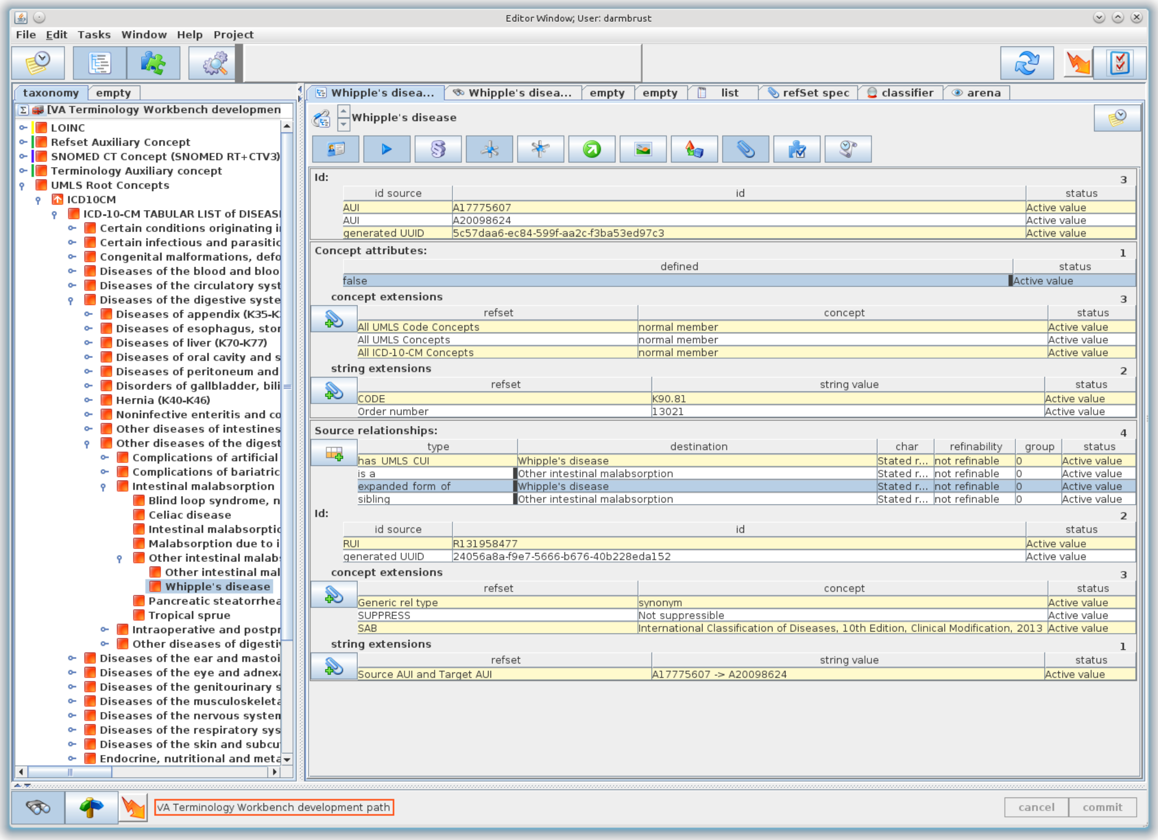
score	status	text	type
	Active v...	Certain conditions originating in the perinatal period (P00-P96)	Fully specified name
	Active v...	conditions that have their origin in the fetal or perinatal period (before birth through the first 28 days after birth) even if morbidity occurs later	Fully specified name

VA Terminology Workbench development path

Figure 3 - Refset Query

5.2.2. AUI-based concepts

Below, the concept “Whipple’s disease” is shown in the Workbench with a subset of attribute view toggles enabled.



[VA Terminology Workbench developmen

- LOINC
- Refset Auxiliary Concept
- SNOMED CT Concept (SNOMED RT+CTV3)
- Terminology Auxiliary concept
- UMLS Root Concepts
  - ICD10CM
    - ICD-10-CM TABULAR LIST of DISEASI
      - Certain conditions originating i
      - Certain infectious and parasitic
      - Congenital malformations, defo
      - Diseases of the blood and bloo
      - Diseases of the circulatory syst
      - Diseases of the digestive syste
        - Diseases of appendix (K35-K
        - Diseases of esophagus, stor
        - Diseases of liver (K70-K77)
        - Diseases of oral cavity and s
        - Diseases of peritoneum and
        - Disorders of gallbladder, bili
        - Hernia (K40-K46)
        - Noninfective enteritis and co
        - Other diseases of intestines
        - Other diseases of the digest
          - Complications of artificial
          - Complications of bariatric
          - Intestinal malabsorption
            - Blind loop syndrome, n
            - Celiac disease
            - Intestinal malabsorptic
            - Malabsorption due to i
            - Other intestinal malab
              - Other intestinal mal
              - Whipple's disease
            - Pancreatic steatorrhea
            - Tropical sprue
          - Intraoperative and postpr
          - Other diseases of digesti
      - Diseases of the ear and mastoi
      - Diseases of the eye and adnexa
      - Diseases of the genitourinary s
      - Diseases of the musculoskeleta
      - Diseases of the nervous systen
      - Diseases of the respiratory sys
      - Diseases of the skin and subcu
      - Endocrine, nutritional and meta

VA Terminology Workbench development path

Whipple's disease

Id: 3

id source	id	status
AUI	A17775607	Active value
AUI	A20098624	Active value
generated UUID	5c57daa6-ec84-599f-aa2c-f3ba53ed97c3	Active value

Concept attributes: 1

defined	status
false	Active value

concept extensions 3

refset	concept	status
All UMLS Code Concepts	normal member	Active value
All UMLS Concepts	normal member	Active value
All ICD-10-CM Concepts	normal member	Active value

string extensions 2

refset	string value	status
CODE	K90.81	Active value
Order number	13021	Active value

Source relationships: 4

type	destination	char	refinability	group	status
has UMLS CUI	Whipple's disease	Stated r...	not refinable	0	Active value
is a	Other intestinal malabsorption	Stated r...	not refinable	0	Active value
expanded form of	Whipple's disease	Stated r...	not refinable	0	Active value
sibling	Other intestinal malabsorption	Stated r...	not refinable	0	Active value

Id: 2

id source	id	status
RUI	R131958477	Active value
generated UUID	24056a8a-f9e7-5666-b676-40b228eda152	Active value

concept extensions 3

refset	concept	status
Generic rel type	synonym	Active value
SUPPRESS	Not suppressible	Active value
SAB	International Classification of Diseases, 10th Edition, Clinical Modification, 2013	Active value

string extensions 1

refset	string value	status
Source AUI and Target AUI	A17775607 -> A20098624	Active value

Figure 4 - AUI-Based Concept

This view shows the following information:

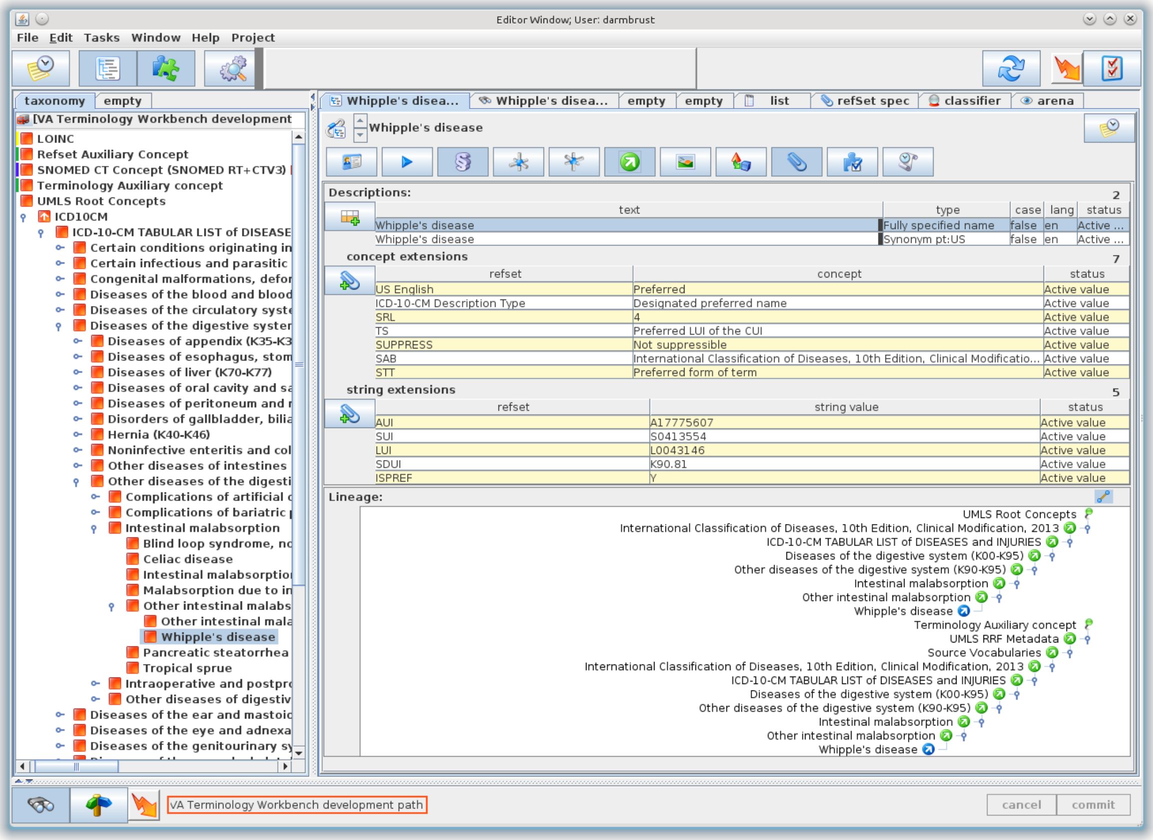
- Id – The Id section shows the identifiers from each AUI that was merged into this concept, and the UUID which was generated from a combination of the CUI/SAB/CODE which is required by the Workbench.

- Concept attributes – The Concept attributes section shows the current status of the Concept (active, inactive) and also any attributes and refset membership information for this concept. On this concept, there are two string extensions (CODE, Order number) and three refset memberships (All UMLS Code Concepts, All UMLS Concepts, All ICD-10-CM Concepts)

- Source relationships – Workbench native “is a” relationships were created from parent/child relationships in the UMLS. The UMLS relationship name is added to the “is a” relationship as concept extension. Other relationship types are used directly from the UMLS.

- In the UMLS, relationships can also have attributes – each of these is included on the relationship as a concept or string extension. A “has_UMLS_CUI” relationship is also created on each AUI based concept, which links to the associated CUI. Other aspects of Relationships are discussed in more detail in section 5.3.

Continuing an overview of the content in the Workbench, below is the same concept with different attribute view toggles enabled.


Editor Window; User: darmbrust
File Edit Tasks Window Help Project
taxonomy
empty
[VA Terminology Workbench development
LOINC
Refset Auxiliary Concept
SNOMED CT Concept (SNOMED RT+CTV3)
Terminology Auxiliary concept
UMLS Root Concepts
ICD10CM
ICD-10-CM TABULAR LIST of DISEASE
Certain conditions originating in
Certain infectious and parasitic
Congenital malformations, defor
Diseases of the blood and blood
Diseases of the circulatory syste
Diseases of the digestive syster
Diseases of appendix (K35-K3
Diseases of esophagus, stom
Diseases of liver (K70-K77)
Diseases of oral cavity and sa
Diseases of peritoneum and r
Disorders of gallbladder, bilia
Hernia (K40-K46)
Noninfective enteritis and col
Other diseases of intestines
Other diseases of the digesti
Complications of artificial c
Complications of bariatric p
Intestinal malabsorption
Blind loop syndrome, no
Celiac disease
Intestinal malabsorptio
Malabsorption due to in
Other intestinal malabs
Other intestinal mala
Whipple's disease
Pancreatic steatorrhea
Tropical sprue
Intraoperative and postpro
Other diseases of digestiv
Diseases of the ear and mastoic
Diseases of the eye and adnexa
Diseases of the genitourinary sy
Whipple's disea...
Whipple's disea...
empty
empty
list
refSet spec
classifier
arena
Whipple's disease
VA Terminology Workbench development path
cancel
commit

Descriptions: 2

text	type	case	lang	status
Whipple's disease	Fully specified name	false	en	Active ...
Whipple's disease	Synonym pt:US	false	en	Active ...

concept extensions 7

refset	concept	status
US English	Preferred	Active value
ICD-10-CM Description Type	Designated preferred name	Active value
SRL	4	Active value
TS	Preferred LUI of the CUI	Active value
SUPPRESS	Not suppressible	Active value
SAB	International Classification of Diseases, 10th Edition, Clinical Modificatio...	Active value
STT	Preferred form of term	Active value

string extensions 5

refset	string value	status
AUI	A17775607	Active value
SUI	S0413554	Active value
LUI	L0043146	Active value
SDUI	K90.81	Active value
ISPREF	Y	Active value

Lineage:

UMLS Root Concepts
International Classification of Diseases, 10th Edition, Clinical Modification, 2013
ICD-10-CM TABULAR LIST of DISEASES and INJURIES
Diseases of the digestive system (K00-K95)
Other diseases of the digestive system (K90-K95)
Intestinal malabsorption
Other intestinal malabsorption
Whipple's disease
Terminology Auxiliary concept
UMLS RRF Metadata
Source Vocabularies
International Classification of Diseases, 10th Edition, Clinical Modification, 2013
ICD-10-CM TABULAR LIST of DISEASES and INJURIES
Diseases of the digestive system (K00-K95)
Other diseases of the digestive system (K90-K95)
Intestinal malabsorption
Other intestinal malabsorption
Whipple's disease

Figure 5 - AUI Based Concept

This view shows the following information:

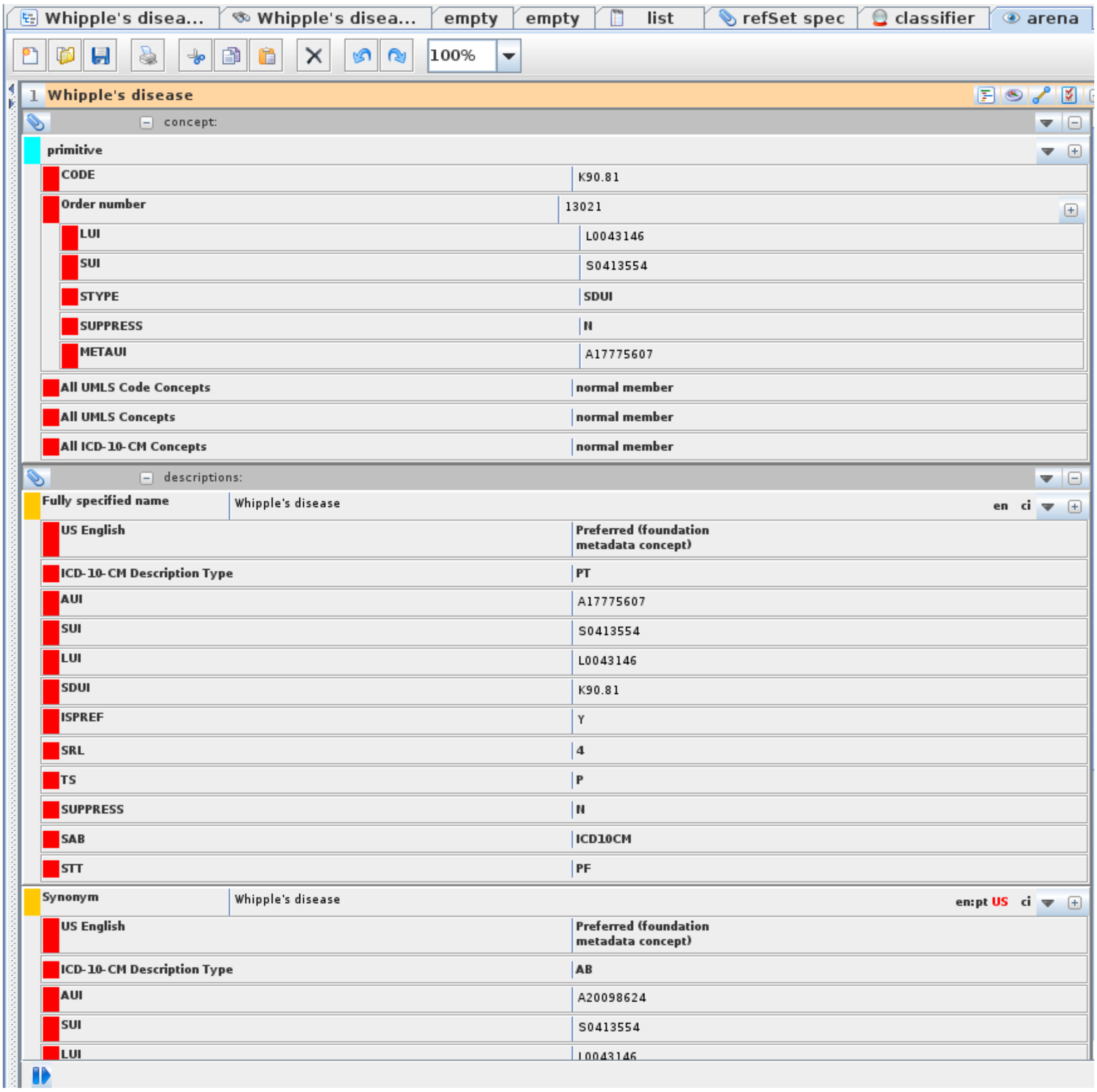
- Descriptions – Each string description for the current group of AUIs is created as a description on the concept. The descriptions are added using the Workbench description types (Fully Specified Name, Synonym). The UMLS description type of each description is added as an extension on each description. In this example, we show that the "Fully specified name" description came from the "ICD-10-CM Description Type" "Designated preferred name".

- Each AUI in the UMLS also has a number of attributes attached to it. These are attached to the description as concept extensions or string extensions. In general, the attribute types attached to the description correspond to the data columns in the RRF release format. The AUI identifier for the description is also attached to the description in the case where more than one AUI has been grouped on a single concept. The AUI extension is omitted in cases where there is only one AUI tied to the concept, as the AUI can then be found in the ID section.

- Lineage – The path(s) to the root concept for this concept.

5.2.2.1. Arena View

The primary Workbench view panels only show one level of attributes on data elements like description and relationship. In order to see nested relationships, you need to switch to the Arena View tab. Below is the same concept again, but in the Arena Viewer tab.



1 Whipple's disease

concept:

primitive

CODE	K90.81	
Order number	13021	
LUI	L0043146	
SUI	S0413554	
STYPE	SDUI	
SUPPRESS	N	
METAUI	A17775607	
All UMLS Code Concepts	normal member	
All UMLS Concepts	normal member	
All ICD-10-CM Concepts	normal member	

descriptions:

Fully specified name	Whipple's disease	en ci
US English	Preferred (foundation metadata concept)	
ICD-10-CM Description Type	PT	
AUI	A17775607	
SUI	S0413554	
LUI	L0043146	
SDUI	K90.81	
ISPREF	Y	
SRL	4	
TS	P	
SUPPRESS	N	
SAB	ICD10CM	
STT	PF	
Synonym	Whipple's disease	en:pt US ci
US English	Preferred (foundation metadata concept)	
ICD-10-CM Description Type	AB	
AUI	A20098624	
SUI	S0413554	
LUI	L0043146	

Figure 6 - Arena Tab

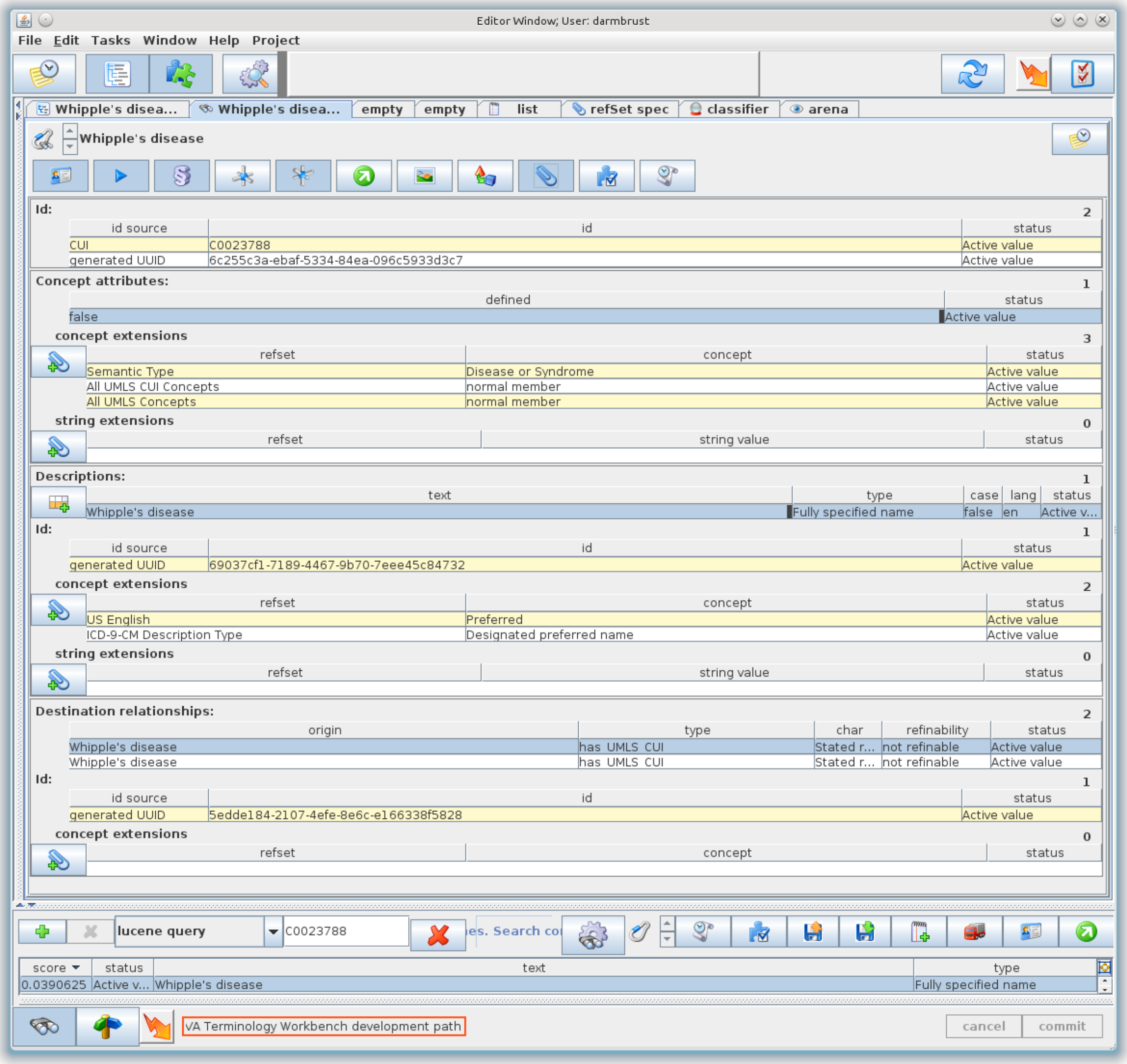
Here you can see that the concept attribute "Order number" has five nested attributes.

5.2.3. CUIs

In the Workbench, one unique concept is created for each CUI that is found in the UMLS. The highest-ranking description for the CUI concept (based on the precedence settings given during MetamorphoSys) that is available in the included terminologies is used as the FSN for this concept.

All attributes, relationships, refset memberships and descriptions which are attached to the CUI in the UMLS are attached to the CUI concept in the Workbench.

CUI concepts in the Workbench are not part of the terminology hierarchy. CUI concept codes can be looked up by ID in the lucene query panel. Figure 7 - CUI Conceptshows a search for CUI “C0023788” in the Workbench.



Whipple's disease

Id: 2

id source	id	status
CUI	C0023788	Active value
generated UUID	6c255c3a-ebaf-5334-84ea-096c5933d3c7	Active value

Concept attributes: 1

defined	status
false	Active value

concept extensions 3

refset	concept	status
Semantic Type	Disease or Syndrome	Active value
All UMLS CUI Concepts	normal member	Active value
All UMLS Concepts	normal member	Active value

string extensions 0

refset	string value	status

Descriptions: 1

text	type	case	lang	status
Whipple's disease	Fully specified name	false	en	Active v...

Id: 1

id source	id	status
generated UUID	69037cf1-7189-4467-9b70-7eee45c84732	Active value

concept extensions 2

refset	concept	status
US English	Preferred	Active value
ICD-9-CM Description Type	Designated preferred name	Active value

string extensions 0

refset	string value	status

Destination relationships: 2

origin	type	char	refinability	status
Whipple's disease	has UMLS CUI	Stated r...	not refinable	Active value
Whipple's disease	has UMLS CUI	Stated r...	not refinable	Active value

Id: 1

id source	id	status
generated UUID	5edde184-2107-4efe-8e6c-e166338f5828	Active value

concept extensions 0

refset	concept	status

lucene query: C0023788

score	status	text	type
0.0390625	Active v...	Whipple's disease	Fully specified name

VA Terminology Workbench development path

Figure 7 - CUI Concept

On this particular CUI concept, you can see the following information:

- Id – The Id section shows the identifiers that were created from the CUI – the first is the CUI itself, and the second is the UUID which is required by the Workbench.

- Concept Attributes – The Concept Attributes section shows the current status of the Concept (active, inactive) and also any attributes and refset membership information for this concept. On this concept, there are no attributes and three refset memberships (Semantic Type – Disease or Syndrome, All UMLS CUI Concepts, All UMLS Concepts)

- Descriptions – A string description is created from the highest ranked AUI description that is linked to this CUI. The description is entered as the Workbench description type "Fully specified name". The source of the description is added as concept extensions to each description. In this example, you can see that the "Fully specified name" description came from the "ICD-9-CM Description Type" "Designated preferred name".

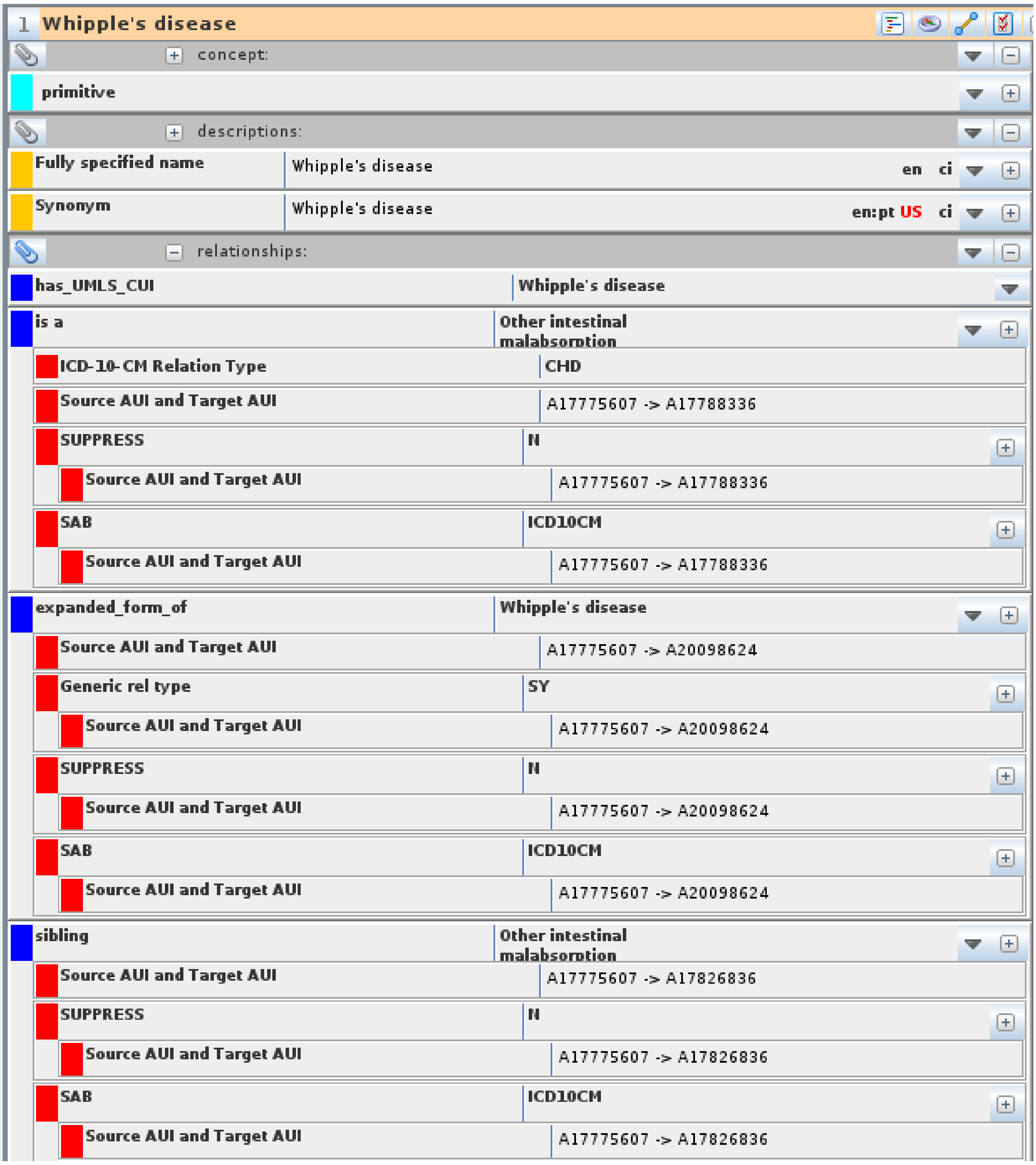
- Destination relationships – The inverse relationships that point to this CUI concept. For CUI concepts, these will include the "has_UMLS_CUI" relationship, which links the AUI based concepts to the CUI based concepts.

5.3. Relationships

In the UMLS, many relationships have two different relationship names – a specific name such as "expanded form of" and a generic name such as "synonym". When a specific relationship name is available – this is used as the Workbench relationship type. The generic relationship name is added as a relationship extension. If no specific relationship name is available, then the generic relationship name is used as the Workbench relationship type.

In the image below, we can see an example of each of these. The relationship "expanded_form_of" is a specific relationship type, so it has a nested "Generic rel type" attribute with a value of "SY" – also known as "Synonym". This is the generic relationship name attached to this particular relationship in the UMLS.

The relationship "sibling" is already a generic relationship type, so it is used directly with no further qualification. There is no specific relationship name available in the UMLS for this particular relationship.



1 Whipple's disease

concept:

primitive	

descriptions:

Fully specified name	Whipple's disease	en	ci
Synonym	Whipple's disease	en:pt US	ci

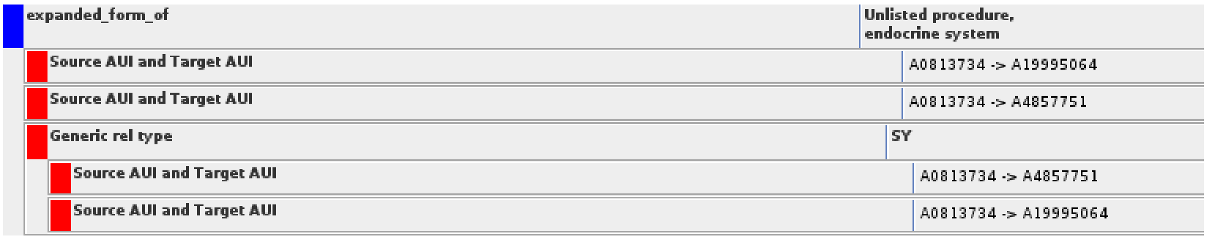
relationships:

- has_UMLS_CUI — Whipple's disease
- is a — Other intestinal malabsorption
  - ICD-10-CM Relation Type — CHD
  - Source AUI and Target AUI — A17775607 -> A17788336
  - SUPPRESS — N
    - Source AUI and Target AUI — A17775607 -> A17788336
  - SAB — ICD10CM
    - Source AUI and Target AUI — A17775607 -> A17788336
- expanded_form_of — Whipple's disease
  - Source AUI and Target AUI — A17775607 -> A20098624
  - Generic rel type — SY
    - Source AUI and Target AUI — A17775607 -> A20098624
  - SUPPRESS — N
    - Source AUI and Target AUI — A17775607 -> A20098624
  - SAB — ICD10CM
    - Source AUI and Target AUI — A17775607 -> A20098624
- sibling — Other intestinal malabsorption
  - Source AUI and Target AUI — A17775607 -> A17826836
  - SUPPRESS — N
    - Source AUI and Target AUI — A17775607 -> A17826836
  - SAB — ICD10CM
    - Source AUI and Target AUI — A17775607 -> A17826836

Figure 8 - Relationship Attributes

The image also shows numerous nested attributes named "Source AUI and Target AUI". These attributes are used to identify the specific AUI values that declared the relationship and the nested attribute values. This is necessary because when multiple AUIs are merged into a single concept in the Workbench, it would be impossible to link back to the actual source of the relationship without this value.

Frequently, as a result of the merged AUIs, duplicate definitions of the same relationships will be created. These duplicates are automatically removed during the conversion – but the Source AUI and Target AUI annotations will remain to identify each unique AUI pair that had the relationship in the UMLS. The image below shows a case where a duplicate relationship was merged.



expanded_form_of	Unlisted procedure, endocrine system
Source AUI and Target AUI	A0813734 -> A19995064
Source AUI and Target AUI	A0813734 -> A4857751
Generic rel type	SY
Source AUI and Target AUI	A0813734 -> A4857751
Source AUI and Target AUI	A0813734 -> A19995064

Only one instance of the “expanded_form_of” relationship was created – but both AUI pairs that listed this relationship in the UMLS are added as attributes. Additionally, each relationship also declared the same “Generic rel type” attribute in this case – this duplication was removed – “Generic rel type” is only listed once, but nested attributes were added to provide a link to each unique source of the relationship information in the UMLS that specified the “Generic rel type”.

The UMLS also declares all relationships in both directions – so the UMLS source data will contain:

```
A -> expanded_form_of-> B
B -> has_expanded_form -> A
```

The Workbench automatically supports reversing relationships – so only one half of the relationships from the UMLS are loaded. The reverse relationships are not necessary. In this example, the relationship `A -> expanded_form_of-> B` will be loaded onto concept A – which is shown in the “Source Relationship” section for concept A. In concept B, the same relationship is visible in the “Destination Relationship” section.

The concept that defines “expanded_form_of” within the terminology metadata hierarchy (see section 5.6.1) carries the inverse relationship name from the UMLS as a secondary description.

5.4. SNOMED CT Relationships

SNOMED CT and the SNOMED CT US Extension are not loaded into the Workbench from the UMLS content, as these terminologies already exist in the Workbench in their native format.

When SNOMED-related terminologies are processed by the loader, they are handled as a special case. As an example, when a relationship is found that crosses from an ICD-9-CM AUI concept to a SNOMED CT AUI concept, the SNOMED concept is mapped directly to the appropriate Workbench UUID identifier for the SNOMED concept referenced in the UMLS.

When a relationship is found that references a CUI concept that is only linked to SNOMED concepts (and therefore, would not normally be loaded because SNOMED concepts are not loaded into the Workbench from the UMLS content) – the necessary SNOMED CUI concept is created. This CUI concept is then linked via the "has_UMLS_CUI" relationship to the appropriate UUID SNOMED concept(s) within the Workbench.

> Note: Normally, the "has_UMLS_CUI" relationship is stored on the AUI based concept, and references the CUI based concept. In cases involving SNOMED CT concepts since we are not creating the UUID based SNOMED concepts, the relationship is loaded in reverse. Meaning, the "has_UMLS_CUI" relationship is stored on the CUI based concept, and it references the UUID SNOMED concept.

Relationships between CUI concepts are then created between the new SNOMED CUI concepts in the same way as any other CUI concept relationship.

5.5. Metathesaurus Relationships

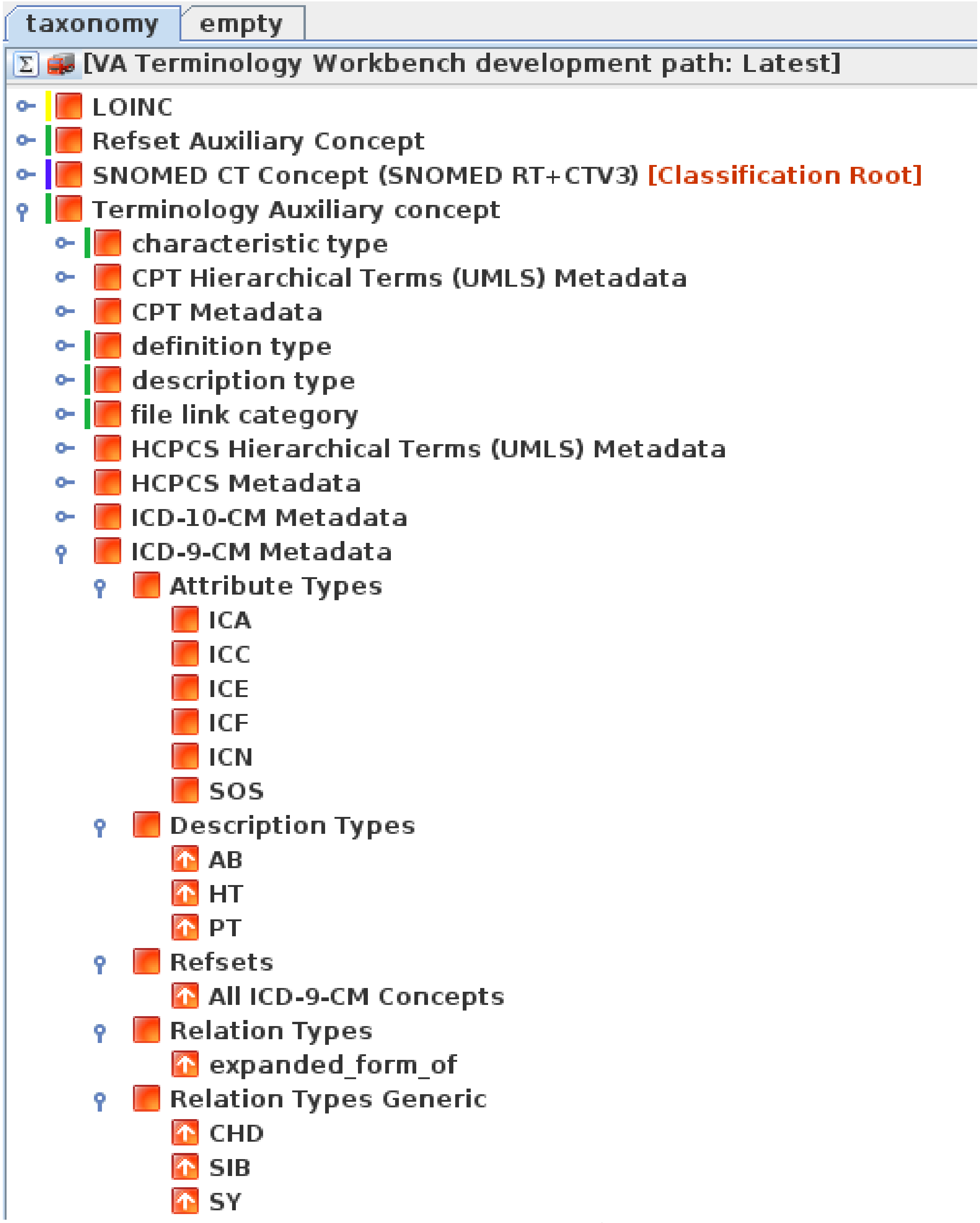
The Metathesaurus (MTH) terminology within the UMLS is another terminology with (optional) special handling. When the special handling is enabled in the loader (which it is by default) the MTH terminology concepts are not loaded. Instead, only the relationships between CUIs from MTH are loaded.

5.6. UMLS metadata

Various metadata hierarchies are created under the "Terminology Auxiliary Concept" when content is loaded from the UMLS.

5.6.1. UMLS terminology-specific metadata

Each terminology that is loaded has a terminology-specific metadata hierarchy under "Terminology Auxiliary Concept". Figure 9 below shows the metadata node for ICD-9-CM. The metadata hierarchy for each terminology is further grouped by functional use. These groupings are described below.



taxonomy | empty

[VA Terminology Workbench development path: Latest]

- LOINC
- Refset Auxiliary Concept
- SNOMED CT Concept (SNOMED RT+CTV3) [Classification Root]
- Terminology Auxiliary concept
  - characteristic type
  - CPT Hierarchical Terms (UMLS) Metadata
  - CPT Metadata
  - definition type
  - description type
  - file link category
  - HCPCS Hierarchical Terms (UMLS) Metadata
  - HCPCS Metadata
  - ICD-10-CM Metadata
  - ICD-9-CM Metadata
    - Attribute Types
      - ICA
      - ICC
      - ICE
      - ICF
      - ICN
      - SOS
    - Description Types
      - AB
      - HT
      - PT
    - Refsets
      - All ICD-9-CM Concepts
    - Relation Types
      - expanded_form_of
    - Relation Types Generic
      - CHD
      - SIB
      - SY

Figure 9 - ICD-10 Metadata.

- Attribute Types – These are the extension types that are used on concepts, descriptions and relationships within the ICD-9-CM terminology. Each of these concepts also doubles as a refset, which can be dragged into the refSet spec tab – where you can see a list of the concepts that use each attribute.

- Description Types – These are the UMLS description types that are used within the ICD-9-CM terminology. These types are used as qualifying types on the core Workbench description types.

- Refsets – These are the member refsets that were created from the ICD-9-CM content. Each terminology that is loaded from the UMLS will have an “All

<terminology name> Concepts" Refset. When this refset is dragged to the refSet spec tab, all AUI based concepts from the specified terminology can be viewed.

- Relation Types – These are the specific UMLS relationship types that are used within the ICD-9-CM terminology as primary relationship types.

- Relationship Types Generic – These are the generic UMLS relationship types that are used within the ICD-9-CM terminology. These types are used as both primary relationship types and as secondary attributes, depending on whether or not the UMLS relationship contains a specific relationship type.

5.6.2. UMLS shared metadata

When data is imported from the UMLS, a "UMLS RRF Metadata" hierarchy node is also created. This hierarchy is shown in Figure 10.

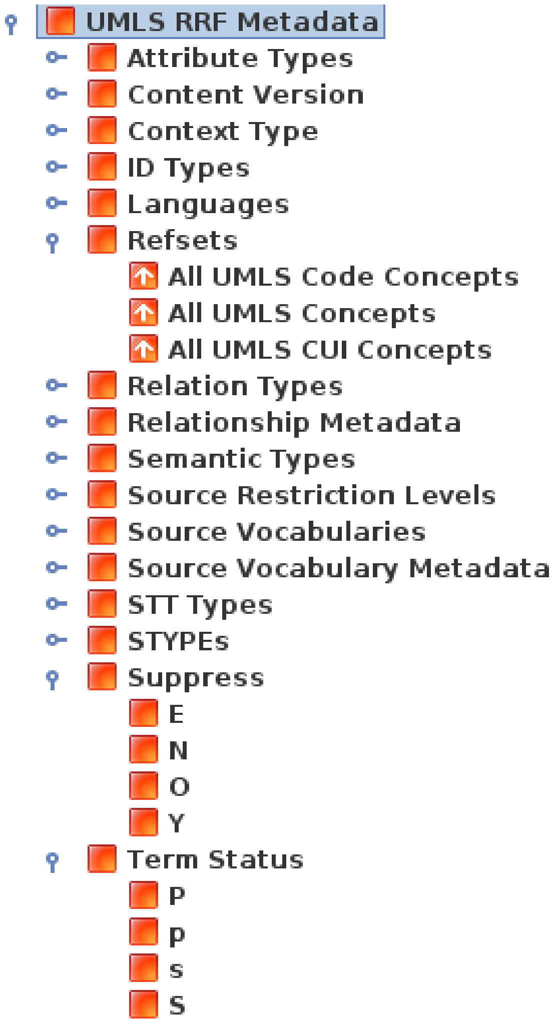


Figure 10 - UMLS Shared Metadata

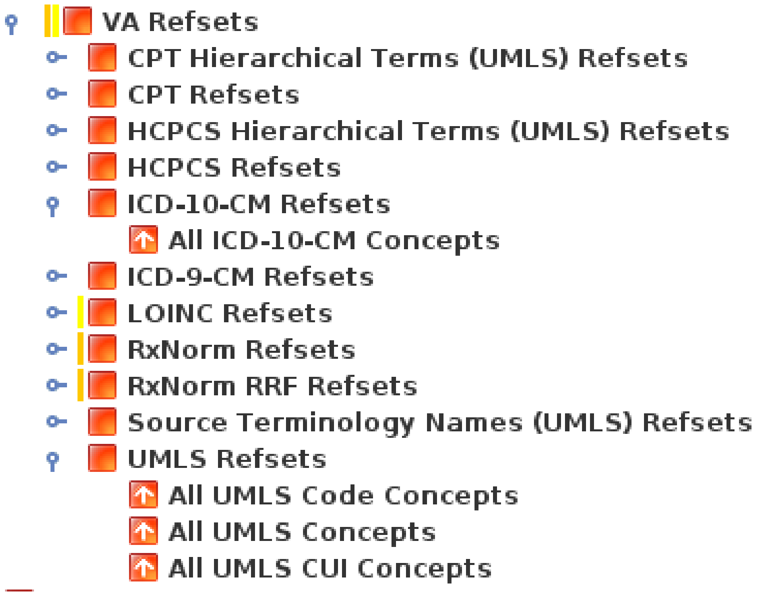
This metadata is used during the creation of Workbench attributes from various data tables and column names within the RRF format. In general, they are created to fulfil requirements of the Workbench data model, and will not be described in detail in this document.

Of note, however, is the Refsets node – which lists three different member refsets, which are created during a UMLS data load.

- All UMLS Code Concepts – This refset contains the AUI based concepts that were created for all terminologies that were imported from the UMLS.
- All UMLS Concepts –This refset contains both the CUI based concepts and the AUI based concepts for all terminologies that were imported from the UMLS.
- All UMLS CUI Concepts – This refset contains all of the CUI based concepts that were created for all terminologies that were imported from the UMLS.

5.6.3. Refsets

Refsets are also stored under the Workbench hierarchy “Refset Auxiliary Concept” -> “refset identity” -> “VA Refsets”.



- VA Refsets
  - CPT Hierarchical Terms (UMLS) Refsets
  - CPT Refsets
  - HCPCS Hierarchical Terms (UMLS) Refsets
  - HCPCS Refsets
  - ICD-10-CM Refsets
    - All ICD-10-CM Concepts
  - ICD-9-CM Refsets
  - LOINC Refsets
  - RxNorm Refsets
  - RxNorm RRF Refsets
  - Source Terminology Names (UMLS) Refsets
  - UMLS Refsets
    - All UMLS Code Concepts
    - All UMLS Concepts
    - All UMLS CUI Concepts

Figure 11 - Refsets

Here, you can see the member refset concepts for each individual terminology, in addition to the UMLS shared refsets.

5.7. Workbench data mapping

The release and version information of the UMLS data is attached to the root node “UMLS Root Concepts” as a set of string annotations.

The rest of the data from a UMLS release is transferred into the Workbench according to the table below.

RRF / UMLS Data File	Workbench Data Type
MRCOLS	This is data about the UMLS release itself – it is used internally in the converter, but not added to the WB.
MRCONSO	This file contains all of the descriptions within the UMLS, and a number of other attributes. The descriptions are loaded onto the appropriate WB concept with a type of "Fully Specified Name" or "Synonym", depending on the precedence settings and the descriptions available on a particular concept. The UMLS description type is added to each description as an annotation. Other attributes from this table are also attached to the description as extensions on the description. A CUI concept is also created for each unique CUI concept found in this file.
MRDEF	This file contains Definitions and a few attributes that attach to the definition. These are loaded into the Workbench on the appropriate concept as descriptions of type Definition. Attributes from this file are added as annotations on each definition.
MRDOC	This is data about the attribute types used in the UMLS release itself – it is used internally in the converter to document the various metadata concepts that are created – but in general is not captured directly into the WB.
MRFILES	This is data about the UMLS release itself – it is used internally in the converter, but not added to the WB.
MRHIER	This file is used to locate the root concept of each terminology that is loaded from the UMLS. The rest of this data duplicates relationship information that we read from MRREL, so it is not loaded.
MRRANK	This file contains the precedence information used to select which description type should become the Fully Specified Name attribute for each WB Concept.
MRREL	Relationships between AUIs and CUIS. Relationships are processed to remove duplicates, and combine inverse relationships into a single relationship – and are then added to the appropriate Workbench Concept. Attributes from this file are added to the relationship as annotations.
MRSAB	This file contains information about each terminology that is available in the UMLS. The relevant data (depending on the terminologies being loaded) is added to the root node for that terminology as string extensions on the concept.
MRSAT	This file contains attributes which may be attached to CUIs, AUIs, or relationships. The attribute types from this file are used to create metadata concepts, and the attribute values are attached to the appropriate Workbench component – Concept, Relationship, etc. This file also contains annotations on each attribute – each of these is added as an annotation on the Workbench Attribute object that is created for the attribute value.
MRSTY	This file contains the Semantic Type information for the CUIs. This data is added to each CUI based concept as a member refset reference.
Other RRF Data files	Not needed at this time for the conversion into the Workbench.

Appendix A

These are the type mapping statistics from the 2013AA release of UMLS, for the requested terminology subset.

Due to the tremendous amount of individual attribute mappings and counts, the data presented here is a high level summary of the statistics from the current release.

The full in-depth statistics are available within the `UMLS-econcept` Maven artifact which is published in step 6 of the Converting the source content step, documented above.

Data Type	Loaded into Workbench
Total Concepts	**618,690**
Additional Concept Identifiers	738,156
Total Relationships	**1,870,238**
Total Descriptions	**738,156**
FSN	618,690
Synonym	119,466
Description Annotations	4,304,917
Total Annotations	**12,626,944**
Concept Annotations	848,560
Other Annotations	11,778,384
Total Refset Members	**977,893**
All CPT Concepts	32,444
All CPT Hierarchical Terms (UMLS) Concepts	1,041
All HCPCS Concepts	5,952
All HCPCS Hierarchical Terms (UMLS) Concepts	347
All ICD-10-CM Concepts	101,954
All ICD-9-CM Concepts	22,401
All Source Terminology Names (UMLS) Concepts	18
All UMLS CUI Concepts	242,711
All UMLS Code Concepts	164,157
All UMLS Concepts	406,868