



Installing Vocabularies from the UMLS Metathesaurus in LexEVS 6.x

Loading from UMLS RRF (Rich Resource Format) files

The National Library of Medicine (NLM) regularly releases the Unified Medical Language System (UMLS) as a set of terminologies in a large set of files referred to as the UMLS Metathesaurus. LexEVS can load the entire set or individual files from this file set using the LexEVS UMLS Batch loader.



The UMLS Metathesaurus tools are briefly discussed on this page. Since all documentation for these are maintained at the website linked to below, only provide brief pointers are provided. Also, LexEVS has a special purpose RRF loader which loads an NCI version of an entire MetaThesaurus. This is explained in [Installing NCI Vocabularies in LexEVS 6.x](#) for the NCI loaders.

Step	Action
1	The UML Metathesaurus is controlled by a license from the NLM. You may obtain a free license of the UMLS Metathesaurus on the NLM license site .
2	Download and install the UMLS Metathesaurus: Loading to LexEVS from these source files requires that they reside locally so it can be accessed quickly from LexEVS. This is not just a download. The UMLS Metathesaurus also has install steps that are documented by NLM on the UMLS website . MetamorphoSys is used to install the UMLS Metathesaurus and to subset terminologies as explained in the subsequent instructions.
3	<p>Subset the desired terminology (recommended):</p> <p>Once the UMLS Metathesaurus is installed, users can either load the entire set of files to LexEVS by pointing to the containing file directory or they can use the UMLS MetamorphoSys tool to subset part of the whole. Subsetting the terminology is the same as creating a smaller set with only your desired concepts. A very common practice is to subset out one of the source terminologies that make up the UMLS. This provides improved performance during loads and is typical when you want only one of the original source coding schemes. Again, the documentation to use MetamorphoSys to subset out a terminology is on the UMLS website but here are a few pointers:</p> <ul style="list-style-type: none">• After the install the Welcome to MetamorphoSys screen has a "Browse my subset" button. If you use it right away you will see the entire UMLS.• Do not browse anything yet or exit back to the Welcome screen.• Under the Advanced menu select Customize My Subset.• Select either of the 2 default subsets as a starting point. Does not matter as you can change it later.• On the Input tab make sure you select just the terminology you want.• On the Output tab select a location for the resultant RRF files.• Under the Done menu select Begin Subset. <p>Once this process is complete you will have a directory with the RRF files required for the LexEVS UMLS Batch loader.</p>
4	<p>Set command line options in the loading script:</p> <p>If you are loading a subset you may or may not need to change the memory options. If you load the entire UMLS Metathesaurus, we usually recommend use of the command line options. This will allow manipulation of the memory allocated for loads of larger terminologies such as SNOMED. Scripting options can be added to the scripts contained at <LEXEVS_HOME>/admin. If a user is working on a Linux environment with a 64 bit architecture, then they can use the LoadUmlsBatch.sh file. On a server class computer with say 16 gigabytes of memory and 8 four core processors users can access fairly substantial resources to load content. Open the .sh file with a text editor and edit the values for -Xmx and -XX:MaxPermSize as follows "-Xmx6000M -XX:MaxPermSize=256M" or more if you have adequate resources available.</p> <p>If you have not set the DB_PRIMARY_KEY value to SEQUENTIAL_INTEGER as described under best practices it could take 33 hours to load a terminology as large as SNOMED which otherwise could complete in 4 hours.</p> <div><p>Memory Handling</p><p>The batch loader is not memory dependent, but at the end of the load the resource is indexed and indexing does require at least 3 GB of memory. Increasing memory can provide faster indexing time.</p></div>
5	<p>Find the SAB (RSAB) in the MRSAB.RRF file:</p> <p>Both the LexEVS Administrative GUI and the LexEVS Administrative commands require the user to enter a SAB or source abbreviation when loading RRF files. This requires that you either know this source abbreviation, or find it in the MRSAB.RRF file contained in the folder of the UMLS installation or the subset you made for the terminology you wish to load. We recommend you open this file in a text editor and search on the terminology name, for instance SNOMED and you should find a line with that name in a row of text separated by a "pipe" character or " ".</p> <div><code>E2960759 C2720507 SNOMEDCT_2011_01_31 SNOMEDCT SNOMED Clinical Terms, 2011_01_31 SNOMEDCT 2011_01_31 2011AA</code></div> <p>The current format of UMLS has the RSAB in column four, which in the case of SNOMED is "SNOMEDCT". Notice that this is a licensed terminology and all use must be in accordance with the licensing agreements.</p>

6	<p>Load the Terminology from the command line referencing the SAB.</p> <pre>./LoadUmlsBatch.sh -in "file:///data/phont/ontologies/2011AA" -s "SNOMEDCT"</pre> <p>The file path is pointing to the directory that contains all of the UMLS RRF files. For UMLS there is not a single source format file.</p>
7	<p>Monitor output (optional): The output from the UMLS batch loader indicates steps of the batch load and can be monitored from the logs at <LEXEVS_HOME>/logs/LexBIG_load_log.text. If you are on Linux this can be done using:</p> <pre>watch -n .1 -d tail LexBIG_load_log.text.</pre> <p>Sample output of an early load step is as follows:</p> <pre> Every 0.1s: tail LexBIG_load_log.txt Fri Nov 11 12:28:53 2011 - Step Property: status=COMPLETED - Step Property: exitStatus=COMPLETED - Step Property: readCount=0 - Step Property: filterCount=0 - Step Property: writeCount=0 readSkipCount=0 - Step Property: writeSkipCount=0 - Step Property: processSkipCount=0 - Step Property: commitCount=1 - Step Property: rollbackCount=0 LB_LOAD_LOGGER DEBUG - 2011-11-11 12:28:38,510 - Step: umlsLoadEntity starting.</pre> <p>Sample output of a final load step:</p> <pre> Every 0.1s: tail LexBIG_load_log.txt Fri Nov 11 12:24:52 2011 - Value: false Property: rrfDir - Value: file:///data/phont/ontologies/2011AA Property: prefix - Value: lbaaaf Property: sab - Value: SNOMEDCT LB_LOAD_LOGGER DEBUG - 2011-11-11 12:22:14,023 - Executing Load Job Id: 1 at Fri Nov 11 12:22:14 CST 2011 LB_LOAD_LOGGER DEBUG - 2011-11-11 12:22:14,033 - Step: loadStagingInsertsTable s tarting.</pre>

Restarting an RRF Load

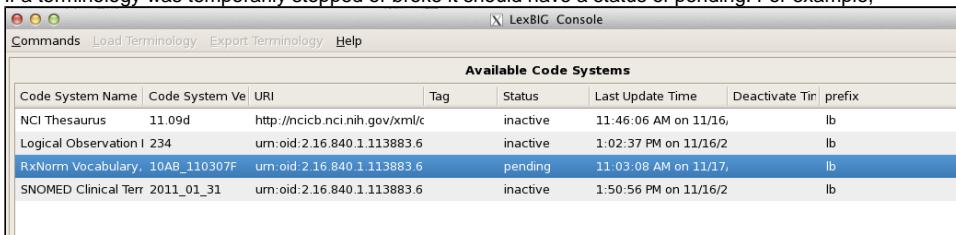


Remember

Killed processes cannot be restarted. The load can recover from an application error, but not from any outside activity that stops the process.

Step	Action
1	Open the LexEVS Administrative GUI.

2 If a terminology was temporarily stopped or broke it should have a status of pending. For example,

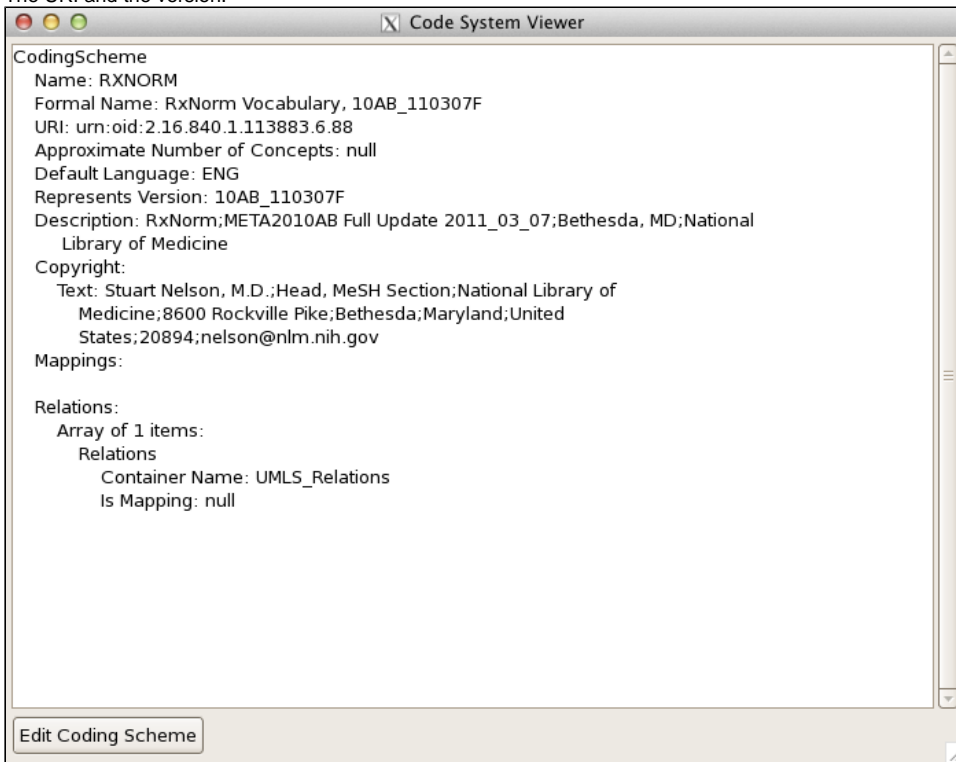


The screenshot shows the LexBIG Console window with a menu bar (Commands, Load Terminology, Export Terminology, Help) and a table titled "Available Code Systems". The table has columns: Code System Name, Code System Ver, URI, Tag, Status, Last Update Time, Deactivate Time, and prefix. The "RxNorm Vocabulary, 10AB_110307F" row is highlighted in blue and has a status of "pending".

Code System Name	Code System Ver	URI	Tag	Status	Last Update Time	Deactivate Time	prefix
NCI Thesaurus	11.09d	http://ncicb.nci.nih.gov/xml/c		inactive	11:46:06 AM on 11/16,		lb
Logical Observation I	234	urn:oid:2.16.840.1.113883.6		inactive	1:02:37 PM on 11/16/2		lb
RxNorm Vocabulary, 10AB_110307F		urn:oid:2.16.840.1.113883.6		pending	11:03:08 AM on 11/17,		lb
SNOMED Clinical Terr	2011_01_31	urn:oid:2.16.840.1.113883.6		inactive	1:50:56 PM on 11/16/2		lb

3 Double click the terminology row in the table.

4 On the resulting window note for the following command line execution:
The URI and the version.



The screenshot shows the Code System Viewer window. It displays the following information for the RXNORM coding scheme:

- Name: RXNORM
- Formal Name: RxNorm Vocabulary, 10AB_110307F
- URI: urn:oid:2.16.840.1.113883.6.88
- Approximate Number of Concepts: null
- Default Language: ENG
- Represents Version: 10AB_110307F
- Description: RxNorm;META2010AB Full Update 2011_03_07;Bethesda, MD;National Library of Medicine
- Copyright:
 - Text: Stuart Nelson, M.D.;Head, MeSH Section;National Library of Medicine;8600 Rockville Pike;Bethesda;Maryland;United States;20894;nelson@nlm.nih.gov
- Mappings:
- Relations:
 - Array of 1 items:
 - Relations
 - Container Name: UMLS_Relations
 - Is Mapping: null

At the bottom of the window is a button labeled "Edit Coding Scheme".

5 Using the LexEVS utilities restart the load of the RRF source and pass in the location, the SAB, the URI, and the version:

```
<LEXEVS_HOME>/admin
```

For Windows installation use the following command:

```
ResumeUmlsBatch.bat -in [file:///home/LargeStorage/ontologies/rrf/RXNORM/2011AA/] \-s RXNORM \-uri "urn:oid:2.16.840.1.113883.6.88" \-version "10AB_110307F"
```

For Linux installation use the following command:

```
./ResumeUmlsBatch.sh \-in [file:///home/LargeStorage/ontologies/rrf/RXNORM/2011AA/] \-s RXNORM \-uri "urn:oid:2.16.840.1.113883.6.88" \-version "10AB_110307F"
```