# Description of premap.pl

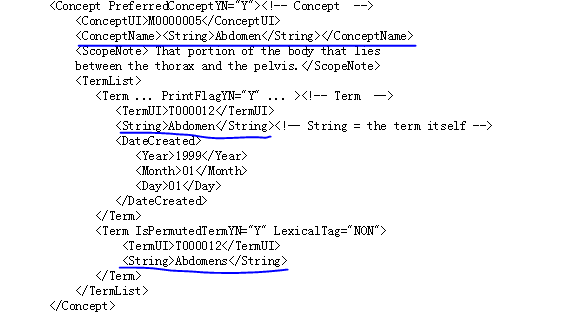
## Things to know before reading the code

We had better learn some knowledge of “MeSH” and “Metathesaurus”. This is very important to help us understand what the “premap.pl” is dong.

### What is “MeSH”?

Mesh is a dictionary which consists of many proper nouns. Any one of these nouns represents a unique medicine concept. This dictionary is stored in three xml files which are “desc2014.xml”, “qual2014.xml” and “supp2014.xml”. Although there are lots of elements in these files, we just need pay attention to three of them, which are **Descriptors**, **Concepts** and **Terms**. It’s important to understand this [three-level concept structure of MeSH data.](https://www.nlm.nih.gov/mesh/xmlmesh.html)

**Concept and Term:**

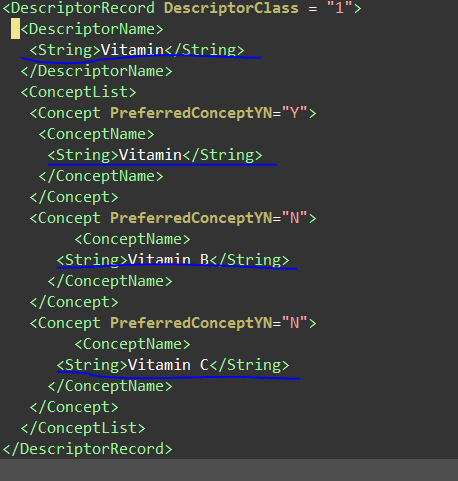


<Concept> usually contains more than one <Term> elements. These terms are synonymous with each other, because that they belong to the same <Concept> means they represent the same concept.

Look at the above illustration. <Concept> has two terms, “Abdomen” and “Abdomens”, and they are synonyms. “Abdomen” has been elected the **preferred name** that uniquely names the concept.

**Descriptor and Concept:**

A “Descriptor” has one or more concepts, the broadest one of which is elected the preferred concept.



This illustration helps us understand the meaning of “broadest”. This “Descriptor” has three concepts, Vitamin, Vitamin B and Vitamin C. Vitamin covers Vitamin B and Vitamin C. But Vitamin B or Vitamin C can’t cover Vitamin. So “Vitamin” is the broadest one and elected preferred concept. Its preferred name is elected the Descriptor name.

### What is Metathesaurus?

There are many dictionaries like MeSH in the world. They are created by different organizations. As the number of dictionaries is increasing, it becomes more and more problematic that the same string represents different concepts in different dictionaries and different strings in various dictionaries represent the same concept actually. In order to solve this problem, UMLS, the creator of MeSH, decides to create a super dictionary to standardize the names of the concepts. [The super dictionary is called Metathesaurus.](https://www.nlm.nih.gov/research/umls/new_users/online_learning/OVR_002.html)

### Mesh and Metathesaurus

Metathesaurus consists of various concepts, which are provided by different dictionaries, [Source Vocabularies](https://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_002.html). When a concept is added to Metathesaurus, it receives a unique identifier, called [CUI](https://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_005.html). CUI isn’t readable at all, so Metathesaurus must also provide a string for every concept to uniquely name the concept.

Given the fact that a concept has different names in various Source Vocabularies, [Metathesaurus selects one of them as the preferred name.](https://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_004.html) Preferred names are computed from a list of ranked source vocabularies. By default, MeSH is ranked 1st in the list, that means the preferred name of a concept in Metathesaurs is the preferred name of it in MeSH.

### Data files in Metathesaurus

Metathersaurus is a large database and its data is stored in tens of [RRF files](http://www.ncbi.nlm.nih.gov/books/NBK9685/). These files can be loaded into MYSQL directly. There are two approaches to make clear what the use of each file is.

**Approach 1: Refer to the files “MRFILES.RRF” and “MRCOLS.RRF”**

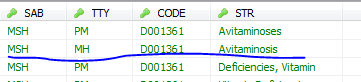
If you are offline, you can refer to the files “MRFILES.RRF” and “MRCOLS.RRF”. There is one row for describing each file briefly in the file. When these RRF files have been loaded into MYSQL, you can refer to the file “MRCOLS.RRF” to make clear what the meaning of each column is.

**Approach 2: Refer to the “UMLS Reference Manual”**

Browsing the web page [Descriptions of Each File](http://www.ncbi.nlm.nih.gov/books/NBK9685/#ch03.I33_Descriptions_of_Each_File) is the best way to understand these RRF files.

### Abbreviations in Metathesaurus.

The values of some columns are various abbreviations. This paragraph just explains my own way to learn about these abbreviations.

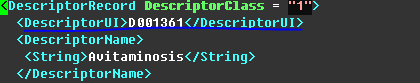


**How do I understand the abbreviation “MH”?**

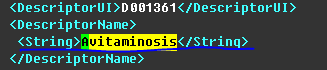
The web page [Abbreviations Used in Data Elements](http://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/release/abbreviations.html) lists all of abbreviations and their corresponding full names. According to the web page, “MH” corresponds to “Main heading”. But what is “Main heading”? This explanation is too short for us. We need more details.

The SAB value, “MSH” suggests that the concept “Avitaminosis” is from MeSH. The CODE value, “D001361” is the Descriptor Unique Identifier. Following is my approach to learning about the abbreviation “MH”:

Step 1: Find “D001361” in MeSH data file “desc2014.xml”;



Step 2: Find the string “Avitaminosis” in the Descriptor Record whose Descriptor Unique Identifier is “D001361”;



Step 3: The string “Avitaminosis” occurs in the element “<DescriptorName>”. So we need to read the introduction in the page [MeSH XML format](https://www.nlm.nih.gov/mesh/xml_data_elements.html). The introduction is comprehensive enough. According to description of the “<DescriptorName>”, we realize that “MH” means the string “Avitaminosis” uniquely represents the Descriptor Record in MeSH.

## What the premap.pl is doing?

### Purpose of premap.pl

In Metathesaurus, a concept belongs to a descriptor provided by MeSH and has many other names provided by other Source Vacabularies. The program “premap.pl” aims to build the mappings between various names and MeSH descriptor names. It also assigns a type to each mapping for classifying these mappings.

### Types of mappings

### Procedure of building mappings

#### Build mappings among MeSH concept names

A concept has several names just in MeSH. So the program needs to build the mappings between various names of a concept and its descriptor name at first.

**Step 1: Handle the concepts in file “supp2014.xml”. Assign the type “8c” to the mappings between CAS names and descriptor names. Assign the type “8b” to the mappings between various concept names and descriptor names.**

Code: Line 162 File premap.pl

$conn->do("REPLACE INTO TERMMAP SELECT A.SUI, A.STR, B.STR MAPPING, IF(A.TTY='CE', '8b', '8c') FROM MRCONSO A, MRCONSO B WHERE A.TTY IN ('CE', 'N1') AND A.CODE=B.CODE AND B.TTY='NM'");

Comment: “A.CODE=B.CODE” indicates that all of the names are from the same descriptor record. If TTY is “CE”, the string is a usual concept name. If TTY is “N1”, the string is a name provided by “[Chemical Abstracts Service](https://www.cas.org/)”. “NM” indicates that the string is a descriptor name in file “supp2014.xml” just like “MH” in file “desc2014.xml”.

**Step 2: Handle the concepts in file “supp2014.xml”. Assign the type “8a” to the mappings when descriptor names are mapped to themselves.**

Code: Line 171 File premap.pl

$conn->do("REPLACE INTO TERMMAP SELECT SUI, STR, STR MAPPING, '8a' FROM MRCONSO WHERE TTY='NM'");

**Step 3: Handle the concepts in file “desc2014.xml”. Assign the type “8d” to the mappings between CAS names and descriptor names.**

Code: Line 177 File premap.pl

$conn->do("REPLACE INTO TERMMAP SELECT A.SUI, A.STR, B.STR MAPPING, '8d' FROM **MRCONSO** A, MRCONSO B WHERE A.TTY='N1' AND A.CODE=B.CODE AND B.TTY='MH'");

Comment: Refer to the comment in Step 1.

Step 4: Assign the type “6” to the mappings between concept names and MeSH ATXs.