Clinical Measurement Ontology (CMO)– Mappings, Thesaurus, Concept and Triple/Publication generation

This activity for the CMO datasource is to ensure all relevant concepts are added to the thesaurus and additional synonyms are added. In addition triples can be created from this source.

The source file is an ontology with the OBO format.

First we need to create the mapping collection, based on inputfile:

clinical_measurement.obo

Understanding the source file

clinical_measurement.obo is a standard OBO file, containing 2682 terms; example:

format-version: 1.2
date: 07:04:2016 17:21
saved-by: jesmith
auto-generated-by: OBO-Edit 2.3.1
synonymtypedef: curation_status "curation_status" RELATED
default-namespace: Clinical_Measurement.ontology
ontology: cmo
data-version: 2.37

[Term]
id: CMO:0001601
name: saccharin intake volume to total fluid intake volume ratio
def: "A calculated measurement of saccharin intake in which the volume, that is, the size of the three dimensional space occupied by a solution of saccharin, a cyclic imine of 2-sulfobenzoic acid which is 500 times sweeter than sugar and used as a nonnutritive sweetener, consumed in a specified period of time is divided by the total volume of all fluids consumed within that period, and presented as a ratio, fraction, quotient or percentage." [Multiple_Dictionaries:http\://www.thefreedictionary.com/]
synonym: "intake volume of saccharin to total fluid intake volume ratio" EXACT []
synonym: "ratio of the volume of saccharin consumed to the total volume of fluid consumed" EXACT []
synonym: "saccharin intake volume to total saccharin and water intake volume ratio" RELATED []

synonym: "saccharin preference score" RELATED []
xref: CHEBI:32111
is_a: CMO:0001600 ! calculated saccharin drink intake volume
relationship: part_of CMO:0001817 ! calculated total pancreatic islet beta cell weight
created_by: JSmith
creation_date: 2013-06-10T13:34:33Z

key	mandatory/optional	type	example
id	M	single value - string :	CMO:0001601
name	M	single value - string	saccharin intake volume to total fluid intake volume ratio
def	O	single value - string	"A calculated measurement of saccharin intake in which the volume, that is, the size of the three dimensional space occupied by a solution of saccharin, a cyclic imine of 2-sulfobenzoic acid which is 500 times sweeter than sugar and used as a nonnutritive sweetener, consumed in a specified period of time is divided by the total volume of all fluids consumed within that period, and presented as a ratio,

			fraction, quotient or percentage." [Multiple_Dictionaries: http\://www.thefreedictionary.com/]
synonym	O	array	{ "intake volume of saccharin to total fluid intake volume ratio" , "ratio of the volume of saccharin consumed to the total volume of fluid consumed", "saccharin intake volume to total saccharin and water intake volume ratio" ,"saccharin preference score"}
xref	O	array	{CHEBI:32111}
is_a	O	array (,) - string	{CMO:0001600}
relationship	O	key , value pair, space separated	part_of CMO:0001817
creation_date	O	dd:mm:yyyy	02:08:2016
created_by	O	string	JSmith

Thesaurus and Concept generation

For this activity, the mapping collection provides:

- New terms / new uuids - insertion in Solr and Concept Collection
- Synonyms - insertion in Solr
- Defintions - insertion in Concept collection

Also triples will be generated from the same CMO mapping collections

CMO mapping collection

key	Mandatory/ Optional	Example	Use	Modifications
id	M	CMO:0001601	synonym	lower case
name	M	saccharin intake volume to total fluid intake volume ratio	preferred term	lower case
def	O	"A calculated measurement of saccharin intake in which the volume, that is, the size of the three dimensional space occupied by a solution of saccharin, a cyclic imine of 2-sulfobenzoic acid which is 500 times sweeter than sugar and used as a nonnutritive sweetener, consumed in a specified period of time is divided by the total volume of all fluids consumed within that period, and presented as a	String - single value ; used as definition in concept collection.	

		ratio, fraction, quotient or percentage." [Multiple_Dictionaries:http\://www.thefreedictionary.com/]		
synonym	O	intake volume of saccharin to total fluid intake volume ratio	synonym	lower case
is_a	O	{CMO:0005329}	Relation (triple generation)	
xref	O	{CHEBI:32111}	Relation (triple generation)	
part_of	O	{CMO:0001817}	Relation (triple generation)	

1) *Is the data source an authority?*

Yes, this an authority for clinical measurements.

a. *If so, do we expect SOLR create events? (new GI)*

Yes, New GI's can be created

b. *If so, do we expect SOLR update events? (additional synonym to existing UUID)*

Yes, new synonyms can be expected.

LOGIC overview

1. identify if a new uuid needs to be created
2. determine preferred term, semantic type
3. create new concept / uuid to Solr
4. add synonyms to Solr
5. Create concept in Concept collection

- determine preferred term
- determine definition

1. Identify if new UUID needs to be created

In the CMO mappings collection, iterate through the records.

1. Verify if the <id> already exists in solr. Example: term: “cmo:0000062” .

If the <id> exists for 1 UUID/GI, verify if the following synonyms in the mappings record exists in Solr for that GI/UIID:

(<synonyms>, <name>).

If so , proceed to next mapping document. If not, add the synonyms according to logic described in “add synonym” section.

1b. If <id> exists for more than 1 GI/UUID, Log the <id> entry in the error log (duplicate) and proceed to the next mapping record

2. If <id> does not exist in Solr, Verify if the concept already exists in Solr by <name> or <synonym> according to the following query:

Term: <name> && semanticcategory: Phenomena, or

Term: <synonym> && semanticcategory: Phenomena

If the document does not contain <name> OR <synonym> , then we cannot establish the concept. Create new UUID/GI for this entry (step 3)

If <name> or <synonym> does exist for this entry, Use the following queries to identify a match:

a. term: <name>|<synonym> && semanticcategory: “Phenomena”
    i. If 1 uuid/GI exists, add the synonyms according to logic described in “add synonym” section.
    ii. if multiple UUID/GIs are returned, add the synonyms according to logic described in “add synonyms” section to each of the UUID/GI’s

3. If the concept does not yet exist in Solr, create uuid and add to Solr:

Add Preferred Term: <name>

id":
"**term**": "<name>",
"**source**": "cmo",
"**knowledgebase**": "cmo",
"**semantictype**": 34
semanticcategory: Phenomena
"**gi**": "generate",
"**preferred**": "T",
"**_version_**":

Add Synonyms: <id> | <synonym>

id":
"**term**": "<id> | <synonym> ",
"**source**": cmo",
"**knowledgebase**": "cmo_id" (when <id>) | "cmo" when <synonym>
"**semantictype**": 34,
semanticcategory: Phenomena

"**gi**": "same as <name>(gi)",
"**preferred**": "F",
"**_version_**":

4. If the Concept / UUID does exist in Solr, verify existing synonyms (<name>, <synonym>) from the mapping collection and add new synonyms which are not present in Solr.

If there are more than 1 hits in Solr with a different UUID, create a "duplicate" error log entry and do not create synonyms:

id":
"**term**": "<id> | <synonym> | ",
"**source**": "cmo",
"**knowledgebase**": "cmo_id" (when <id>) | else "cmo"
"**semantictype**": 34,

semanticcategory: Phenomena

"**gi**": "same as <name>(gi)",
"**preferred**": "F",
"**_version_**":

Concept Generation in concept collection

Trigger for concept generation in Mongo collection

Source	Knowledgebase	Semantic Type	Semantic Group	Preferred T/F
cmo	cmo	34	Phenomena	T

when Solr record contains source=cmo && Preferred = T -> create a new Concept entry in Mongo.

For that gi, identify key to the mapping collection to get the definition:
The term of the Solr record (kb=cmo_id) is the <id> of the CMO mapping collection:

Identify Source-Mongo document and KEY

Solr Criteria	Mongo Document Key	Concept attribute	Value type	Error handling
knowledgebase_id	term	<id>	string	do not create concept

So the term of cmo_id (example: cmo:0000062 is the key to the Mongo mapping collection to fetch the following data to create the concept record:

Mapping criteria

Attribute	Source key	Concept target key
name	name	name
definition	<def>	definition
semantictype		34

Access parameter

Source Mongo collection	RD	RT
cmo	0	111

Triple Generation

There are 3 types of triples generated:

Triple 1: <id> - is a - <is_a>

Triple 2: <id> - part of <part_of>

Triple 3 <id> - associated_with - <xref>

Triple 1:

For each element in the is_a array, create a triple:

Subject: <id> ;

solr query: term: <id> && semanticcategory:Phenomena

Create a triple for each uuid/GI

Predicate (constant): is a

Object: for each element in the array <is_a>

Solr query: term: <is_a> && semanticcategory: Phenomena

Create a triple for each uuid/GI

Measure: none

Triple 2:

For each element in "part_of" key , create a triple:

Subject: <id> ;

solr query: term: <id> && semanticcategory:Phenomena

Create a triple for each uuid/GI

Predicate (constant): is a

Object: for each element in the array <part_of>

Solr query: term: <is_a> && semanticcategory: Phenomena

Create a triple for each uuid/GI

Measure: none

Triple 3:

For each element in "xref" key , create a triple:

Subject: <id> ;

solr query: term: <id> && semanticcategory:Phenomena

Create a triple for each uuid/GI

Predicate (constant): is a

Object: for each element in the array <part_of>

Solr query: term: <xref>

Create a triple for each uuid/GI

Publication Generation

For all triples generated out of a mapping record (so per <id>) 1 publication is created with the following characteristics:

Measure: Publicationtype= ontology

Scientific value = 7

Institution= Animal QTLdb

Publicationtitle=CMO/<name>

Publication ID = CMO/<id>

Publicationdate= <creation_date>

Publicationsource: CMO/<name>

URL : http://bioportal.bioontology.org/ontologies/CMO/<id>

Source Mongo collection	RD	RT
cmo	0	111

Data Processing System