Vertebrate Trait Ontology (VTO)– Mappings, Thesaurus, Concept and Triple/Publication generation

This activity for the VTO 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:

VT_v6.9.obo, found at development server: /inputdata/ontologies/vT_v6.9.obo

Understanding the source file

VT_v6.9.obo is a standard OBO file, containing 3489 terms; example:
format-version: 1.2
date: 02:08:2016 15:11
saved-by: caripark
auto-generated-by: OBO-Edit 2.1
default-namespace: Trait.ontology
RGDformat-version: 1.2

[Term]
id: VT:0000062
name: heart muscle trabeculae morphology trait
def: "Any measurable or observable characteristic related to the shape, structure, color or pattern of the supporting bundles of muscular fibers lining the walls of the heart." [ISBN:0-683-40008-8, **MP:0002189**]
synonym: "cardiac muscle trabeculae morphology trait" EXACT []
synonym: "myocardial trabeculae morphology trait" EXACT []
synonym: "trabeculae carneae morphology trait" EXACT []
is_a: VT:0005329 ! heart muscle morphology trait
xref: GO:0001503 "heart muscle trabeculae"

key	mandatory/optional	type	example
id	M	single value - string :	VT:0000062
name	M	single value - string	heart muscle trabeculae morphology trait
def	O	single value - string	"Any measurable or observable characteristic related to the shape, structure, color or pattern of the supporting bundles of muscular fibers lining the walls of the heart." [ISBN:0-683-40008-8, **MP:0002189**]
synonym	O	array	cardiac muscle trabeculae morphology trait, myocardial trabeculae morphology trait, trabeculae carneae morphology trait
xref	O	array	{GO:0001503 }
is_a	O	array (,) - string	{VT:0005329}
date	From file header <date>	dd:mm:yyyy	02:08:2016

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

In a next activity, also triples will be generated from the same VT mapping collections

VT mapping collection

key	Mandatory/ Optional	Example	Use	Modifications
id	M	VT:0000062	synonym	lower case
name	M	heart muscle trabeculae morphology trait	preferred term	lower case
def	O	"Any measurable or observable characteristic related to the shape, structure, color or pattern of the supporting bundles of muscular fibers lining the walls of the heart."	String - single value ; used as definition in concept collection.	Remove def-reference - between square brackets
def_xref	O	[ISBN:0-683-40008-8, **MP:0002189**]	Synonym - identify existing concepts	Array of strings - {[ISBN:0-683-40008-8, MP:0002189}
synonym	O	N-acetyl-p-aminophenolcardiac muscle trabeculae	synonym	lower case

		morphology trait, myocardial trabeculae morphology trait, trabeculae carneae morphology trait		
is_a	O	{VT:0005329}	Relation (triple generation)	
xref	O	{GO:0001503 }	synonym	

1) *Is the data source an authority?*

Yes, this an authority for traits / phenotypes. Some entries map onto mammelian phenotype ontology concepts (mp:xxxxxxx) or GO concepts (go:xxxxxxxx)

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 VTO mappings collection, iterate through the records.

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

(<x-ref>, <def_xref>).

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 <x-ref> and <def_xref>.

If the document does not contain <x-ref> or <def_xref> , then we cannot establish the concept. Create new UUID/GI for this entry (step 3)

If <x-ref> or <def_xref> does exist for this entry, Use the following queries to identify a match:

a. term: <x-ref>|<def_xref> && semanticcategory: “Disorders”
    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**": "vt",
"**knowledgebase**": "vt",
"**semantictype**": 33
semanticcategory: Disorders
"**gi**": "generate",
"**preferred**": "T",
"**_version_**":

Add Synonyms: <id> | <xref> |<def_xref> | <synonym>
id":
"**term**": "<id> | <def_xref> | <synonym> | <xref>",
"**source**": vt",
"**knowledgebase**": "vt_id" (when <id>) | "vt" when <xref>|<def_xref>|<synonym>
"**semantictype**": 33,
semanticcategory: Disorders

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

4. If the Concept / UUID does exist in Solr, verify existing synonyms (<id>,<xref>, <def_xref>,<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> | <def_xref> | <synonym> | <xref>",
"**source**": "vt",
"**knowledgebase**": "vt_id" (when <id>) | else "vt"
"**semantictype**": 33,
semanticcategory: Disorders

"**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
vt	vt	33	Disorders	T

when Solr record contains source=vt && 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=vt_id) is the <id> of the chebi 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 vt_id (example: vt: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		33

Access parameter

Source Mongo collection	RD	RT
chebi	0	102

Triple Generation

There are 2 types of triples generated:

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

Triple 1:
For each element in the is_a array, create a triple:

Subject: <id> ;

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

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: Disorders

Create a triple for each uuid/GI

Measure: none

Publication Generation

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

Data Processing System

Measure: Publicationtype= ontology

Scientific value = 7

Institution= Animal QTLdb

Publicationtitle=VT/<name>

Publication ID = VT/<id>

Publicationdate= <date>

Publicationsource: VT/<name>

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

Source Mongo collection	**RD**	**RT**
vt	0	102