Chebi– Mappings, Thesaurus, Concept and Triple/Publication generation

This activity for the Chebi 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:

chebi.obo, found at development server: /inputdata/ontologies/chebi.obo

Understanding the source file

chebi.obo is a standard OBO file, containing 103478 terms; example:

format-version: 1.2
data-version: 142
date: 01:08:2016 12:55
saved-by: chebi
default-namespace: chebi_ontology
remark: ChEBI subsumes and replaces the Chemical Ontology first
synonymtypedef: IUPAC_NAME "IUPAC NAME"
synonymtypedef: FORMULA "FORMULA"
synonymtypedef: SMILES "SMILES"
synonymtypedef: InChI "InChI"
synonymtypedef: InChIKey "InChIKey"
synonymtypedef: BRAND_NAME "BRAND NAME"
synonymtypedef: INN "INN"
remark: developed by Michael Ashburner & Pankaj Jaiswal.
remark: Author: ChEBI curation team
remark: ChEBI Release version 142
remark: For any queries contact chebi-help@ebi.ac.uk
ontology: chebi

[Term]
id: CHEBI:46195

name: paracetamol
alt_id: CHEBI:2386
alt_id: CHEBI:46191
def: "A member of the class of phenols that is 4-aminophenol in which one of the hydrogens attached to the amino group has been replaced by an acetyl group." []
synonym: "N-acetyl-p-aminophenol" RELATED [ChEBI]
synonym: "paracetamol" RELATED INN [WHO MedNet]
synonym: "acetaminophene" RELATED [ChEBI]
synonym: "4-(Acetylamino)phenol" RELATED [ChemIDplus]
synonym: "Paracetamol" EXACT [KEGG COMPOUND]
synonym: "paracetamolum" RELATED INN [ChemIDplus]
synonym: "Acenol" RELATED [ChemIDplus]
synonym: "Acetaminofen" RELATED [ChemIDplus]
synonym: "acetaminofen" RELATED [ChemIDplus]
synonym: "APAP" RELATED [DrugBank]
synonym: "p-Acetylaminophenol" RELATED [ChemIDplus]
synonym: "Panadol" RELATED BRAND_NAME [ChEBI]
synonym: "paracetamol" RELATED INN [WHO MedNet]
synonym: "p-acetamidophenol" RELATED [NIST Chemistry WebBook]
synonym: "4'-hydroxyacetanilide" RELATED [ChemIDplus]
synonym: "p-hydroxyphenolacetamide" RELATED [NIST Chemistry WebBook]
synonym: "N-(4-hydroxyphenyl)acetamide" EXACT IUPAC_NAME [IUPAC]
synonym: "p-acetaminophenol" RELATED [NIST Chemistry WebBook]
synonym: "Tylenol" RELATED BRAND_NAME [KEGG DRUG]
synonym: "p-hydroxyacetanilide" RELATED [NIST Chemistry WebBook]
synonym: "paracetamol" RELATED INN [KEGG DRUG]
synonym: "Acetaminophen" RELATED [KEGG COMPOUND]
synonym: "4-acetamidophenol" RELATED [NIST Chemistry WebBook]
synonym: "C8H9NO2" RELATED FORMULA [ChEBI]
synonym: "151.16260" RELATED FORMULA [ChEBI]
synonym: "InChI=1S/C8H9NO2/c1-6(10)9-7-2-4-8(11)5-3-7/h2-5,11H,1H3,(H,9,10)" RELATED InChI [ChEBI]
synonym: "RZVAJINKPMORJF-UHFFFAOYSA-N" RELATED InChIKey [ChEBI]
synonym: "CC(=O)Nc1ccc(O)cc1" RELATED SMILES [ChEBI]
xref: PMID:16716555 "Europe PMC"
xref: PMID:21108564 "Europe PMC"
xref: KEGG:D00217

xref: LINCS:LSM-5533
xref: DrugBank:DB00316
xref: PMID:22770225 "Europe PMC"
xref: PMID:11304127 "Europe PMC"
xref: PMID:7602118 "Europe PMC"
xref: Reaxys:2208089 "Reaxys"
xref: Wikipedia:Acetaminophen
xref: PMID:25128677 "Europe PMC"
xref: HMDB:HMDB01859
xref: MetaCyc:CPD-7669
xref: CAS:103-90-2 "KEGG COMPOUND"
xref: KEGG:C06804
xref: CAS:103-90-2 "NIST Chemistry WebBook"
xref: CAS:103-90-2 "ChemIDplus"
xref: Beilstein:2208089 "Beilstein"
xref: PDBeChem:TYL
relationship: has_role CHEBI:50629
relationship: has_role CHEBI:50630
relationship: has_role CHEBI:35481
relationship: has_role CHEBI:35493
relationship: has_role CHEBI:35475
is_a: CHEBI:33853
relationship: has_role CHEBI:73263
relationship: has_role CHEBI:35703
relationship: has_role CHEBI:78298
relationship: has_role CHEBI:85234
relationship: has_functional_parent CHEBI:17602
is_a: CHEBI:22160

key	mandatory/optional	type	example
id	M	single value - string : KEY TO MAPPING DOC	CHEBI:46195

alt_id	O	array	{CHEBI:2386, CHEBI:46191 }
name	M	single value - string	paracetamol
def	O	single value - string	"A member of the class of phenols that is 4-aminophenol in which one of the hydrogens attached to the amino group has been replaced by an acetyl group."
synonym	O	array of key value pairs :NOTE ; 3 special synonyms need to be singled out as separate mapping keys: inchi, inchikey, smiles - see table 2	{"N-acetyl-p-aminophenol", [ChEBI] ;"RZVAJINKPMORJF-UHFFFAOYSA-N", InChIKey}
xref	O	array (;)- of key:value pairs	{ KEGG:D00217 ; LINCS:LSM-5533 ; DrugBank:DB00316 }
is_a	O	array (,) - string	{CHEBI:33853, CHEBI:22160}
relationship	O	Array (;) - key, value pairs	has_role, CHEBI:73263 ; has_functional_parent , CHEBI:17602
date	From file header <date>	dd:mm:yyyy	01:08:2016

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 Plant Ontology mapping collections

Chebi mapping collection

key	Mandatory/ Optional	Example	Use	Modifications
id	M	CHEBI:46195	synonym	lower case
alt_id	O	CHEBI:2386	synonym	lower case
name	M	paracetamol	preferred term	lower case
def	O	"A member of the class of phenols that is 4-aminophenol in which one of the hydrogens attached to the amino group has been replaced by an acetyl group."	String - single value ; used as definition in concept collection.	
synonym	O	N-acetyl-p-aminophenol	synonym	lower case
synonym_inchikey	O	RZVAJINKPMORJF-UHFFFAOYSA-N	synonym	
synonym_inchi	O	InChI=1S/C8H9NO2/c1-6(10)9-7-2-4-8(11)5-3-7/h2-5,11H,1H3,(H,9,1	synonym	

		0)		
synonym_smiles	O	CC(=O)Nc1ccc(O)cc1	synonym	
is_a	O	CHEBI:22160	Relation (triple generation)	
xref	O	{KEGG:C06804, CAS:103-90-2 }	synonym	remove keys -> solr query term: c0007180 D018879 ; only use for selected keys (see below)
relationship	O	{has_role, CHEBI:78298 ; has_role, CHEBI:85234 ; has_functional_parent, CHEBI:17602}	Relation (triple generation)	

1) *Is the data source an authority?*

Yes, this an authority for chemicals (small molecules)

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 Chebi mappings collection, iterate through the records.

1. Verify if the <id> already exists in solr. Example: term: “chebi:32636” .

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

(<alt_id>, <synonym_inchikey>).

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 <synonym_inchikey>.

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

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

a. term: <synonym_inchikey> && semanticcategory: “Chemicals & drugs”
    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**": "chebi",
"**knowledgebase**": "chebi",
"**semantictype**": 104
semanticcategory: Chemicals & drugs
"**gi**": "generate",
"**preferred**": "T",
"**_version_**":

Add Synonyms: <id> , <alt_id>, <synonym_inchikey>, <xref value>, (where xref key= {HMDB, Drugbank, CAS, KEGG, Metacyc}

id":
"**term**": "<id> | <alt_id> | <synonym_inchikey> | <xref>",
"**source**": "chebi",
"**knowledgebase**": "chebi_id" (when <id>) | "inchikey" (when <synonym_inchikey>) | <xref key> (when xref) | "chebi", when <alt_id>
"**semantictype**": 104,
semanticcategory: Chemicals & drugs

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

4. If the Concept / UUID does exist in Solr, verify existing synonyms (<id>, <alt id>, <synonym_inchikey> <xref>{hmdb, drugbank, CAS, KEGG, Metacyc}) 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" erro log entry and do not create synonyms:

id":
"**term**": "<id> | <alt_id> | <synonym_inchikey> | <xref>",
"**source**": "chebi",
"**knowledgebase**": "chebi_id" (when <id>) | "inchikey" (when <synonym_inchikey>) | <xref key> (when xref) | "chebi", when <alt_id>
"**semantictype**": 104,
semanticcategory: Chemicals & drugs

"**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
chebi	chebi_id	104	Chemicals & rugs	T

when Solr record contains source=chebi && 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=chebi_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 chebi_id (example: chebi:36081 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		104
Measure: inchi	synonym_inchi	measure
Measure: smiles	synonym_smiles	measure

Access parameter

Source Mongo collection	RD	RT
chebi	0	101

Triple Generation

There are 2 types of triples generated:

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

Triple 2 : <id> - <mapped predicated based on Relation-key> - <relation value>

Triple 1:

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

Subject: <id> ;

solr query: term: <id> && semanticcategory: Chemicals & drugs

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: Chemicals & drugs

Create a triple for each uuid/GI

Measure: none

Triple 2:

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

Subject: <id> ;

solr query: term: <id> && semanticcategory: Chemicals & drugs

Create a triple for each uuid/GI

Predicate: according to the mapping table based on Relationship: key

Relationship : Key	**Predicate**
has_part	contains
is_conjugate_base_of	converts_to
is_conjugate_acid_of	converts_to
is_tautomer_of	converts_to
is_enantiomer_of	variant_of
has_functional_parent	functionally_related_to
has_parent_hydride	variant_of
is_substituent_group_from	derivative_of
has_role	classified_as

Object: for each element in the array <relationship: value>

Solr query: term: <relationship value> && semanticcategory: Chemicals & drugs

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:

Measure: Publicationtype= ontology

Scientific value = 7

Institution= EMBL-EBI

Publicationtitle=Chebi/<name>

Publicatin ID = Chebi/<id>

Publicationdate= <date>

Publicationsource: Chebi/<name>

URL : http://www.ebi.ac.uk/chebi/searchId.do?chebiId=<id>

Source Mongo collection	RD	RT
chebi	0	101