Tutorial Slides

Francisco M. Couto

March 28, 2021

Biomedical Text Processing using Semantics tutorial
at the 43rd European Conference on Information Retrieval (ECIR 2021)

Source:
F. Couto, Data and Text Processing for Health and Life Sciences.
No. 1137 in Advances in Experimental Medicine and Biology, Springer, 2019
http://labs.rd.ciencias.ulisboa.pt/book/

Semantics Introduction

Lack of use of standard nomenclatures

- different labels (synonyms, acronyms)
- sharing the same label (homonyms)
- requires sense disambiguation to select the correct meaning

Disease acronym *ATS* may represent

- *Andersen-Tawil syndrome*
- or the *X-linked Alport syndrome*

Solution: ontologies and semantic similarity

What?

In 1993 definition of ontology:

> an explicit specification of a conceptualization

In 1997 and 1998 refined to:

> a formal, explicit specification of a shared conceptualization

Conceptualization

an abstract view of the concepts

and the relationships of a given domain

Shared conceptualization

a group of individuals agree (common agreement)

Specification is a representation of that conceptualization

using a given language.

needs to be formal and explicit

so computers can deal with it

Languages

Web Ontology Language (OWL)

most common languages to specify ontologies

Open Biomedical Ontology (OBO)

principles to ensure high quality, formal rigor
and interoperability between other OBO ontologies

Concepts are defined as OWL classes

that may include multiple properties, such as labels

official name, acronyms, exact synonyms, and even related terms

Class *malignant hyperthermia*

synonym *anesthesia related hyperthermia*.

Andersen-Tawil syndrome and *X-linked Alport syndrome*

share *ATS* as an exact synonym

Formality

Different levels of formality
    such as controlled vocabularies, taxonomies
    and thesaurus may include logical axioms.

Controlled vocabularies are list of terms
    without specifying any relation between them

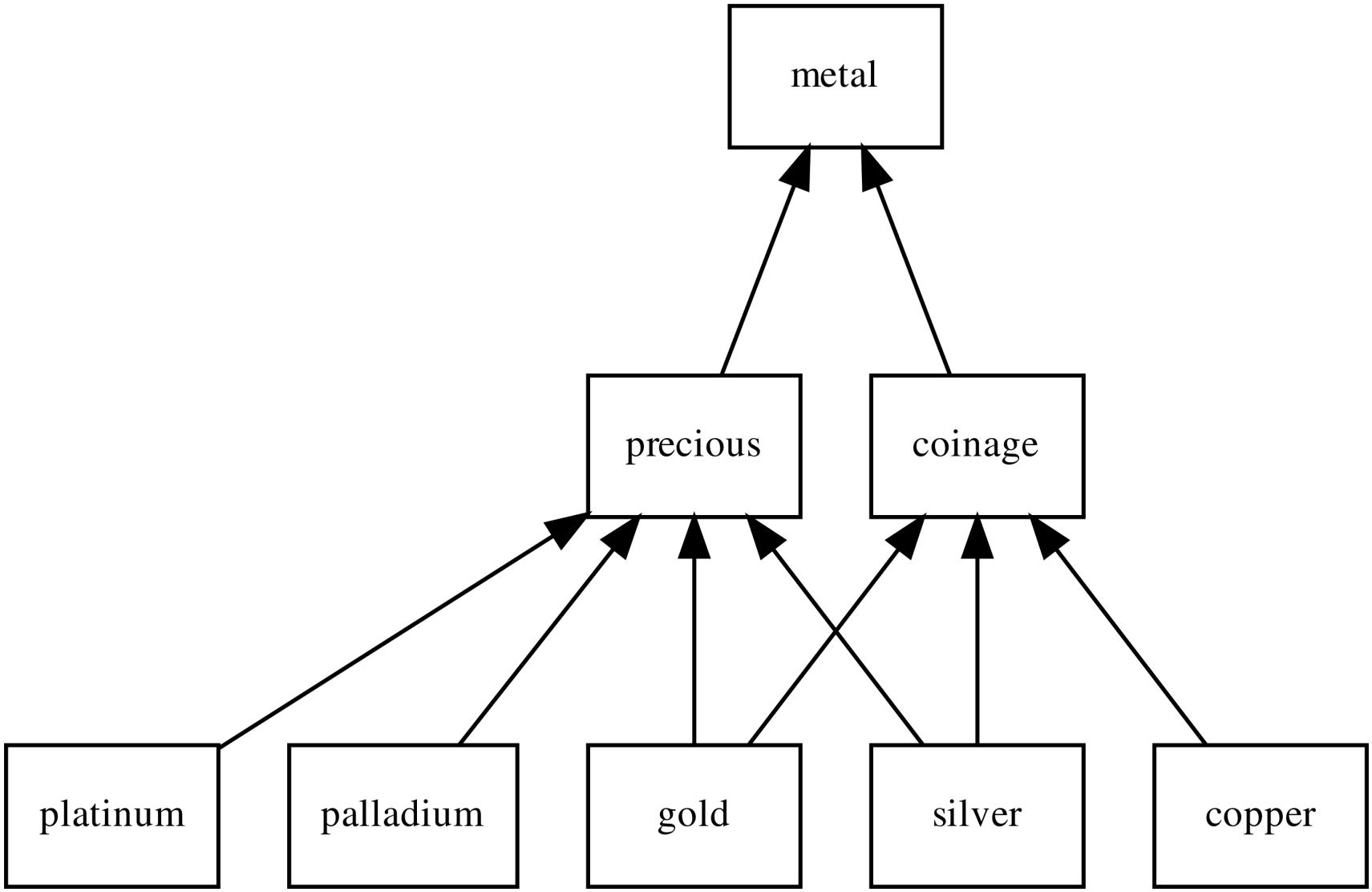
Taxonomies are controlled vocabularies
    that include subsumption relations
    malignant hyperthermia is a *muscle tissue disease*

is-a or subclass relations

- are normally the backbone of ontologies.
- include multiple inheritance
- organized as a directed acyclic graphs (DAG)

Thesaurus

- includes other types of relations besides subsumption
- *caffeine* has role *mutagen*.



DAG representing a classification of metals with multiple inheritance

Gold related documents

Find texts related to *gold*

a corpus with one distinct document mentioning each metal

except for *gold* that no document mentions

which documents should we read first?

silver is probably the most related

shares two parents, *precious* and *coinage*.

platinum, *palladium* or *copper*?

- depends on our information need
- previous searches or reads

Last searches were *coinage*

- *copper* is probably the second-most related

Importance of these semantic resources

- development of the knowledge graph by Google

Where?

BioPortal (Dec. 2018)

- more than 750 ontologies
- more than 9 million classes (2018)

Search for *caffeine*

- large list of ontologies that define it
- conceptualizations of *caffeine* in different domains
- alternative perspectives
- Interoperability property with links to similar classes

OBO initiative

- tackle this somehow disorderly spread of definitions
- each OBO ontology covers a clearly specified scope

OBO ontologies

Success of Gene Ontology (GO)

- describe molecular function, biological process and cellular component
- gene-products for different species

Disease Ontology (DO)

- human disease terms
- phenotype characteristics
- and related medical vocabulary disease concepts

Chemical Entities of Biological Interest (ChEBI)

- classification of molecular entities
- with biological interest
- focus on small chemical compounds

Popular controlled vocabularies

International Classification of Diseases (ICD)
- by World Health Organization (WHO)
- generic clinical terms

Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT)
- highly comprehensive and detailed

Medical Subject Headings (MeSH)
- classifying biomedical and health-related information and documents

Unified Medical Language System (UMLS)
- large resource integrate most biomedical vocabularies
- 2015AB release more than three million concepts

Ontobee

- repository of ontologies (most OBO ontologies)
- 187 ontologies (Dec. 2018)

Outside the biomedical domain

- W3C SWEO Linking Open Data community project
- W3C Library Linked Data Incubator Group

How?

Find ontology home page
- download the most recent release
- the original format
- select the subset of the ontology

ChEBI provides three versions:
- LITE, CORE and FULL

If not interested in chemical data and structures
- that is available in CORE
- LITE is probably the best solution
- may miss synonyms from FULL version

OWL

OWL language prevailing language to represent ontologies

OWL extends RDF Schema (RDFS)
with more complex statements using description logic

RDFS is an extension of RDF
with additional statements
such as class-subclass or property-subproperty relationships

RDF is a data model
- stores information in statements
- represented as triples: subject, predicate and object

RDF data encoded using Extensible Markup Language (XML)
- named RDF/XML

XML is a self-descriptive mark-up language
- composed of data elements

XML example

caffeine is a drug

may treat the condition of sleepiness

without being an official treatment:

```
<treatment category="non-official">
  <drug>caffeine</drug>
  <condition>sleepiness</condition>
</treatment>
```

Hierarchical structure of data elements:

`<` - new data element

`</` - data element will end

property `category` with value `non-official`

Large XML files are almost unreadable by humans

N3 and Turtle
- legible encoding languages for RDF

Most biomedical ontologies in OWL using XML encoding

URI

The Uniform Resource Identifier (URI)

standard global identifier of classes

class `caffeine` in ChEBI identified by:

`http://purl.obolibrary.org/obo/CHEBI_27732`

A URI is a URL if we open it in a web browser

and obtain a resource describing that class

XPath

XPath (XML Path Language)

a powerful tool to extract information from XML and HTML documents following their hierarchical structure

Query Examples

`//dbReference`

elements of type `dbReference` descendants of something

Using `https://www.uniprot.org/uniprot/P21817.xml`

```
<dbReference type="NCBI Taxonomy" id="9606"/>
...
<dbReference type="PubMed" id="27586648"/>
```

`/entry//dbReference`

equivalent to the previous query
specifying `dbReference` descendants of `entry`

`/entry/reference/citation/dbReference`

equivalent to the previous query
specifying the full path

```
//dbReference/*
```

any child elements of `dbReference`

```
<property type="protein sequence ID" value="AAA60294.1"/> ... <
  property type="match status" value="5"/>
```

```
//dbReference/property[1]
```

first `property` of each `dbReference`

```
<property type="protein sequence ID" value="AAA60294.1"/> ... <
  property type="entry name" value="MIR"/>
```

```
//dbReference/property[2]
```

second `property` of each `dbReference`

```
<property type="molecule type" value="mRNA"/> ... <property type=
  "match status" value="5"/>
```

```
//dbReference/property[3]
```

third `property` of each `dbReference`

```
<property type="molecule type" value="Genomic_DNA"/> ... <
  property type="project" value="UniProtKB"/>
```

```
//dbReference/property/@type
```

all type attributes of `property`

```
type="protein sequence ID" type="molecule type" type="protein
  sequence ID" ... type="entry name" type="match status"
```

```
//dbReference/property[@type="protein sequence ID"]
```

the previous `property` elements

with attribute type equal to *protein sequence ID*

```
<property type="protein sequence ID" value="AAA60294.1"/> ... <
  property type="protein sequence ID" value="ENSP00000352608"/>
```

```
//dbReference/property[@type="protein sequence ID"]/@value
```

string of each attribute *value* of previous `property` elements

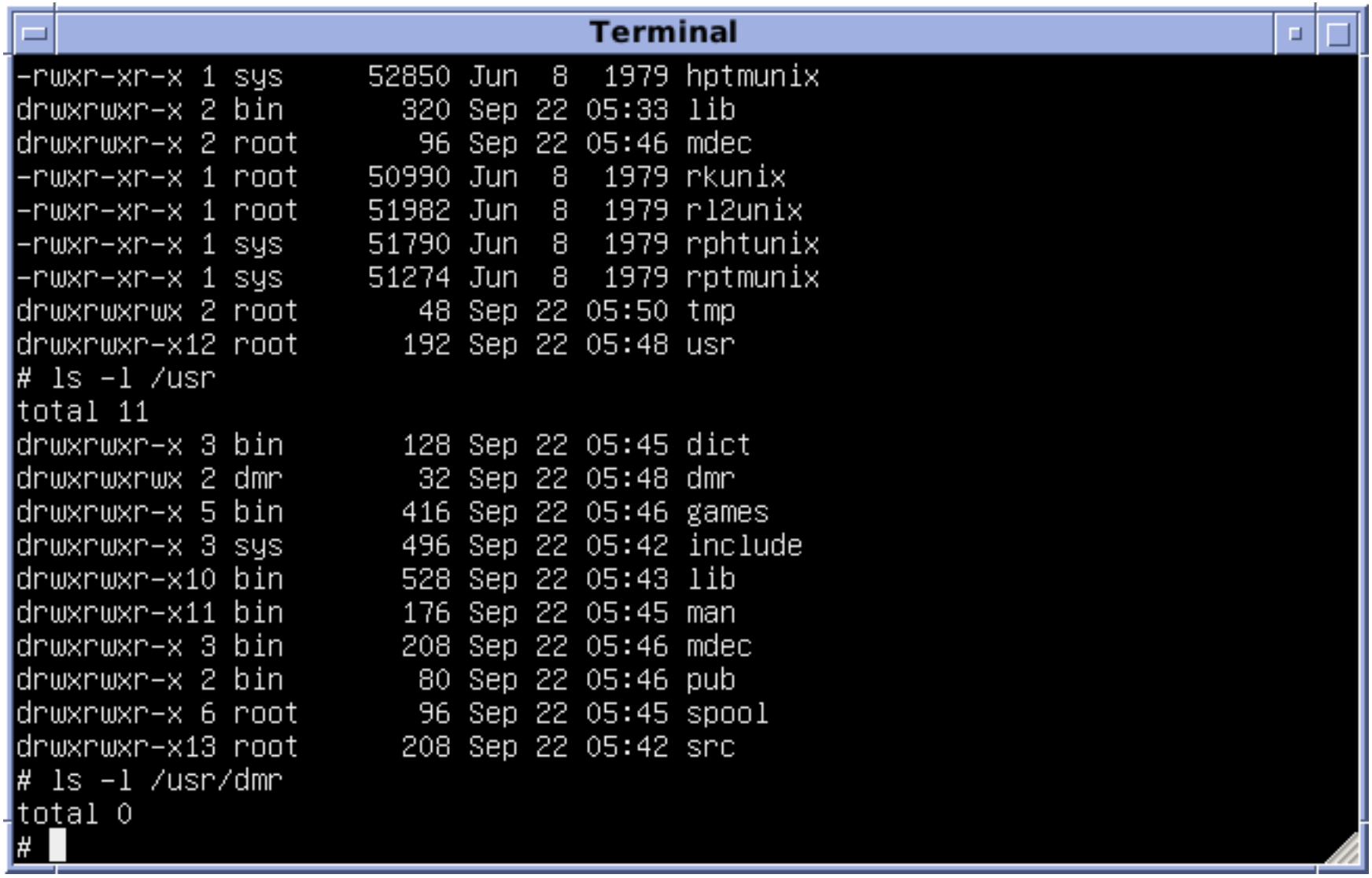
```
value="AAA60294.1" value="AAC51191.1" ... value="ENSP00000352608"
```

```
//sequence/text()
```

the contents inside `sequence`

```
MGDAEGEDEVQFLRTDDEVVLQCSATVLKEQLKLCLAAEGFGNRLCFLEPTSNAQNVPPD
...
LEEHNLANYMFFLMYLINKDETEHTGQESYVWKMYQERCWDFFPAGDCFRKQYEDQLS
```

Shell Scripting



Screenshot of a Terminal application

(Source: `https://en.wikipedia.org/wiki/Unix`)

Command line tools

- `curl`: a tool to download data and text from the web;
- `grep`: a tool to search our data and text;
- `gawk`: a tool to manipulate our data and text;
- `sed`: a tool to edit our data and text;
- `xargs`: a tool to repeat the same step for multiple data items;
- `xmllint`: a tool to search in XML data files using XPath.

- `cat`: a tool to get the content of file;
- `tr`: a tool to replace one character by another;
- `sort`: a tool to sort multiple lines;
- `head`: a tool to select only the first lines.

Check the commands

```
$ curl -O http://labs.rd.ciencias.ulisboa.pt/book/testshell20190520
    .zip
$ unzip testshell20190520.zip
$ chmod u+x testshell.sh
$ ./testshell.sh
```

`-O` saves to file name as remote file
last part of URL

Problems

Watch the following videos during coffee break:

- Linux: `https://youtu.be/4cO9vvbxWUU`
- Windows Subsystem: `https://youtu.be/VNlksnYDE0Y`
- Mobaxterm: `https://youtu.be/yI1No5_o-Kw`
- MacOS: `https://youtu.be/SELYgZvAZbU`

Ontologies

Retrieving both *ChEBI* and *DO* ontologies

OWL files:

```
$   curl -L -O http://purl.obolibrary.org/obo/doid.owl
$   curl -L -O http://purl.obolibrary.org/obo/chebi/chebi_lite.owl.
      gz
$   gunzip chebi_lite.owl.gz
```

`-L` follows redirects

Check links on BioPortal or OBO Foundry

Class label

OWL files use XML syntax

check entities:

```
$ grep '>malignant hyperthermia<' doid.owl
$ grep '>caffeine<' chebi_lite.owl
```

```
<rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string
  ">malignant hyperthermia</rdfs:label>
<rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string
  ">caffeine</rdfs:label>
```

Property label (*rdfs:label*),

inside class definition

Class definition

Retrieve the full class definition with `xmllint`:

```
$ xmllint --xpath "//*[local-name()='label' and text()='malignant
    hyperthermia']/.." doid.owl
```

The XPath query

find the label *malignant hyperthermia*

then `..` the parent element, `Class` element

Semantics of *malignant hyperthermia*

much more than its label:

```
<owl:Class rdf:about="http://purl.obolibrary.org/obo/DOID_8545">
      <rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/
        obo/DOID_0050736"/>
      <rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/
        obo/DOID_66"/>
      <rdfs:subClassOf>
         <owl:Restriction>
            <owl:onProperty rdf:resource="http://purl.obolibrary.
              org/obo/IDO_0000664"/>
            <owl:someValuesFrom rdf:resource="http://purl.
              obolibrary.org/obo/GENO_0000147"/>
         </owl:Restriction>
      </rdfs:subClassOf>
      <obo:IAO_0000115
      ...
       <oboInOwl:hasDbXref rdf:datatype="http://www.w3.org/2001/
         XMLSchema#string">UMLS_CUI:C0024591</oboInOwl:hasDbXref>
      <oboInOwl:hasExactSynonym rdf:datatype="http://www.w3.org
        /2001/XMLSchema#string">anesthesia related hyperthermia</
```

```
        oboInOwl:hasExactSynonym>
      <oboInOwl:hasExactSynonym rdf:datatype="http://www.w3.org
        /2001/XMLSchema#string">malignant hyperpyrexia due to
        anesthesia</oboInOwl:hasExactSynonym>
      <oboInOwl:hasOBONamespace rdf:datatype="http://www.w3.org
        /2001/XMLSchema#string">disease_ontology</
        oboInOwl:hasOBONamespace>
      <oboInOwl:id rdf:datatype="http://www.w3.org/2001/XMLSchema
        #string">DOID:8545</oboInOwl:id>
      <oboInOwl:inSubset rdf:resource="http://purl.obolibrary.org
        /obo/doid#DO_MGI_slim"/>
      <oboInOwl:inSubset rdf:resource="http://purl.obolibrary.org
        /obo/doid#DO_rare_slim"/>
      <oboInOwl:inSubset rdf:resource="http://purl.obolibrary.org
        /obo/doid#NCIthesaurus"/>
      <rdfs:comment rdf:datatype="http://www.w3.org/2001/
        XMLSchema#string">Xref MGI.
OMIM mapping confirmed by DO. [SN].</rdfs:comment>
      <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#
        string">malignant hyperthermia</rdfs:label>
      </owl:Class>
```

Class: malignant hyperthermia

Term IRI: http://purl.obolibrary.org/obo/DOID_8545

Definition: A muscle tissue disease that is characterized by a drastic and uncontrolled increase in skeletal muscle oxidative metabolism, which overwhelms the body's capacity to supply oxygen, remove carbon dioxide, and regulate body temperature. [database_cross_reference: url:http://en.wikipedia.org/wiki/Malignant_hyperthermia][database_cross_reference: url:http://en.wikipedia.org/wiki/Malignant_hyperthermia] [database_cross_reference: url:http://en.wikipedia.org/wiki/Malignant_hyperthermia][database_cross_reference: url:http://en.wikipedia.org/wiki/Malignant_hyperthermia][database_cross_reference: url:http://en.wikipedia.org/wiki/Malignant_hyperthermia]

Annotations

- **database_cross_reference:** ICD9CM:995.86; MESH:D008305; ICD10CM:T88.3; UMLS_CUI:C0024591; ORDO:423; CSP2005:2871-4352; GARD:6964; MTHICD9_2006:995.86; NCI:C84869; OMIM:PS145600
- **has_exact_synonym:** anesthesia related hyperthermia; malignant hyperpyrexia due to anesthesia
- **has_obo_namespace:** disease_ontology
- **http://www.w3.org/2000/01/rdf-schema#comment:** Xref MGI. OMIM mapping confirmed by DO. [SN].
- **id:** DOID:8545
- **in_subset:** DO MGI slim; DO rare slim; NCIthesaurus

Class Hierarchy

Thing
+ disease
+ disease of anatomical entity
+ musculoskeletal system disease
+ muscular disease
+ muscle tissue disease
- distal arthrogryposis
- rippling muscle disease 2
- rippling muscle disease 1
- myostatin-related muscle hypertrophy
- myotonia congenita
+ myopathy
- malignant hyperthermia

Class description of *malignant hyperthermia* in the Human Disease Ontology

(Source: `http://www.ontobee.org/`)

malignant hyperthermia subclass of (specialization)
entries 0050736
and 66 *muscle tissue disease*

malignant hyperthermia a special case of *muscle tissue disease*

Retrieve full class definition of *caffeine*:

```
$ xmllint --xpath "//*[local-name()='label' and text()='caffeine
    ']/.." chebi_lite.owl
```

Semantics of *caffeine* differs from *malignant hyperthermia*

- still share many properties
- such as `subClassOf`

```
<owl:Class rdf:about="http://purl.obolibrary.org/obo/CHEBI_27732"
  >
      <rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/
        obo/CHEBI_26385"/>
      <rdfs:subClassOf rdf:resource="http://purl.obolibrary.org/
        obo/CHEBI_27134"/>
      <rdfs:subClassOf>
         <owl:Restriction>
            <owl:onProperty rdf:resource="http://purl.obolibrary.
              org/obo/RO_0000087"/>
            <owl:someValuesFrom rdf:resource="http://purl.
              obolibrary.org/obo/CHEBI_25435"/>
         </owl:Restriction>
      </rdfs:subClassOf>
      ...
      <rdfs:subClassOf>
         <owl:Restriction>
            <owl:onProperty rdf:resource="http://purl.obolibrary.
              org/obo/RO_0000087"/>
            <owl:someValuesFrom rdf:resource="http://purl.
              obolibrary.org/obo/CHEBI_85234"/>
```

```
    </owl:Restriction>
</rdfs:subClassOf>
<obo:IAO_0000115 rdf:datatype="http://www.w3.org/2001/
  XMLSchema#string">A trimethylxanthine in which the three
  methyl groups are located at positions 1, 3, and 7. A
  purine alkaloid that occurs naturally in tea and coffee.<
  /obo:IAO_0000115>
<oboInOwl:hasAlternativeId rdf:datatype="http://www.w3.org
  /2001/XMLSchema#string">CHEBI:22982</
  oboInOwl:hasAlternativeId>
<oboInOwl:hasAlternativeId rdf:datatype="http://www.w3.org
  /2001/XMLSchema#string">CHEBI:3295</
  oboInOwl:hasAlternativeId>
<oboInOwl:hasAlternativeId rdf:datatype="http://www.w3.org
  /2001/XMLSchema#string">CHEBI:41472</
  oboInOwl:hasAlternativeId>
<oboInOwl:hasOBONamespace rdf:datatype="http://www.w3.org
  /2001/XMLSchema#string">chebi_ontology</
  oboInOwl:hasOBONamespace>
<oboInOwl:id rdf:datatype="http://www.w3.org/2001/XMLSchema
  #string">CHEBI:27732</oboInOwl:id>
```

```
        <oboInOwl:inSubset rdf:resource="http://purl.obolibrary.org
          /obo/chebi#3_STAR"/>
        <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#
          string">caffeine</rdfs:label>
</owl:Class>
```

Class: caffeine

Term IRI: http://purl.obolibrary.org/obo/CHEBI_27732

Definition: A trimethylxanthine in which the three methyl groups are located at positions 1, 3, and 7. A purine alkaloid that occurs naturally in tea and coffee.

Annotations

- **database_cross_reference:** PMID:15257305; PMID:10822912; PMID:18421070; PMID:16528931; PMID:22770225; PMID:12943586; PMID:17957400; PMID:8679661; PMID:12397877; KNApSAcK:C00001492; PMID:14521986; PMID:11815511; PMID:11431501; PMID:20164568; Beilstein:17705; PMID:11209966; PMID:9132918; PMID:11410911; PMID:16709440; PMID:11014293; PMID:18625110; Gmelin:103040; MetaCyc:1-3-7-TRIMETHYLXANTHINE; PMID:19879252; KEGG:C07481; PMID:12457274; PMID:10803761; PMID:19088793; HMDB:HMDB0001847; PMID:7689104; PMID:14607010; KEGG:D00528; PMID:16143823; PMID:11949272; DrugBank:DB00201; PMID:15280431; PMID:10884512; PMID:17387608; PMID:16856769; PMID:19084078; PMID:16644114; PMID:10924888; PMID:10796597; PMID:11022879; LINCS:LSM-2026; PMID:10510174; PMID:16805851; PMID:8347173; PDBeChem:CFF; PMID:7441110; PMID:16391865; PMID:9218278; PMID:15840517; PMID:9067318; PMID:18258404; Drug_Central:463; PMID:19418355; PMID:17508167; PMID:17724925; PMID:12574990; PMID:10983026; PMID:15718055; Reaxys:17705; PMID:19007524; Wikipedia:Caffeine; PMID:9063686; PMID:18647558; PMID:18068204; CAS:58-08-2; PMID:17132260; PMID:20470411; PMID:8332255; PMID:11312039; PMID:15681408; PMID:17932622; PMID:19047957; PMID:12915014
- **has_alternative_id:** CHEBI:22982; CHEBI:41472; CHEBI:3295
- **has_exact_synonym:** CAFFEINE; Caffeine; 1,3,7-trimethyl-3,7-dihydro-1H-purine-2,6-dione; caffeine
- **has_obo_namespace:** chebi_ontology
- **has_related_synonym:** Thein; guaranine; cafeine; theine; 1-methyltheobromine; 1,3,7-trimethyl-2,6-dioxopurine; 3,7-Dihydro-1,3,7-trimethyl-1H-purin-2,6-dion; 1,3,7-trimethylxanthine; anhydrous caffeine; 1,3,7-Trimethylxanthine; 7-methyltheophylline; Coffein; cafeina; 1,3,7-trimethylpurine-2,6-dione; mateina; methyltheobromine; Koffein; teina
- **http://purl.obolibrary.org/obo/chebi/charge:** 0
- **http://purl.obolibrary.org/obo/chebi/formula:** C8H10N4O2
- **http://purl.obolibrary.org/obo/chebi/inchi:** InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3
- **http://purl.obolibrary.org/obo/chebi/inchikey:** RYYVLZVUVIJVGH-UHFFFAOYSA-N
- **http://purl.obolibrary.org/obo/chebi/mass:** 194.19076
- **http://purl.obolibrary.org/obo/chebi/monoisotopicmass:** 194.080
- **http://purl.obolibrary.org/obo/chebi/smiles:** Cn1cnc2n(C)c(=O)n(C)c(=O)c12
- **http://www.geneontology.org/formats/oboInOwl#id:** CHEBI:27732
- **in_subset:** http://purl.obolibrary.org/obo/chebi#3_STAR

Class Hierarchy

Thing
+ chemical entity
+ molecular entity
+ main group molecular entity
+ p-block molecular entity
+ carbon group molecular entity
+ organic molecular entity
+ organic molecule
+ organic cyclic compound
+ organic heterocyclic compound
+ organic heteropolycyclic compound
+ organic heterobicyclic compound
+ imidazopyrimidine
+ purines
+ purine alkaloid
+ methylxanthine
+ trimethylxanthine
- 8-(3-chlorostyryl)caffeine
- caffeine

Class description of *caffeine* in ChEBI

(Source: `http://www.ontobee.org/`)

caffeine specialization of
    26385 *purine alkaloid* and 27134 *trimethylxanthine*

Have additional subclass relationships
    not subsumption (*is-a*).

Related Classes

Superclasses & Asserted Axioms

- muscle tissue disease
- autosomal dominant disease
- has material basis in some autosomal dominant inheritance

Related classes of *malignant hyperthermia* in the Human Disease Ontology

(Source: `http://www.ontobee.org/`)

Superclasses & Asserted Axioms

- has role some human blood serum metabolite
- has role some mouse metabolite
- has role some plant metabolite
- has role some fungal metabolite
- has role some environmental contaminant
- has role some adjuvant
- has role some food additive
- has role some ryanodine receptor agonist
- has role some adenosine receptor antagonist
- has role some ryanodine receptor modulator
- has role some EC 3.1.4.* (phosphoric diester hydrolase) inhibitor
- has role some EC 2.7.11.1 (non-specific serine/threonine protein kinase) inhibitor
- has role some adenosine A2A receptor antagonist
- has role some central nervous system stimulant
- has role some psychotropic drug
- has role some diuretic
- has role some xenobiotic
- has role some mutagen
- purine alkaloid
- trimethylxanthine

Related classes of *caffeine* in ChEBI

(Source: `http://www.ontobee.org/`)

```
<rdfs:subClassOf>
 <owl:Restriction>
  <owl:onProperty rdf:resource="http://purl.obolibrary.org/obo/
    RO_0000087"/>
  <owl:someValuesFrom rdf:resource="http://purl.obolibrary.org/obo
    /CHEBI_25435"/>
 </owl:Restriction>
</rdfs:subClassOf>
```

relationship between *caffeine* and
the entry CHEBI:25435 (*mutagen*)
defined by RO:0000087 (*has role*)
of the *Relations Ontology*.

Means *caffeine has role mutagen*

Search *has role* in OWL:

```
$ xmllint --xpath "//*[local-name()='ObjectProperty'][@*[local-name
    ()='about']='http://purl.obolibrary.org/obo/RO_0000087']"
    chebi_lite.owl
```

Finds `ObjectProperty`

selects the ones with `about` attribute

with the relation URI as value.

Neither transitive or cyclic:

```
<owl:ObjectProperty rdf:about="http://purl.obolibrary.org/obo/
  RO_0000087">
      ...
      <oboInOwl:id rdf:datatype="http://www.w3.org/2001/XMLSchema
        #string">has_role</oboInOwl:id>
      <oboInOwl:is_cyclic rdf:datatype="http://www.w3.org/2001/
        XMLSchema#boolean">false</oboInOwl:is_cyclic>
      <oboInOwl:is_transitive rdf:datatype="http://www.w3.org
        /2001/XMLSchema#boolean">false</oboInOwl:is_transitive>
      ...
      <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#
        string">has role</rdfs:label>
  </owl:ObjectProperty>
```

ObjectProperty: has role

Term IRI: http://purl.obolibrary.org/obo/RO_0000087

Annotations

- **database_cross_reference:** RO:0000087
- **has_obo_namespace:** chebi_ontology
- **http://www.geneontology.org/formats/oboInOwl#id:** has_role
- **http://www.geneontology.org/formats/oboInOwl#is_cyclic:** false
- **http://www.geneontology.org/formats/oboInOwl#is_transitive:** false
- **shorthand:** has_role

Description of *has role* property

(Source: `http://www.ontobee.org/`)

URIs and Labels

Standardize the process

- scripts convert label into URI
- and vice-versa

Internal ontology processing using URIs

- then convert to labels

URI of a label

Get URI of *malignant hyperthermia*:

```
$ xmllint --xpath "//*[local-name()='label' and text()='malignant
    hyperthermia']/../@*[local-name()='about']" doid.owl

  rdf:about="http://purl.obolibrary.org/obo/DOID_8545"
```

`@*[local-name()='about']`

extracts the URI specified

as an attribute of that class.

Only the value, add `string`:

```
$ xmllint --xpath "string(//*[local-name()='label' and text()='
    malignant hyperthermia']/../@*[local-name()='about'])" doid.owl

  http://purl.obolibrary.org/obo/DOID_8545
```

`string` returns only one attribute value

- even if many are matched
- assuming *malignant hyperthermia* is unambiguous

Alternative:

- add `tr '"''n'` to separate URI
- add `grep 'http'` to keep lines with URI

Get URI of *caffeine*:

```
$ xmllint --xpath "string(//*[local-name()='label' and text()='
    caffeine']/../@*[local-name()='about'])" chebi_lite.owl | tr
    '"' '\n' | grep 'http'
```

Script *geturi.sh*:

```
1 OWLFILE=$1
2 xargs -I {} xmllint --xpath "//*[local-name()='label' and text()
    ='{}']/../@*[local-name()='about']" $OWLFILE | \
3 tr '"' '\n' | grep 'http'
```

Multiple labels as standard input

- OWL file to find URIs as argument
- `xargs` process each line of standard input

Execute:

```
$ chmod u+x geturi.sh
$ echo 'malignant hyperthermia' | ./geturi.sh doid.owl
$ echo 'caffeine' | ./geturi.sh chebi_lite.owl

  http://purl.obolibrary.org/obo/DOID_8545
  http://purl.obolibrary.org/obo/CHEBI_27732
```

Execute using multiple labels:

```
$ echo -e 'malignant hyperthermia\nmuscle tissue disease' | ./geturi.sh doid.owl
$ echo -e 'caffeine\npurine alkaloid\ntrimethylxanthine' | ./geturi.sh chebi_lite.owl

http://purl.obolibrary.org/obo/DOID_8545
http://purl.obolibrary.org/obo/DOID_66

http://purl.obolibrary.org/obo/CHEBI_27732
http://purl.obolibrary.org/obo/CHEBI_26385
http://purl.obolibrary.org/obo/CHEBI_27134
```

Label of a URI

Get label disease 8545:

```
$ xmllint --xpath "//*[local-name()='Class'][@*[local-name()='about
  ']='http://purl.obolibrary.org/obo/DOID_8545']/*[local-name()='
  label']/text()" doid.owl

  malignant hyperthermia
```

`@*[local-name()='label']`

selects element describes label

Problem if multiple matches

`text()` all labels in same line

alternative add `tr` and `grep`

Get label of compound 27732:

```
$ xmllint --xpath "//*[local-name()='Class'][@*[local-name()='about
  ']='http://purl.obolibrary.org/obo/CHEBI_27732']/*[local-name()
  ='label']/text()" chebi_lite.owl

  caffeine
```

Script *getlabels.sh*:

```
1 OWLFILE=$1
2 xargs -I {} xmllint --xpath "//*[local-name()='Class'][@*[local-
    name()='about']='{}']/*[local-name()='label']" $OWLFILE | \
3 tr '<>' '\n' | \
4 grep -v -e ':label' -e '^$'
```

Multiple URIs as standard input

- OWL file to find labels as argument
- `xargs` process each line of standard input
- `text` not adds newline after each match
- split in multiple lines using `tr`
- filtering `:label` keyword or are empty `^$`

Execute:

```
$ chmod u+x getlabels.sh
$ echo 'http://purl.obolibrary.org/obo/DOID_8545' | ./getlabels.sh
    doid.owl
$ echo 'http://purl.obolibrary.org/obo/CHEBI_27732' | ./getlabels.
    sh chebi_lite.owl

  malignant hyperthermia
  caffeine
```

Execute with multiple URIs:

```
$ echo -e 'http://purl.obolibrary.org/obo/DOID_8545\nhttp://purl.
    obolibrary.org/obo/DOID_66' | ./getlabels.sh doid.owl
$ echo -e 'http://purl.obolibrary.org/obo/CHEBI_27732\nhttp://purl.
    obolibrary.org/obo/CHEBI_26385\nhttp://purl.obolibrary.org/obo/
    CHEBI_27134' | ./getlabels.sh chebi_lite.owl

malignant hyperthermia
muscle tissue disease

caffeine
purine alkaloid
trimethylxanthine
```

Test both scripts:

```
$ echo -e 'malignant hyperthermia\nmuscle tissue disease' | ./geturi.sh doid.owl | ./getlabels.sh doid.owl
$ echo -e 'caffeine\npurine alkaloid\ntrimethylxanthine' | ./geturi.sh chebi_lite.owl | ./getlabels.sh chebi_lite.owl

malignant hyperthermia
muscle tissue disease

caffeine
purine alkaloid
trimethylxanthine
```

URIs as input:

```
$ echo -e 'http://purl.obolibrary.org/obo/DOID_8545\nhttp://purl.
    obolibrary.org/obo/DOID_66' | ./getlabels.sh doid.owl | ./
    geturi.sh doid.owl
$ echo -e 'http://purl.obolibrary.org/obo/CHEBI_27732\nhttp://purl.
    obolibrary.org/obo/CHEBI_26385\nhttp://purl.obolibrary.org/obo/
    CHEBI_27134' | ./getlabels.sh chebi_lite.owl | ./geturi.sh
    chebi_lite.owl

  http://purl.obolibrary.org/obo/DOID_8545
  http://purl.obolibrary.org/obo/DOID_66

  http://purl.obolibrary.org/obo/CHEBI_27732
  http://purl.obolibrary.org/obo/CHEBI_26385
  http://purl.obolibrary.org/obo/CHEBI_27134
```

Synonyms

Not always mentioned using official label

- text alternative labels
- represented by `hasExactSynonym`

Synonyms of a disease:

```
$ xmllint --xpath "//*[local-name()='Class'][@*[local-name()='about
  ']='http://purl.obolibrary.org/obo/DOID_8545']/*[local-name()='
  hasExactSynonym']" doid.owl

<oboInOwl:hasExactSynonym rdf:datatype="http://www.w3.org/2001/
  XMLSchema#string">anesthesia related hyperthermia</
  oboInOwl:hasExactSynonym>
<oboInOwl:hasExactSynonym rdf:datatype="http://www.w3.org/2001/
  XMLSchema#string">malignant hyperpyrexia due to anesthesia</
  oboInOwl:hasExactSynonym>
```

Both primary label and synonyms:

```
$ xmllint --xpath "//*[local-name()='Class'][@*[local-name()='about
  ']='http://purl.obolibrary.org/obo/DOID_8545']/*[local-name()='
  hasExactSynonym' or local-name()='label']" doid.owl

<oboInOwl:hasExactSynonym rdf:datatype="http://www.w3.org/2001/
  XMLSchema#string">anesthesia related hyperthermia</
  oboInOwl:hasExactSynonym>
<oboInOwl:hasExactSynonym rdf:datatype="http://www.w3.org/2001/
  XMLSchema#string">malignant hyperpyrexia due to anesthesia</
  oboInOwl:hasExactSynonym>
<rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string
  ">malignant hyperthermia</rdfs:label>
```

Update *getlabels.sh*:

```
1 OWLFILE=$1
2 xargs -I {} xmllint --xpath "//*[local-name()='Class'][@*[local-
    name()='about']='{}']/*[local-name()='hasExactSynonym' or local
    -name()='hasRelatedSynonym' or local-name()='label']" $OWLFILE
    | \
3 tr '<>' '\n' | \
4 grep -v -e ':label' -e ':hasExactSynonym' -e 'hasRelatedSynonym'
    -e '^$'
```

Adding the `hasExactSynonym` keyword and `hasRelatedSynonym`

Execute:

```
$ echo -e 'http://purl.obolibrary.org/obo/DOID_8545' | ./getlabels.sh doid.owl

anesthesia related hyperthermia
malignant hyperpyrexia due to anesthesia
malignant hyperthermia
```

URI of synonyms

Send output to *geturi.sh*:

```
$ echo 'http://purl.obolibrary.org/obo/DOID_8545' | ./getlabels.sh ↩
    doid.owl | ./geturi.sh doid.owl
```

XPath warnings for the two synonyms:

```
XPath set is empty
XPath set is empty
http://purl.obolibrary.org/obo/DOID_8545
```

Ignore these mismatches:

```
$ echo 'http://purl.obolibrary.org/obo/DOID_8545' | ./getlabels.sh
    doid.owl | ./geturi.sh doid.owl 2>/dev/null
```

Or update *geturi.sh* to include synonyms:

```
1 OWLFILE=$1
2 xargs -I {} xmllint --xpath "//*[(local-name()='hasExactSynonym'
    or local-name()='hasRelatedSynonym' or local-name()='label')
    and text()='{}']/../@*[local-name()='about']" $OWLFILE | \
3 tr '"' '\n' | grep 'http'
```

Execute:

```
$ echo 'http://purl.obolibrary.org/obo/DOID_8545' | ./getlabels.sh
    doid.owl | ./geturi.sh doid.owl

  http://purl.obolibrary.org/obo/DOID_8545
  http://purl.obolibrary.org/obo/DOID_8545
  http://purl.obolibrary.org/obo/DOID_8545
```

Avoid repetitions:

```
$ echo 'http://purl.obolibrary.org/obo/DOID_8545' | ./getlabels.sh ↩
    doid.owl | ./geturi.sh doid.owl | sort -u

http://purl.obolibrary.org/obo/DOID_8545
```

Parent Classes

Parent classes of *malignant hyperthermia*:

```
$ xmllint --xpath "//*[local-name()='Class'][@*[local-name()='about
    ']='http://purl.obolibrary.org/obo/DOID_8545']/*[local-name()='
    subClassOf']/@*[local-name()='resource']" doid.owl
```

`[local-name()='subClassOf']` gets subclass

`@*[local-name()='resource']` gets attribute with URI

Output URIs parents of 8545:

```
rdf:resource="http://purl.obolibrary.org/obo/DOID_0050736"
rdf:resource="http://purl.obolibrary.org/obo/DOID_66"
```

Execute for *caffeine*:

```
$ xmllint --xpath "//*[local-name()='Class'][@*[local-name()='about
  ']='http://purl.obolibrary.org/obo/CHEBI_27732']/*[local-name()
  ='subClassOf']/@*[local-name()='resource']" chebi_lite.owl

rdf:resource="http://purl.obolibrary.org/obo/CHEBI_26385"
rdf:resource="http://purl.obolibrary.org/obo/CHEBI_27134"
```

No longer can use `string`

multiple parents

and `string` only returns first match

Get only URIs:

```
$ xmllint --xpath "//*[local-name()='Class'][@*[local-name()='about
  ']='http://purl.obolibrary.org/obo/CHEBI_27732']/*[local-name()
  ='subClassOf']/@*[local-name()='resource']" chebi_lite.owl | tr
   '"' '\n' | grep 'http'

  http://purl.obolibrary.org/obo/CHEBI_26385
  http://purl.obolibrary.org/obo/CHEBI_27134
```

Script *getparents.sh*:

```
1 OWLFILE=$1
2 xargs -I {} xmllint --xpath "//*[local-name()='Class'][@*[local-
    name()='about']='{}']/*[local-name()='subClassOf']/@*[local-
    name()='resource']" $OWLFILE | \
3 tr '"' '\n' | grep 'http'
```

Multiple URIs given as standard input

OWL file to find parents as argument

Parents of *malignant hyperthermia*:

```
$ chmod u+x getparents.sh
$ echo 'http://purl.obolibrary.org/obo/DOID_8545' | ./getparents.sh
    doid.owl

http://purl.obolibrary.org/obo/DOID_0050736
http://purl.obolibrary.org/obo/DOID_66
```

Labels of parents

Redirect the output:

```
$ echo 'http://purl.obolibrary.org/obo/DOID_8545' | ./getparents.sh
    doid.owl | ./getlabels.sh doid.owl

autosomal dominant disease
muscle tissue disease
```

Same with *caffeine*:

```
$ echo 'http://purl.obolibrary.org/obo/CHEBI_27732' | ./getparents.sh chebi_lite.owl | ./getlabels.sh chebi_lite.owl

purine alkaloid
trimethylxanthine
```

Related classes

All related classes

besides *subClassOf*:

```
$ xmllint --xpath "//*[local-name()='Class'][@*[local-name()='about
  ']='http://purl.obolibrary.org/obo/CHEBI_27732']/*[local-name()
  ='subClassOf']//*[local-name()='someValuesFrom']/@*[local-name
  ()='resource']" chebi_lite.owl | tr '"' '\n'| grep 'http'
```

Related classes are

in attribute *resource*

of *someValuesFrom* element

inside *subClassOf* element

Related classes of *caffeine*:

```
http://purl.obolibrary.org/obo/CHEBI_25435
http://purl.obolibrary.org/obo/CHEBI_35337
http://purl.obolibrary.org/obo/CHEBI_35471
http://purl.obolibrary.org/obo/CHEBI_35498
http://purl.obolibrary.org/obo/CHEBI_35703
...
http://purl.obolibrary.org/obo/CHEBI_75771
http://purl.obolibrary.org/obo/CHEBI_76924
http://purl.obolibrary.org/obo/CHEBI_76946
http://purl.obolibrary.org/obo/CHEBI_78298
http://purl.obolibrary.org/obo/CHEBI_85234
```

Labels of related classes

Add *getlabels.sh*:

```
$ xmllint --xpath "//*[local-name()='Class'][@*[local-name()='about
  ']='http://purl.obolibrary.org/obo/CHEBI_27732']/*[local-name()
  ='subClassOf']//*[local-name()='someValuesFrom']/@*[local-name
  ()='resource']" chebi_lite.owl | tr '"' '\n'| grep 'http' | ./
  getlabels.sh chebi_lite.owl

  mutagen
  central nervous system stimulant
  psychotropic drug
  diuretic
  xenobiotic
  ryanodine receptor modulator
  EC 3.1.4.* (phosphoric diester hydrolase) inhibitor
  EC 2.7.11.1 (non-specific serine/threonine protein kinase)
    inhibitor
  adenosine A2A receptor antagonist
  adjuvant
```

```
food additive
ryanodine receptor agonist
adenosine receptor antagonist
mouse metabolite
plant metabolite
fungal metabolite
environmental contaminant
human blood serum metabolite
```

Ancestors

Chain invocations of *getparents.sh*

- until no matches (root)
- avoid cyclic relations (infinite loop)
- consider only parent relations

Grandparents

Parents of parents also generalizations

Grandparents of *malignant hyperthermia*:

```
$ echo 'malignant hyperthermia' | ./geturi.sh doid.owl | ./
    getparents.sh doid.owl | ./getparents.sh doid.owl

  http://purl.obolibrary.org/obo/DOID_0050739
  http://purl.obolibrary.org/obo/DOID_0080000
```

Their labels:

```
$ echo 'malignant hyperthermia' | ./geturi.sh doid.owl | ./
    getparents.sh doid.owl | ./getparents.sh doid.owl | ./getlabels
    .sh doid.owl

  autosomal genetic disease
  muscular disease
```

Root class

Not have any parent

disease and *chemical entity*

highly generic terms

Check root class:

```
$ echo 'disease' | ./geturi.sh doid.owl | ./getparents.sh doid.owl
$ echo 'chemical entity' | ./geturi.sh chebi_lite.owl | ./
    getparents.sh chebi_lite.owl
```

Warning confirming root class:

```
XPath set is empty
```

Recursion

Script *getancestors.sh*:

```
1 OWLFILE=$1
2 CLASSES=$(cat -)
3 [[ -z "$CLASSES" ]] && exit
4 PARENTS=$(echo "$CLASSES" | ./getparents.sh $OWLFILE | sort -u)
5 echo "$PARENTS"
6 echo "$PARENTS" | ./getancestors.sh $OWLFILE
```

List of URIs as standard input

- invokes *getparents.sh* recursively
- until reaches root class

Standard input in variable `CLASSES` to use twice:

- check input is empty (line 3)
- get parents classes (line 4).

Input empty then script ends

- base case of the recursion
- otherwise run indefinitely

Output in variable `PARENTS` to use twice

- output these direct parents (line 5)
- get ancestors of parents (line 6)

Invoking *getancestors.sh* inside *getancestors.sh*

- defines recursion step
- at some time reach classes without parents (root classes)
- then script ends

`echo` of variables `CLASSES` and `PARENTS`

- inside commas so newline chars preserved

Test with *malignant hyperthermia*:

```
$ chmod u+x getancestors.sh
$ echo 'http://purl.obolibrary.org/obo/DOID_8545' | ./getancestors.sh doid.owl

http://purl.obolibrary.org/obo/DOID_0050736
http://purl.obolibrary.org/obo/DOID_66
http://purl.obolibrary.org/obo/DOID_0050739
http://purl.obolibrary.org/obo/DOID_0080000
http://purl.obolibrary.org/obo/DOID_0050177
http://purl.obolibrary.org/obo/DOID_17
http://purl.obolibrary.org/obo/DOID_630
http://purl.obolibrary.org/obo/DOID_7
http://purl.obolibrary.org/obo/DOID_4
```

Warning when reaches root class:

```
XPath set is empty
```

Redirect warnings:

```
$ echo 'malignant hyperthermia' | ./geturi.sh doid.owl | ./getancestors.sh doid.owl 2>/dev/null | ./getlabels.sh doid.owl
```

Ancestors of *malignant hyperthermia*:

```
autosomal dominant disease
muscle tissue disease
autosomal genetic disease
muscular disease
monogenic disease
musculoskeletal system disease
genetic disease
disease of anatomical entity
disease
```

First two ancestors direct parents

last one the root class.

prints the parents before invoking itself

Same with *caffeine*:

```
$ echo 'caffeine' | ./geturi.sh chebi_lite.owl | ./getancestors.sh
    chebi_lite.owl | ./getlabels.sh chebi_lite.owl | sort -u
```

Repeated classes

- using different branches
- add `sort -u`

Ancestors of *caffeine*:

```
alkaloid
aromatic compound
bicyclic compound
carbon group molecular entity
chemical entity
cyclic compound
heteroarene
heterobicyclic compound
heterocyclic compound
heteroorganic entity
heteropolycyclic compound
imidazopyrimidine
main group molecular entity
methylxanthine
molecular entity
molecule
nitrogen molecular entity
organic aromatic compound
organic cyclic compound
organic heterobicyclic compound
```

```
organic heterocyclic compound
organic heteropolycyclic compound
organic molecular entity
organic molecule
organonitrogen compound
organonitrogen heterocyclic compound
p-block molecular entity
pnictogen molecular entity
polyatomic entity
polycyclic compound
purine alkaloid
purines
trimethylxanthine
```

My Lexicon

Labels and related classes from ontology

Create *do_8545_lexicon.txt*:

```
$ echo 'malignant hyperthermia' | ./geturi.sh doid.owl | ./
    getlabels.sh doid.owl > do_8545_lexicon.txt
```

Lexicon for *malignant hyperthermia*

with all its labels

Ancestors labels

Add to lexicon:

```
$ echo 'malignant hyperthermia' | ./geturi.sh doid.owl | ./
    getancestors.sh doid.owl | ./getlabels.sh doid.owl >>
    do_8545_lexicon.txt
```

>> and not >

append lines to file

Check contents:

```
$ cat do_8545_lexicon.txt | sort -u

anesthesia related hyperthermia
autosomal dominant disease
autosomal genetic disease
disease
disease of anatomical entity
genetic disease
malignant hyperpyrexia due to anesthesia
malignant hyperthermia
monogenic disease
muscle tissue disease
muscular disease
musculoskeletal system disease
```

Same for *caffeine* in *chebi_27732_lexicon.txt*:

```
$ echo 'caffeine' | ./geturi.sh chebi_lite.owl | ./getlabels.sh ↵
    chebi_lite.owl > chebi_27732_lexicon.txt
$ echo 'caffeine' | ./geturi.sh chebi_lite.owl | ./getancestors.sh ↵
    chebi_lite.owl | ./getlabels.sh chebi_lite.owl >> ↵
    chebi_27732_lexicon.txt
```

Check contents:

```
$ cat chebi_27732_lexicon.txt | sort -u

alkaloid
aromatic compound
bicyclic compound
caffeine
...
```

This lexicon is much larger.

Merging labels

Merging two lexicons in *lexicon.txt*:

```
$ cat do_8545_lexicon.txt chebi_27732_lexicon.txt | sort -u > ↩
    lexicon.txt
```

Corpus

Retrieve text file with sentences
from abstracts related to caffeine:

```
$ curl -O http://labs.rd.ciencias.ulisboa.pt/book/archive20191126.zip
$ unzip archive20191126.zip chebi_27732_sentences.txt
```

Generated in previous chapters of the book

Recognize any mention in *chebi_27732_sentences.txt*:

```
$ grep -w -i -F -f lexicon.txt chebi_27732_sentences.txt
```

`-F` option

our lexicon is list of fixed strings

not includes regular expressions.

Some results not include direct mention
to *caffeine* or *malignant hyperthermia*

Example *molecule* ancestor of *caffeine*:

```
The remainder of the molecule is hydrophilic and presumably
  constitutes the cytoplasmic domain of the protein.
```

Example *disease* ancestor of *malignant hyperthermia*:

```
Our data suggest that divergent activity profiles may cause
  varied disease phenotypes by specific mutations.
```

Ancestors matched

Ancestors being matched:

```
$ grep -o -w -F -f lexicon.txt chebi_27732_sentences.txt | sort -u

caffeine
disease
malignant hyperthermia
molecule
```

Text limited and using official labels

- missing acronyms and simple variations (plural)
- solution use a stemmer
- all ancestors besides subsumption
- add some regular expressions

Generic Lexicon

Recognizing any disease
represented in ontology
in our sentences
related to *caffeine*

Get all labels without restricting to any URI:

```
$ xmllint --xpath "//*[local-name()='Class']/*[local-name()='
    hasExactSynonym' or local-name()='hasRelatedSynonym' or local-
    name()='label']" doid.owl
```

Script *getalllabels.sh*:

```
1 OWLFILE=$1
2 xmllint --xpath "//*[local-name()='Class']/*[local-name()='
    hasExactSynonym' or local-name()='hasRelatedSynonym' or local-
    name()='label']" $OWLFILE | \
3 tr '<>' '\n' | \
4 grep -v -e ':label' -e ':hasExactSynonym' -e 'hasRelatedSynonym'
    -e '^$' | \
5 sort -u
```

Execute:

```
$ chmod u+x getalllabels.sh
$ ./getalllabels.sh doid.owl

11-beta-hydroxysteroid dehydrogenase deficiency type 2
11p partial monosomy syndrome
1,4-phenylenediamine allergic contact dermatitis
...
Zoophilia
Zoophobia
zygomycosis
```

Redirect to *diseases.txt*:

```
$ ./getalllabels.sh doid.owl > diseases.txt
```

Check how many labels:

```
$ wc -l diseases.txt
```

More than 34 thousand labels

Recognize lexicon entries:

```
$ grep -n -w -E -f diseases.txt chebi_27732_sentences.txt

  grep: Unmatched ) or \)
```

Error because lexicon contains special characters
also used by regular expressions (parentheses)

Replace `-E` by `-F`:

```
$ grep -n -o -w -F -f diseases.txt chebi_27732_sentences.txt

1:malignant hyperthermia
2:malignant hyperthermia
9:central core disease
10:disease
10:myopathy
...
1092:malignant hyperthermia
1092:central core disease
1103:malignant hyperthermia
1104:malignant hyperthermia
1106:central core disease
1106:myopathy
```

Problematic entries

Expressions enclosed by parentheses or brackets:

```
Post measles encephalitis (disorder)
Glaucomatous atrophy [cupping] of optic disc
```

Separation characters (commas or colons)

to represent a specialization

```
Tapeworm infection: intestinal taenia solum
Tapeworm infection: pork
Pemphigus, Benign Familial
ATR, nondeletion type
```

Comma also part of term:

```
46,XY DSD due to LHB deficiency
```

& to represent ampersand:

```
Gonococcal synovitis &amp;/or tenosynovitis
```

But alternatives already included:

```
Gonococcal synovitis and tenosynovitis
Gonococcal synovitis or tenosynovitis
```

Not trivial to devise rules

- that fully solve these issues
- will be exceptions to any rule

Special characters frequency

Check the impact:

```
$ grep -c -F '(' diseases.txt
$ grep -c -F ',' diseases.txt
$ grep -c -F '[' diseases.txt
$ grep -c -F ':' diseases.txt
$ grep -c -F '&amp;' diseases.txt
```

Parentheses and commas most frequent

more than one thousand entries

Completeness

Check presence of *ATR*

acronym *alpha thalassemia-X-linked intellectual disability syndrome*

```
$ grep -E '^ATR' diseases.txt

ATR-16 syndrome
ATR, nondeletion type
ATR syndrome, deletion type
ATR syndrome linked to chromosome 16
ATR-X syndrome
```

A single *ATR* mention will not be recognized:

```
$ echo 'The ATR syndrome is an alpha thalassemia that has material
    basis in mutation in the ATRX gene on Xq21' | grep -w 'ATR'
```

Removing special characters

Remove parentheses and brackets:

```
$ tr -d '[](){}' < diseases.txt
```

Miss shorter labels such as *Post measles encephalitis*,
but at least will recognize:

```
$ tr -d '[](){}' < diseases.txt | grep 'Post measles encephalitis
  disorder'
```

Alternative create multiple entries in the lexicon
or transform the labels in regular expressions

Removing extra terms

Remove text after separation char:

```
$ tr -d '[](){}' < diseases.txt | sed -E 's/[,:;] .*$//'
```

Enforces a space after the separation char

avoids: *46,XY DSD due to LHB deficiency*

Recognize both *ATR* and *ATR syndrome*:

```
$ tr -d '[](){}' < diseases.txt | sed -E 's/[,:;] .*$//' | grep -E ↩
    '^ATR'
```

Removing extra spaces

Remove leading or trailing spaces:

```
$ tr -d '[](){}' < diseases.txt | sed -E 's/[,:;] .*$//; s/^ *//; s
    / *$//'
```

More replacement expressions to `sed`
separated by semicolon

Update *getalllabels.sh*:

```
1 OWLFILE=$1
2 xmllint --xpath "//*[local-name()='Class']/*[local-name()='
    hasExactSynonym' or local-name()='hasRelatedSynonym' or local-
    name()='label']" $OWLFILE | \
3 tr '<>' '\n' | \
4 grep -v -e ':label' -e ':hasExactSynonym' -e 'hasRelatedSynonym'
    -e '^$' | \
5 tr -d '[](){}' | \
6 sed -E 's/[,:;] .*$//; s/^ *//; s/ *$//' | sort -u
```

Generate fixed lexicon:

```
$ ./getalllabels.sh doid.owl > diseases.txt
```

Check number of entries:

```
$ wc -l diseases.txt
```

About 33 thousand labels

less because fixes made duplicate entries

Disease recognition

Recognize entries:

```
$ grep -n -o -w -F -f diseases.txt chebi_27732_sentences.txt
```

Labels recognized:

```
$ grep -o -w -F -f diseases.txt chebi_27732_sentences.txt | sort -u
```

47 diseases related *caffeine*:

```
Andersen-Tawil syndrome
arrhythmogenic right ventricular cardiomyopathy
...
scoliosis
syndrome
T cell
```

Check how many labels recognized:

```
$ grep -o -w -F -i -f diseases.txt chebi_27732_sentences.txt | sort
    -u | wc -l
```

66 labels being recognized.

Check new labels recognized:

```
$ grep -o -w -F -i -f diseases.txt chebi_27732_sentences.txt | sort
    -u > diseases_recognized_ignorecase.txt
$ grep -o -w -F -f diseases.txt chebi_27732_sentences.txt | sort -u
    > diseases_recognized.txt
$ grep -v -F -f diseases_recognized.txt
  diseases_recognized_ignorecase.txt

all
All
Arrhythmogenic right ventricular dysplasia
can
Catecholaminergic polymorphic ventricular tachycardia
Central Core Disease
```

```
defect
Disease
dyskinesia
face
fever
hypotonia
Malignant hyperthermia
Malignant Hyperthermia
March
ORF
total
```

Correct matches

Some only recognized by case insensitive match

```
$ grep -i '^dyskinesia$' diseases.txt
```

Lexicon only name with first character in uppercase:

```
Dyskinesia
```

Check in text:

```
$ grep -w -i 'dyskinesia' chebi_27732_sentences.txt
```

Only lowercase:

```
... non-kinesigenic dyskinesia ...
```

Incorrect matches

Case insensitive match create other problems

CAN for *Crouzon syndrome-acanthosis nigricans syndrome*:

```
$ grep -i '^CAN$' diseases.txt
```

Check how many times *CAN* is recognized:

```
$ grep -n -o -w -i -F -f diseases.txt chebi_27732_sentences.txt |
    grep -i ':CAN' | wc -l
```

18 times

Which type of matches:

```
$ grep -o -w -i -F -f diseases.txt chebi_27732_sentences.txt | grep
    -i -E '^CAN$' | sort -u
```

Incorrect mentions:

```
can
```

18 mismatches by case insensitive match

Entity Linking

What recognized labels represent

Find what *AD2* represents:

```
$ echo "AD2" | ./geturi.sh doid.owl | ./getlabels.sh doid.owl
```

Clearly *Alzheimer disease*:

```
AD2
Alzheimer disease 2, late onset
Alzheimer disease associated with APOE4
Alzheimer disease-2
Alzheimer's disease 2
```

Modified labels

Labels modified by previous fixes:

```
$ echo "ATR" | ./geturi.sh doid.owl

  XPath set is empty
```

Solution keep track of the original label

Ambiguity

Classes acronym *PDR* may represent:

```
$ echo "KOS" | ./geturi.sh doid.owl

  http://purl.obolibrary.org/obo/DOID_0111456
  http://purl.obolibrary.org/obo/DOID_0111712
```

Two distinct diseases:

- *Kaufman oculocerebrofacial syndrome* (DOID:0111456)
- *Kagami-Ogata syndrome* (DOID:0111712)

Alternative labels:

```
$ echo "http://purl.obolibrary.org/obo/DOID_0111456" | ./getlabels.sh doid.owl
$ echo "http://purl.obolibrary.org/obo/DOID_0111712" | ./getlabels.sh doid.owl
```

Both containing *PDR* as expected:

```
KOS
blepharophimosis ptosis intellectual disability syndrome
oculocerebrofacial syndrome, Kaufman type
Kaufman oculocerebrofacial syndrome

KOS
Kagami-Ogata syndrome
```

Surrounding entities

Select class closer in meaning
    to other classes in surrounding text

Assuming entities in same text
    semantically related

Example:

```
KOS is a syndromic intellectual disability
```

Identify the diseases:

```
$ echo 'KOS is a syndromic intellectual disability' | grep -o -w -F
    -f diseases.txt

KOS
syndromic intellectual disability
```

Find URIs:

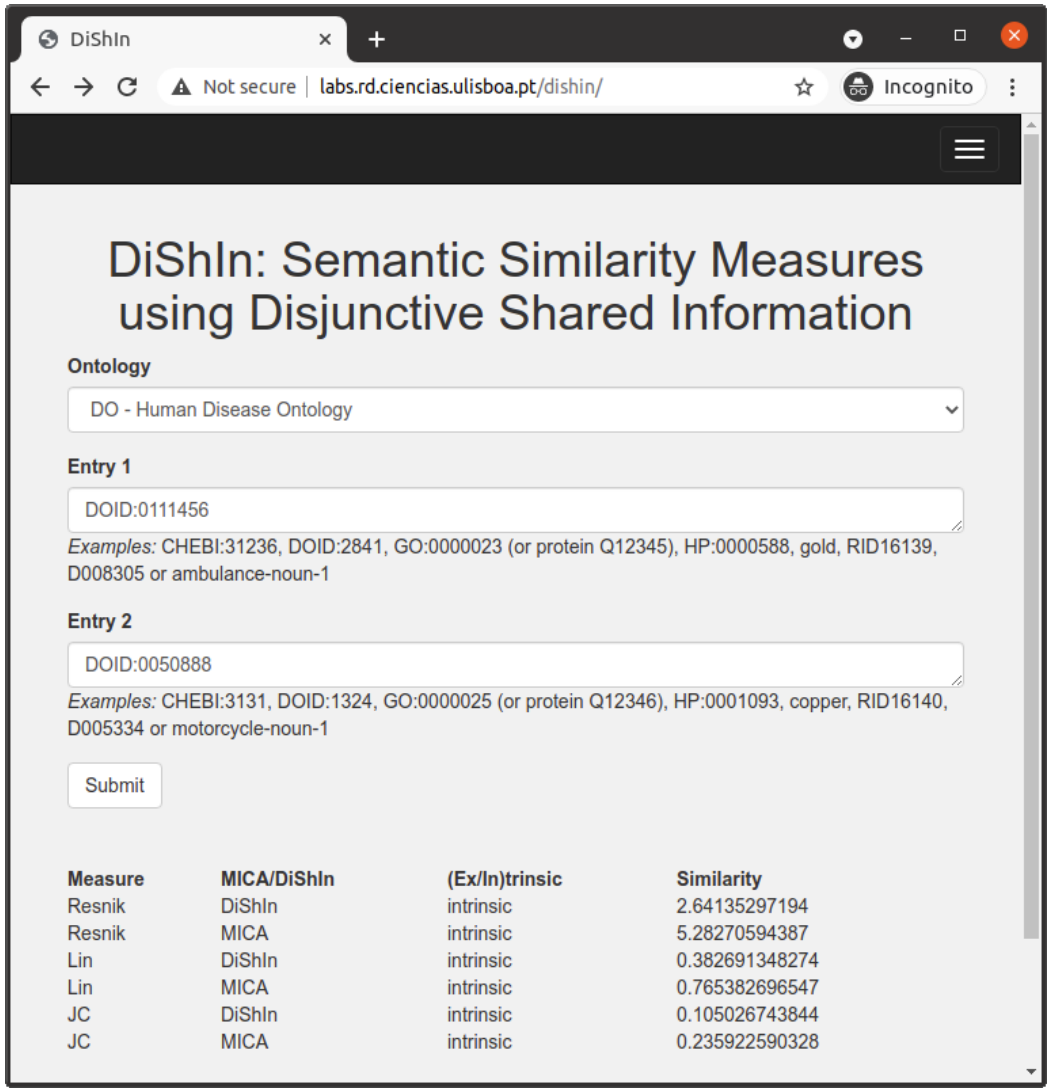
```
$ echo 'KOS is a syndromic intellectual disability' | grep -o -w -F
    -f diseases.txt | ./geturi.sh doid.owl

http://purl.obolibrary.org/obo/DOID_0111456
http://purl.obolibrary.org/obo/DOID_0111712
http://purl.obolibrary.org/obo/DOID_0050888
```

Syndromic intellectual disability (DOID:0050888)

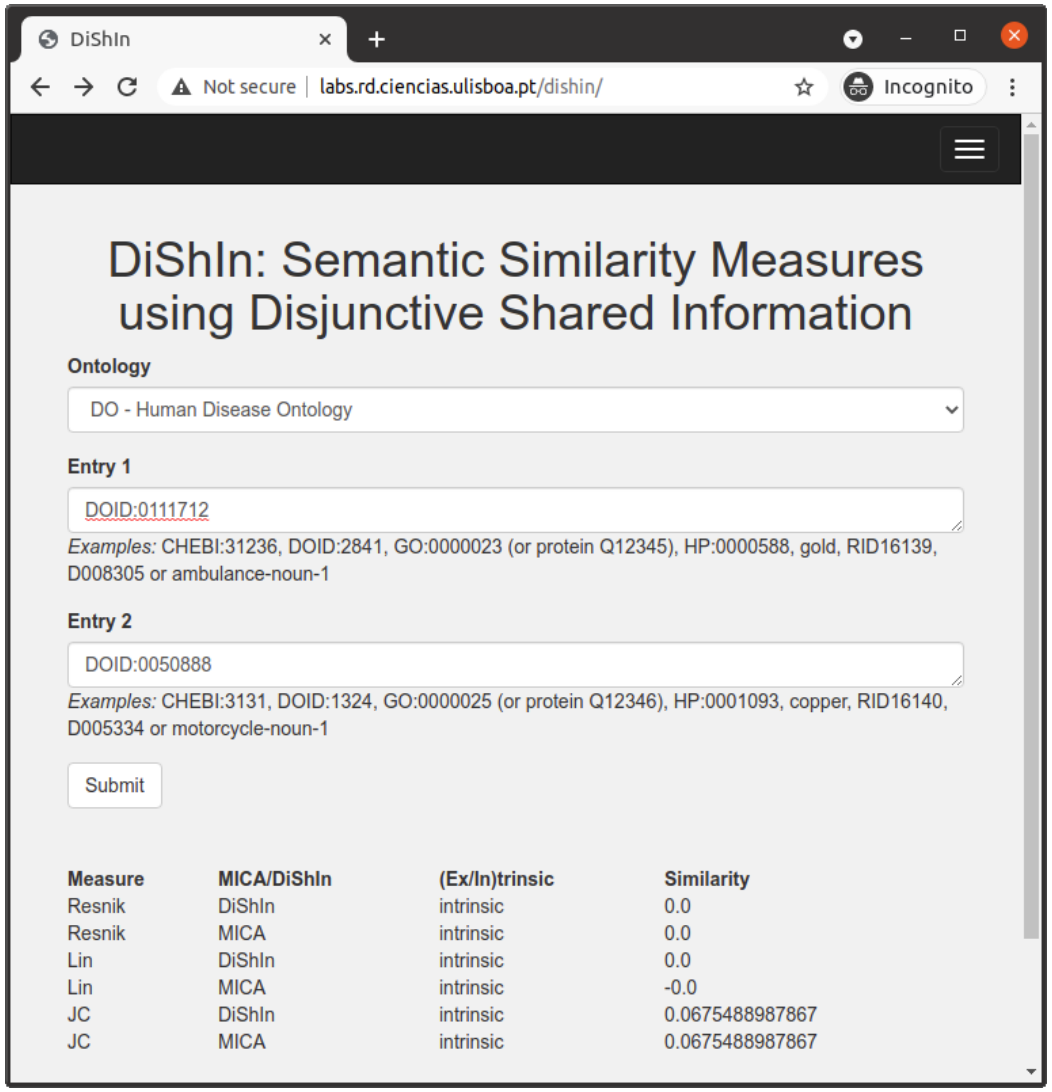
Semantic similarity

Solve ambiguity problems
- quantify how close two classes are
- in terms of semantics
- encoded in a given ontology

Use `http://labs.rd.ciencias.ulisboa.pt/dishin/`
- to calculate semantic similarity between:
- *Kaufman oculocerebrofacial syndrome* (DOID:0111456)
- *Syndromic intellectual disability* (DOID:0050888)
- and
- *Kagami-Ogata syndrome* (DOID:0111712)
- *Syndromic intellectual disability* (DOID:0050888)



Semantic similarity between *Kaufman oculocerebrofacial syndrome* (DOID:0111456) and *Syndromic intellectual disability* (DOID:0050888)



Semantic similarity between *Kagami-Ogata syndrome* (DOID:0111712) and *Syndromic intellectual disability* (DOID:0050888)

Measures

DiShIn provides three measures

- Resnik, Lin and Jiang-Conrath
- last two values between 0 and 1,
- Jiang-Conrath distance converted similarity

Syndromic intellectual disability more similar to
Kaufman oculocerebrofacial syndrome
than to *Kagami-Ogata syndrome*

Semantic similarity can identify
Kaufman oculocerebrofacial syndrome correct linked entity
for *KOS* in this text

DiShIn installation

Execute DiShIn as a command line

need to install python (or python3)

and SQLite

Download DiShIn and latest database version:

```
$ curl -O http://labs.rd.ciencias.ulisboa.pt/dishin/dishin.py
$ curl -O http://labs.rd.ciencias.ulisboa.pt/dishin/ssm.py
$ curl -O http://labs.rd.ciencias.ulisboa.pt/dishin/doid202005.db.gz
$ gunzip -N doid202005.db.gz
```

DiShIn execution

Semantic similarity between:

Kaufman oculocerebrofacial syndrome (DOID:0111456)

Syndromic intellectual disability (DOID:0050888)

and

Kagami-Ogata syndrome (DOID:0111712)

Syndromic intellectual disability (DOID:0050888)

Execute:

```
$ python dishin.py doid.db DOID_0111456 DOID_0050888
$ python dishin.py doid.db DOID_0111712 DOID_0050888
```

Semantic similarity between *Kaufman oculocerebrofacial syndrome* (DOID:0111456) and *Syndromic intellectual disability* (DOID:0050888)

```
Resnik   DiShIn  intrinsic      2.64135297194
Resnik   MICA    intrinsic      5.28270594387
Lin      DiShIn  intrinsic      0.382691348274
Lin      MICA    intrinsic      0.765382696547
JC       DiShIn  intrinsic      0.105026743844
JC       MICA    intrinsic      0.235922590328
```

Semantic similarity between *Kagami-Ogata syndrome* (DOID:0111712) and *Syndromic intellectual disability* (DOID:0050888

```
Resnik   DiShIn  intrinsic      0.0
Resnik   MICA    intrinsic      0.0
Lin      DiShIn  intrinsic      0.0
Lin      MICA    intrinsic      -0.0
JC       DiShIn  intrinsic      0.0675488987867
JC       MICA    intrinsic      0.0675488987867
```

Large lexicons

Online tool MER

- a shell script
- easily executed as a command line
- efficiently recognize and link entities
- using large lexicons

MER installation

Install it locally:

```
$ git clone git://github.com/lasigeBioTM/MER
```

Copy Human Disease Ontology:

```
$ cp doid.owl MER/data/
$ cd MER
```

Lexicon files

Create lexicon:

```
$ (cd data; ../produce_data_files.sh doid.owl)
$ rm data/doid.owl
```

Check the contents:

```
$ tail data/doid*

==> data/doid_links.tsv <==
ziziphus mauritiana fruit allergy http://purl.obolibrary.org/obo/
  DOID_0060507
zlotogora-ogur syndrome http://purl.obolibrary.org/obo/
  DOID_0080400
zlotogora-zilberman-tenenbaum syndrome http://purl.obolibrary.org
  /obo/DOID_0060773
zollinger-ellison syndrome http://purl.obolibrary.org/obo/
  DOID_0050782
zoophilia http://purl.obolibrary.org/obo/DOID_9336
```

```
zoophobia http://purl.obolibrary.org/obo/DOID_600
zunich-kaye syndrome http://purl.obolibrary.org/obo/DOID_0112152
zunich neuroectodermal syndrome http://purl.obolibrary.org/obo/
  DOID_0112152
zygodactyly 1 http://purl.obolibrary.org/obo/DOID_0111820
zygomycosis http://purl.obolibrary.org/obo/DOID_8485

==> data/doid.txt <==
ziziphus mauritiana fruit allergy
zlotogora-ogur syndrome
zlotogora-zilberman-tenenbaum syndrome
zollinger-ellison syndrome
zoophilia
zoophobia
zunich-kaye syndrome
zunich neuroectodermal syndrome
zygodactyly 1
zygomycosis

==> data/doid_word1.txt <==
xpid
```

```
xpv
xrn
xscid
yaba
yaws
zaspopathy
zoophilia
zoophobia
zygomycosis

==> data/doid_word2.txt <==
zellweger syndrome
zemuron allergy
zika fever
zinacef allergy
zinsser.cole.engman syndrome
zlotogora.ogur syndrome
zlotogora.zilberman.tenenbaum syndrome
zollinger.ellison syndrome
zunich.kaye syndrome
zygodactyly 1
```

```
==> data/doid_words2.txt <==
y.linked monogenic
y.linked sertoli
y.linked spermatogenic
yolk sac
young adult.onset
zeta.associated.protein 70
zika virus
zikv congenital
ziziphus mauritiana
zunich neuroectodermal

==> data/doid_words.txt <==
yolk sac tumour
yolk sac tumour of the cns
young adult.onset dhmn
young adult.onset distal hereditary motor neuropathy
zeta.associated.protein 70 deficiency
zika virus congenital syndrome
zika virus disease
```

```
zikv congenital infection
ziziphus mauritiana fruit allergy
zunich neuroectodermal syndrome
```

MER execution

Execute MER:

```
$ cat ../chebi_27732_sentences.txt | tr -d "'" | xargs -I {} ./
    get_entities.sh '{}' doid
```

Removed single quotes

special characters to `xargs`.

get_entities.sh script inside MER folder

not the one created before

Large number of matches:

```
89 111 malignant hyperthermia http://purl.obolibrary.org/obo/
  DOID_8545
144 164 central core disease http://purl.obolibrary.org/obo/
  DOID_3529
13 20 disease http://purl.obolibrary.org/obo/DOID_4
47 55 myopathy http://purl.obolibrary.org/obo/DOID_423
0 20 Central core disease http://purl.obolibrary.org/obo/
  DOID_3529
267 274 disease http://purl.obolibrary.org/obo/DOID_4
254 274 central core disease http://purl.obolibrary.org/obo/
  DOID_3529
48 70 malignant hyperthermia http://purl.obolibrary.org/obo/
  DOID_8545
...
```

First two numbers represent

- the start and end position of match
- followed by label and its URI

Create *diseases_recognized.tsv*:

```
$ cat ../chebi_27732_sentences.txt | tr -d "'" | xargs -I {} ./
    get_entities.sh '{}' doid > ../diseases_recognized.tsv
```

	A	B	C	D
1	89	111	malignant hyperthermia	http://purl.obolibrary.org/obo/DOID_8545
2	144	164	central core disease	http://purl.obolibrary.org/obo/DOID_3529
3	13	20	disease	http://purl.obolibrary.org/obo/DOID_4
4	47	55	myopathy	http://purl.obolibrary.org/obo/DOID_423
5	0	20	Central core disease	http://purl.obolibrary.org/obo/DOID_3529
6	267	274	disease	http://purl.obolibrary.org/obo/DOID_4
7	254	274	central core disease	http://purl.obolibrary.org/obo/DOID_3529
8	48	70	malignant hyperthermia	http://purl.obolibrary.org/obo/DOID_8545

The *diseases_recognized.tsv* file opened in a spreadsheet application