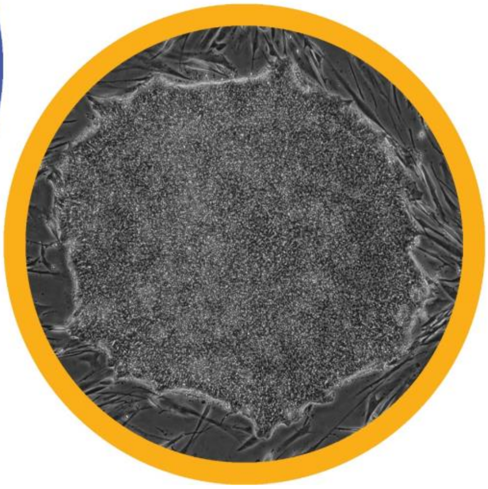
Human Pluripotent Stem Cell Registry

Semantic Queries in hPSCreg®

1.3

25.05.2023



The Human Pluripotent Stem Cell Registry (hPSCreg) has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement 726320

Contents

Introduction

Each cell line in the Human Pluripotent Stem Cell Registry (hPSCreg®) is described by a detailed dataset, including user-entered data and metadata. We focused out, that a simple dataset might not be sufficient to display all associated data to characterise a cell line. Some items, like diseases or gene mutation, requires a more complex and comprehensive data description method.

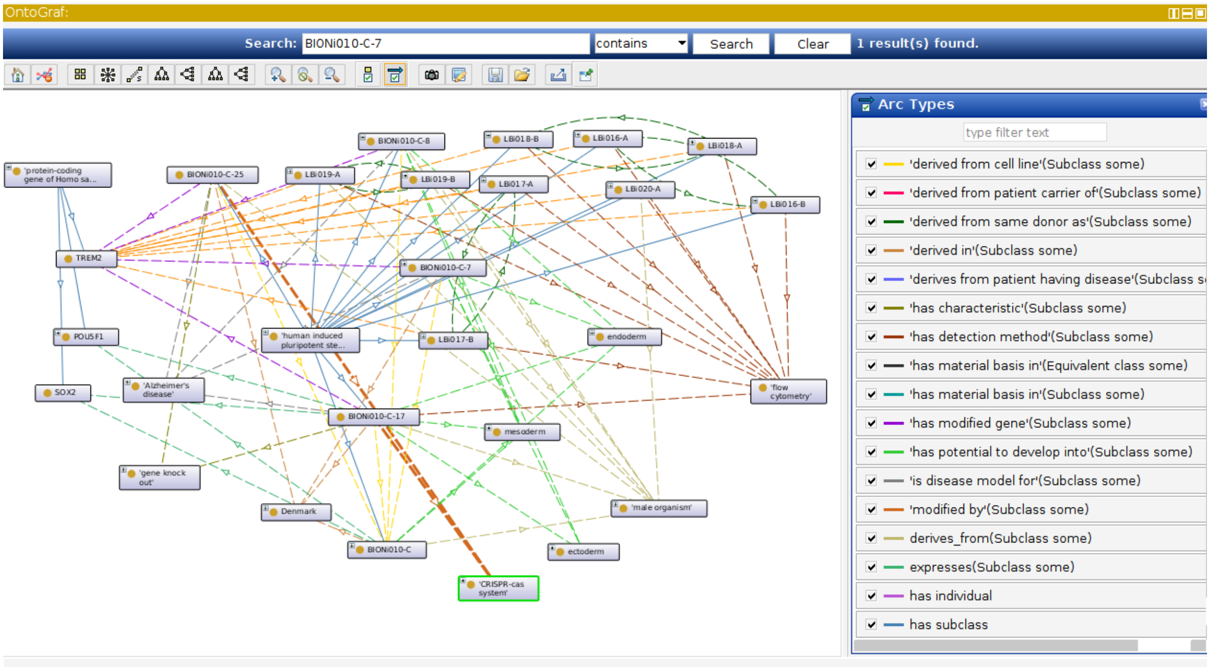
We decided to use a semantic data description by an ontology to archive this goal. For more information about ontologies, please consult Appendix 1.

hPSCreg® Ontology

The aim of the hPSCreg® Ontology (available at https://hpscreg.eu/ontologies/) is to provide fully semantic descriptions of the data and metadata of the human pluripotent stem cell lines (hPSCs) registered in hPSCreg® (human pluripotent stem cell registry) and to make the cell lines more discoverable for users.

As this ontology describes cell lines, it is based on the Cell Line Ontology[1]. Several commonly available ontologies have been imported to enable the most comprehensive possible descriptions of all important metadata. Those include information about cell types, cell lines diseases, employed experimental methods, anatomical entities, genes and proteins.

The following picture shows a short excerpt of the global description of a cell line including some associated metadata.



Search: BIONi010-C-7 contains

1 result(s) found.

Arc Types

☑ 'derived from cell line'(Subclass some)
☑ 'derived from patient carrier of'(Subclass some)
☑ 'derived from same donor as'(Subclass some)
☑ 'derived in'(Subclass some)
☑ 'derives from patient having disease'(Subclass s
☑ 'has characteristic'(Subclass some)
☑ 'has detection method'(Subclass some)
☑ 'has material basis in'(Equivalent class some)
☑ 'has material basis in'(Subclass some)
☑ 'has modified gene'(Subclass some)
☑ 'has potential to develop into'(Subclass some)
☑ 'is disease model for'(Subclass some)
☑ 'modified by'(Subclass some)
☑ derives_from(Subclass some)
☑ expresses(Subclass some)
☑ has individual
☑ has subclass

Figure 1: Semantic description of a specific cell line

[1] https://www.ebi.ac.uk/ols/ontologies/clo

Semantic Linkage to Diseases

An important information is the connection of a cell line to a certain disease.

This feature is particularly important to provide users who search for cell lines with the most appropriate matches, e.g. matches that relate to a specific disease context or genetic mutation/variant.

This connection can exist in two different ways. On the one hand, we have information about the donor of the line and his/her diseases (affected or unaffected). Thus, cell lines can be linked to diseases, which have been diagnosed in the donor, or cell lines can possess disease-related mutations, which have been typed in the donor, who carries the disease mutation.

On the other hand, a line itself can be genetically modified and in this way serve as a role model (or “experimental tool”) for investigating disease mechanisms (see Figure 2).

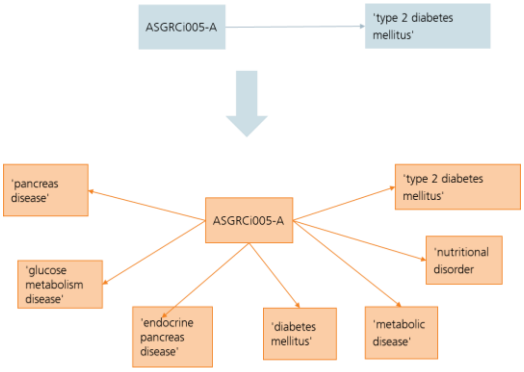
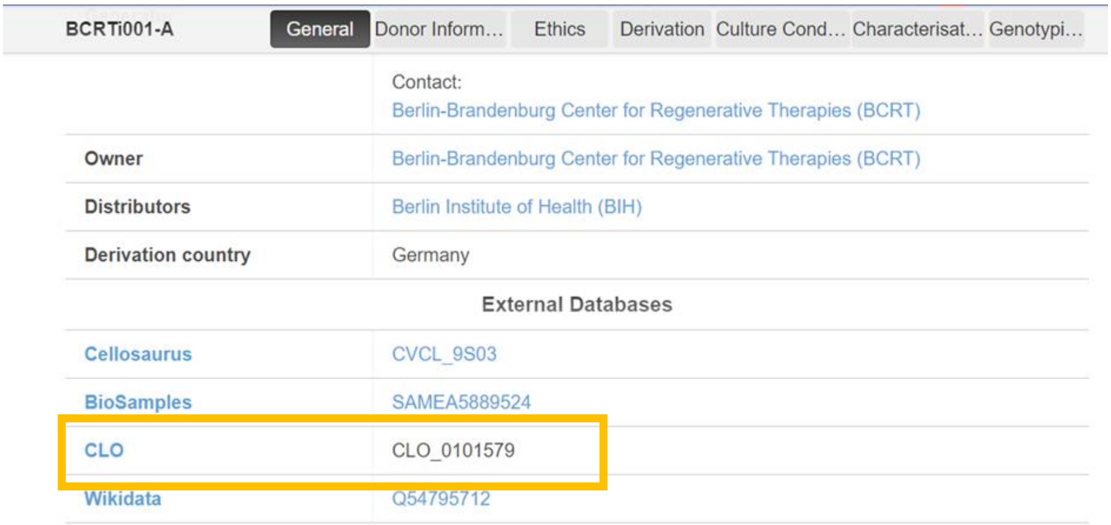

ASGRCi005-A
'type 2 diabetes mellitus'
'pancreas disease'
ASGRCi005-A
'type 2 diabetes mellitus'
'nutritional disorder'
'glucose metabolism disease'
'endocrine pancreas disease'
'diabetes mellitus'
'metabolic disease'

Figure 2: Linkage and detailed sematic information of a disease

Cell Line Identifiers

Every cell line in the hPSCreg® Ontology is described by a CLO_ID, because of its relation to the Cell Line Ontology. This CLO_ID is also part of the cell line’s metadata in the hPSCreg® user interface (see Figure 3).



	Contact: Berlin-Brandenburg Center for Regenerative Therapies (BCRT)
Owner	Berlin-Brandenburg Center for Regenerative Therapies (BCRT)
Distributors	Berlin Institute of Health (BIH)
Derivation country	Germany
External Databases	
Cellosaurus	CVCL_9S03
BioSamples	SAMEA5889524
CLO	CLO_0101579
Wikidata	Q54795712

Figure 3: CLO_ID of cell line BCRTi001-A in the hPSCreg® user interace

IRIs in the hPSCreg® Ontology

Every class in the ontology has a unique identifier called IRI[2] .

The following example explains the IRI behaviour in hPSCreg®:

- Cell line name: BCRTi001-A
- IRI: http://purl.obolibary.com/obo/CLO_0101579
  - General part: http://purl.obolibary.com/obo/ (will not change for ontologies, that a part of the OBO-Foundry[3])
  - Variable part: CLO_0101579 (specific ID)

An easy way to analyse the content of the hPSCreg® ontology to use the software programme Protégé (see Appendix 2).

SPARQL Queries

Instead of using Protégé, the related information of a cell line can also be accessed by SPARQL (see Appendix 1 for a short introduction in SPARQL).

SPARQL Interface hPSCreg Platform

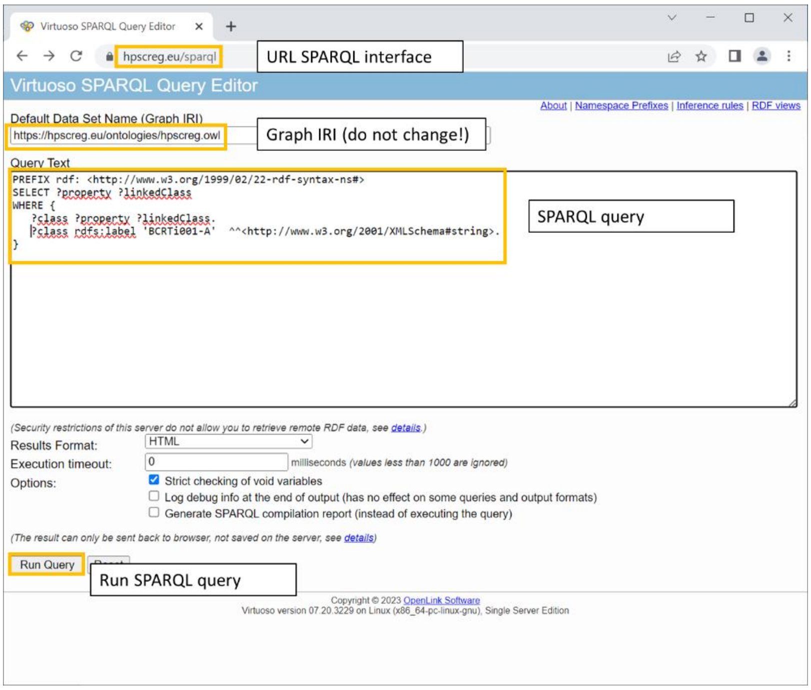
The SPARQL interface of the hPSCreg platform can be reached via https://hpscreg.eu/sparql (see Figure 4).

In the field "*Query Text*", you can enter your SPARQL Query

"*Run Query*" will show you the result of Query in the Result view (see Figure 4).

[2] Internationalized Resource Identifiers (IRIs) (w3.org)

[3] https://obofoundry.org/



Virtuoso SPARQL Query Editor
hpscreg.eu/sparql
URL SPARQL interface
Virtuoso SPARQL Query Editor
About | Namespace Prefixes | Inference rules | RDF views
Graph IRI (do not change!)
SPARQL query
Run Query
Run SPARQL query

Default Data Set Name (Graph IRI)
https://hpscreg.eu/ontologies/hpscreg.owl

Query Text

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?property ?linkedClass
WHERE {
   ?class ?property ?linkedClass.
   ?class rdfs:label 'BCRTi001-A'  ^^<http://www.w3.org/2001/XMLSchema#string>.
}
```

(Security restrictions of this server do not allow you to retrieve remote RDF data, see details.)

Results Format: HTML

Execution timeout: 0 milliseconds (values less than 1000 are ignored)

Options:
☑ Strict checking of void variables
☐ Log debug info at the end of output (has no effect on some queries and output formats)
☐ Generate SPARQL compilation report (instead of executing the query)

(The result can only be sent back to browser, not saved on the server, see details)

Copyright © 2023 OpenLink Software
Virtuoso version 07.20.3229 on Linux (x86_64-pc-linux-gnu), Single Server Edition

Figure 4: SPARQL interface hPSCreg®

SPARQL Examples

The following lines will show some SPARQL examples for querying relevant cell line information available in the hPSCreg® Ontology. The queries can be easily adopted by changing the highlighted part.

- Get all cell lines with a related donor disease

  IRIS of donor diseases in the ontology (can be replaced in owl:onProperty part):

○ Donor has disease:	http://purl.obolibrary.org/obo/CLO_0000015
○ Embryo has disease:	http://purl.obolibrary.org/obo/CLO_0000006
○ Embryo is carrier of disease:	http://purl.obolibrary.org/obo/CLO_0000005
○ Patient is carrier of disease:	http://purl.obolibrary.org/obo/CLO_0000003

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
SELECT DISTINCT (STR(?clname) AS ?line) (STR(?dislab) AS ?disease)
WHERE {
  ?dis rdfs:label ?label.
  ?label bif:contains "'neurodegenerative disease'".
  ?sub rdfs:subClassOf* ?dis.
  ?cell rdfs:subClassOf ?rest.
  ?rest owl:onProperty <http://purl.obolibrary.org/obo/CLO_0000015>.
  ?rest owl:someValuesFrom ?sub.
  ?sub rdfs:label ?dislab.
  ?cell rdfs:label ?clname.
}
GROUP by ?cell
ORDER by ?line
```

Result (shortened):

line	Disease
RCPCMi004-A	Parkinson's disease
RCPCMi005-A	Parkinson's disease
RCPCMi008-A	spinocerebellar ataxia type 17
RIi009-A	retinitis pigmentosa
RIi010-A	Leber congenital amaurosis

- Get all cell lines with a genetically modified gene related to a specific disease

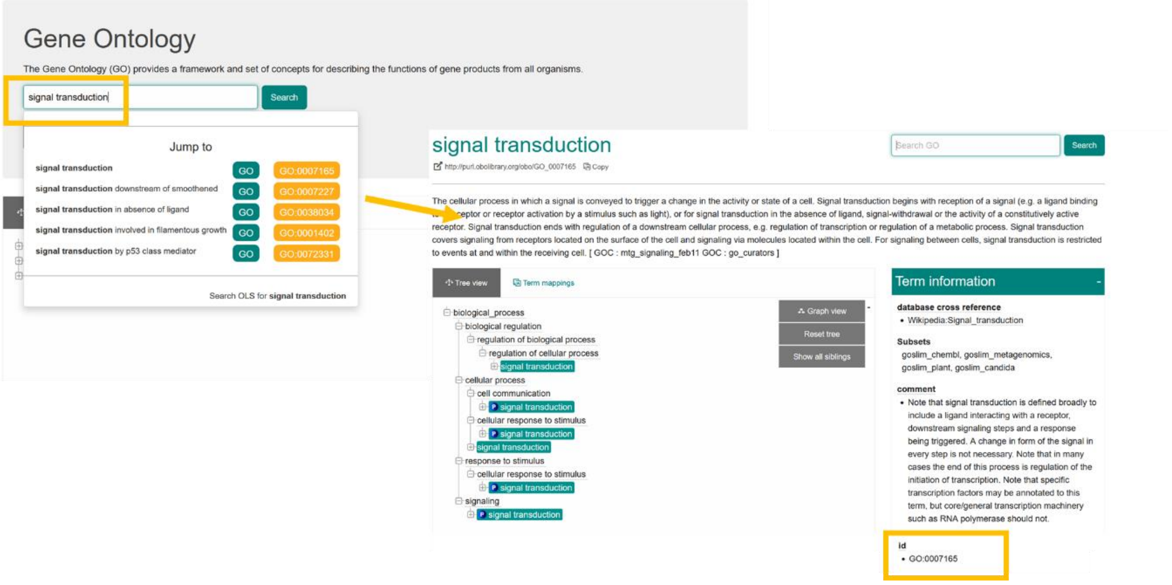
```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
SELECT DISTINCT (STR(?clname) AS ?line) (STR(?dislab) AS ?disease)
WHERE {
  ?dis rdfs:label ?label.
  ?label bif:contains "'neurodegenerative disease'".
  ?sub rdfs:subClassOf* ?dis.
  ?cell rdfs:subClassOf ?rest.
  ?rest owl:onProperty <http://purl.obolibrary.org/obo/CLO_0000179>.
  ?rest owl:someValuesFrom ?sub.
  ?sub rdfs:label ?dislab.
  ?cell rdfs:label ?clname.
}
GROUP by ?cell
 ORDER by ?line
```

Result (excerpt):

line	Disease
MPIi003-A-1	Parkinson's disease
RCi004-A-1	Huntington's disease
SCHi001-A-1	adrenoleukodystrophy
SCTCi014-A-1	age related macular degeneration
SCTCi015-A-1	age related macular degeneration

- Get all cell lines with a modified gene that plays a role in a specific biological process

1. Get ID of modifying gene from Gene Ontology
   The ID can retrieve this URL: https://www.ebi.ac.uk/ols/ontologies/go.
   Type in the name and select "search".
2. Copy the ID from the result page and paste it in the Query below.


Gene Ontology

The Gene Ontology (GO) provides a framework and set of concepts for describing the functions of gene products from all organisms.

signal transduction

Jump to

signal transduction	GO	GO:0007165
signal transduction downstream of smoothened	GO	GO:0007227
signal transduction in absence of ligand	GO	GO:0038034
signal transduction involved in filamentous growth	GO	GO:0001402
signal transduction by p53 class mediator	GO	GO:0072331

Search OLS for **signal transduction**

signal transduction

http://purl.obolibrary.org/obo/GO_0007165 Copy

The cellular process in which a signal is conveyed to trigger a change in the activity or state of a cell. Signal transduction begins with reception of a signal (e.g. a ligand binding to a receptor or receptor activation by a stimulus such as light), or for signal transduction in the absence of ligand, signal-withdrawal or the activity of a constitutively active receptor. Signal transduction ends with regulation of a downstream cellular process, e.g. regulation of transcription or regulation of a metabolic process. Signal transduction covers signaling from receptors located on the surface of the cell and signaling via molecules located within the cell. For signaling between cells, signal transduction is restricted to events at and within the receiving cell. [GOC : mtg_signaling_feb11 GOC : go_curators]

```
biological_process
  biological regulation
    regulation of biological process
      regulation of cellular process
        signal transduction
  cellular process
    cell communication
      signal transduction
    cellular response to stimulus
      signal transduction
    signal transduction
  response to stimulus
    cellular response to stimulus
      signal transduction
  signaling
    signal transduction
```

Term information

database cross reference
- Wikipedia:Signal_transduction

Subsets
goslim_chembl, goslim_metagenomics, goslim_plant, goslim_candida

comment
- Note that signal transduction is defined broadly to include a ligand interacting with a receptor, downstream signaling steps and a response being triggered. A change in form of the signal in every step is not necessary. Note that in many cases the end of this process is regulation of the initiation of transcription. Note that specific transcription factors may be annotated to this term, but core/general transcription machinery such as RNA polymerase should not.

id
- GO:0007165

Figure 5: Get Id from Gene Ontology

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX obo: <http://purl.obolibrary.org/obo/>
SELECT DISTINCT (STR(?clname) AS ?line)
WHERE {
    ?class rdfs:subClassOf ?rest.
    ?rest owl:onProperty <http://purl.obolibrary.org/obo/CLO_0100021>.
    ?rest owl:someValuesFrom ?val.
    ?val obo:OGG_0000000029 ?gA.
    filter contains(?gA,"GO_0007165").
    ?class rdfs:label ?clname.
}
```

Result (excerpt):

line
MHHi001-A-7
BIONi010-C-5
BIONi010-C-9
BIHi005-A-5
TMOi001-A-1

Appendix 1 – Introduction to Ontologies and SPARQL

The following lines will give a short overview about ontologies explained by a simplified example (see Figure 6).

Ontologies

An ontology consist of elements (*classes*) that exist in a specific domain and *properties* to describe them. Properties are relationships to link two classes or *attributes* to describe a class.

The easiest way to link a class to another is the *subClassOf* property. A subClass is a more precise description to a superclass, like creature -> animal -> dog -> poodle.

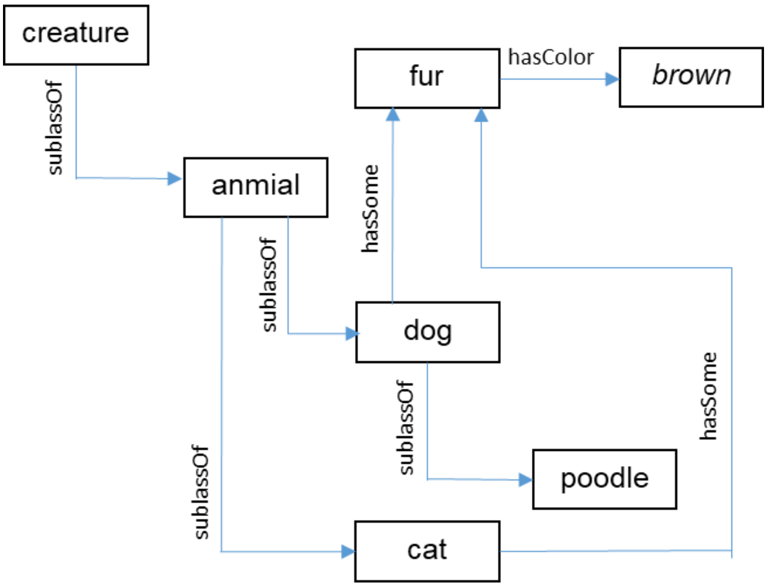


Figure 6: ontology example

It is also possible to link more than two classes. Dogs and cats are both animals.

Properties can also be a bit more complex. As you can see in the example, dogs and cats can have fur. However, fur is a subClass neither of dogs nor of cats. This connection can be realised by a specific property, which is called “hasSome” here.

So, the elements in an ontology are represented by a *graph* structure. Each element of ontology can be described by *triples* (class – property – class – property….).

Ontologies provide more features, which are going beyond this example. Detailed information can be found here https://www.w3.org/standards/semanticweb/ontology.

SPARQL

SPARQL is a query language to receive information from ontologies in RDF format (https://www.w3.org/RDF/). As these datasets are described in triples, its queries have to be constructed in that manner.

The next lines shows some simplified examples.

- all triples of a dataset:

```
SELECT * WHERE {
    graph ?g {
        ?class ?property ?linkedClass .
    }
}
```

Result:

?class	?property	?linkedClass
animal	subClassOf	creature
dog	hasSome	fur
Cat	subClassOf	animal
….	….	…..

- classes with linked by a specific property

```
SELECT ?class
WHERE {
    ?class hasSome fur.
 }
```

Result:

?class
cat
dog

- All subclasses

```
SELECT ?class
WHERE {
    ?class subClassOf dog.
}
```

Result:

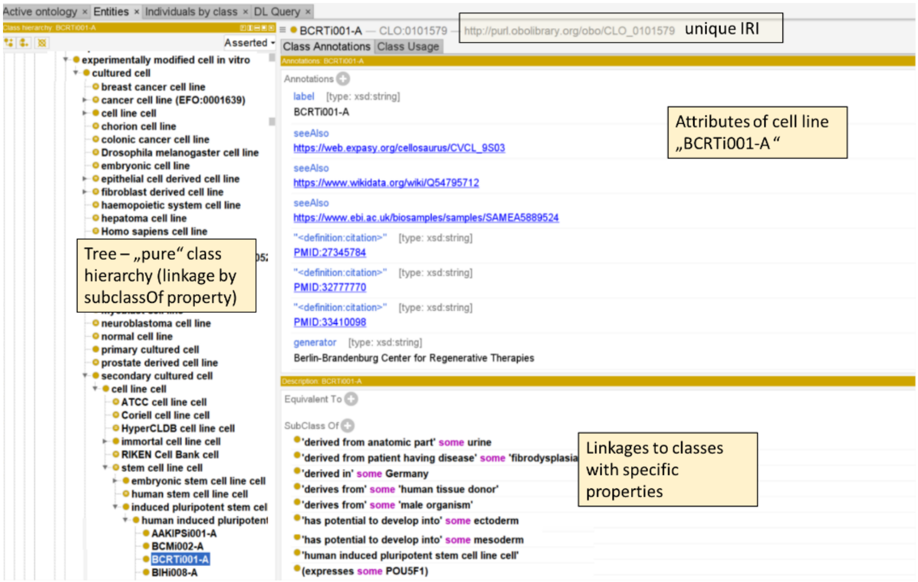
?class
poodle

SPARQL is a very comprehensive query language. More information can be found here: https://www.w3.org/TR/rdf-sparql-query/.

Appendix 2 – Protégé

Protégé GUI

An easy way to analyse the content of the hPSCreg® ontology is the tool Protégé (https://protege.stanford.edu/software.php), as displayed in the following screen. This tool can be installed on every computer and run as a stand – alone application.



- experimentally modified cell in vitro
  - cultured cell
    - breast cancer cell line
    - cancer cell line (EFO:0001639)
    - cell line cell
    - chorion cell line
    - colonic cancer cell line
    - Drosophila melanogaster cell line
    - embryonic cell line
    - epithelial cell derived cell line
    - fibroblast derived cell line
    - haemopoietic system cell line
    - hepatoma cell line
    - Homo sapiens cell line
    - neuroblastoma cell line
    - normal cell line
    - primary cultured cell
    - prostate derived cell line
    - secondary cultured cell
      - cell line cell
        - ATCC cell line cell
        - Coriell cell line cell
        - HyperCLDB cell line cell
        - immortal cell line cell
        - RIKEN Cell Bank cell
        - stem cell line cell
          - embryonic stem cell line cell
          - human stem cell line cell
          - induced pluripotent stem cel
            - human induced pluripotent
              - AAKIPSi001-A
              - BCMi002-A
              - BCRTi001-A
              - BIHi008-A

Tree – „pure“ class hierarchy (linkage by subclassOf property)

BCRTi001-A — CLO:0101579 — http://purl.obolibrary.org/obo/CLO_0101579 unique IRI

Annotations

label [type: xsd:string]
BCRTi001-A

seeAlso
https://web.expasy.org/cellosaurus/CVCL_9S03

seeAlso
https://www.wikidata.org/wiki/Q54795712

seeAlso
https://www.ebi.ac.uk/biosamples/samples/SAMEA5889524

"<definition:citation>" [type: xsd:string]
PMID:27345784

"<definition:citation>" [type: xsd:string]
PMID:32777770

"<definition:citation>" [type: xsd:string]
PMID:33410098

generator [type: xsd:string]
Berlin-Brandenburg Center for Regenerative Therapies

Attributes of cell line „BCRTi001-A“

Equivalent To

SubClass Of

- 'derived from anatomic part' some urine
- 'derived from patient having disease' some 'fibrodysplasia
- 'derived in' some Germany
- 'derives from' some 'human tissue donor'
- 'derives from' some 'male organism'
- 'has potential to develop into' some ectoderm
- 'has potential to develop into' some mesoderm
- 'human induced pluripotent stem cell line cell'
- (expresses some POU5F1)

Linkages to classes with specific properties

Figure 7: Details of cell line "BCRTi001-A" in Protégé.