European Bank for induced pluripotent Stem Cells

Semantic Queries in EBiSC

1.4

31.05.2023

The EBiSC – European Bank for induced pluripotent Stem Cells project has received support from the Innovative Medicines Initiative Joint Undertaking under grant agreements n° 115582 and 821362, resources of which are composed of financial contribution from the European Union and EFPIA companies' in kind contribution.
www.imi.europa.eu

Contents

Introduction

Each cell line in the European Bank for induced pluripotent Stem Cells (EBiSC) catalogue is described by a detailed dataset, including data and metadata provided by the cell line depositor. 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.

EBiSC Ontology

The aim of the EBiSC Ontology (available at https://ebisc.org/ontologies/ebisc.owl) is to provide fully semantic descriptions of the data and metadata of pluripotent Stem Cells registered in the EBiSC platform and to make the cell lines more discoverable for EBiSC 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. They 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.

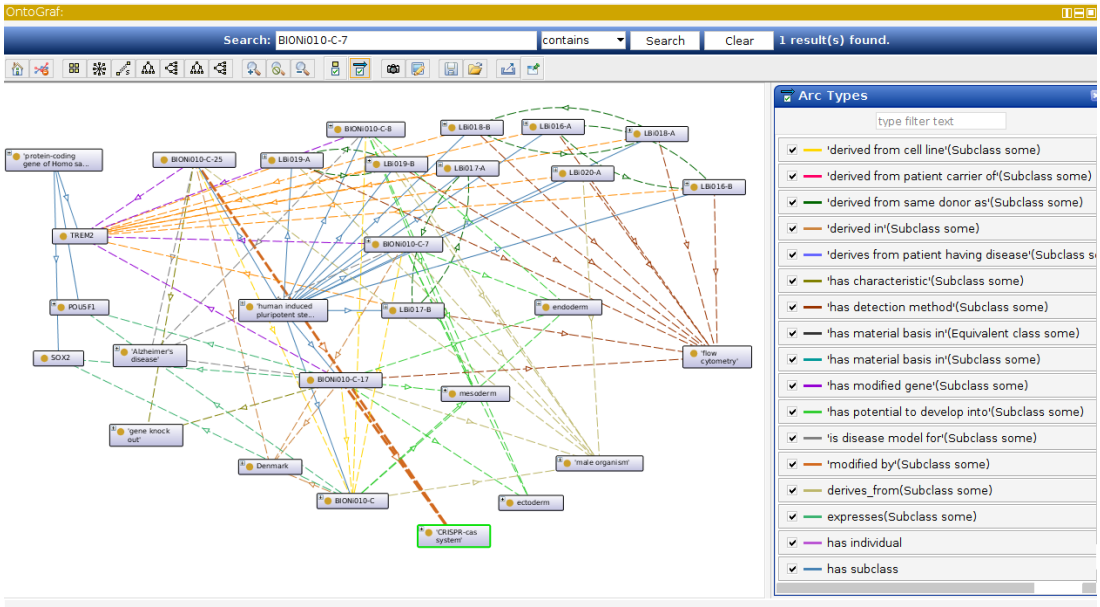


Figure 1: Semantic description of a specific cell line

[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).

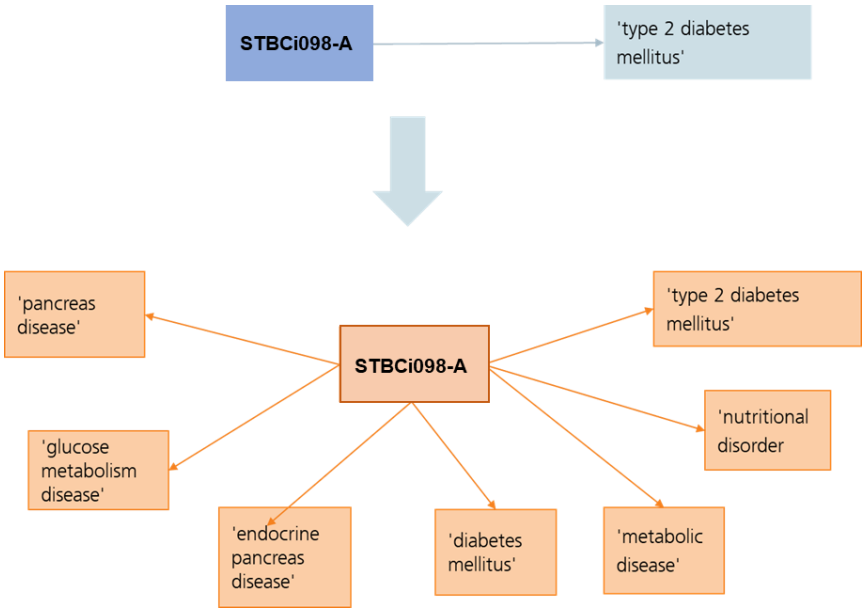
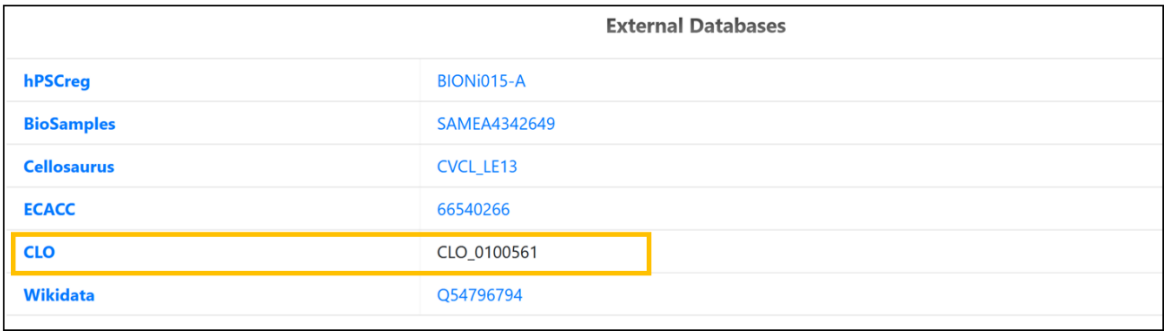


Figure 2: Linkage and detailed sematic information of a disease

Cell Line Identifiers

Every cell line in the EBiSC 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 EBiSC user interface (see Figure 3).



External Databases	
hPSCreg	BIONi015-A
BioSamples	SAMEA4342649
Cellosaurus	CVCL_LE13
ECACC	66540266
CLO	CLO_0100561
Wikidata	Q54796794

Figure 3: CLO_ID of cell line BIONi015-A in the EBiSC catalogue

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

The following example explains the IRI behaviour in EBiSC:

- Cell line name: STBCi098-A
- IRI: http://purl.obolibrary.org/obo/CLO_0101911
    - General part: http://purl.obolibary.com/obo/ (will not change for ontologies, that a part of the OBO-Foundry[3])
    - Variable part: CLO_0101911 (specific ID)

An easy way to analyse the content of the EBiSC 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 EBiSC Platform
The SPARQL interface of the EBiSC platform can be reached via https://ebisc.org/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/

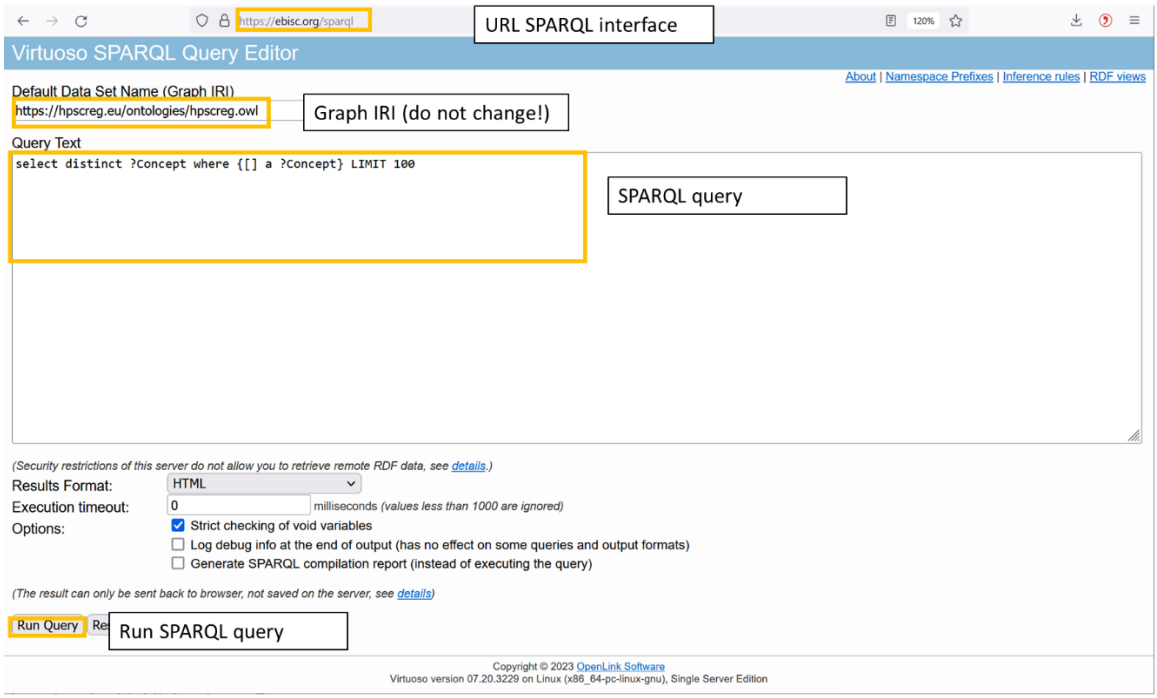


Figure 4: SPARQL interface for EBiSC

SPARQL Examples

The following lines will show some SPARQL examples for querying relevant cell line information available in the EBiSC 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
  - 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.
  ?cell rdfs:seeAlso ?so.
  filter contains(STR(?so),"ebisc").

}
GROUP by ?cell
ORDER by ?line
```

Result (excerpt):

line	disease
BIONi010-C-41	myotonic dystrophy type 1
BIONi010-C-42	myotonic dystrophy
BIONi010-C-43	myotonic dystrophy
CBRCULi002-A	myotonic dystrophy type 1
CENSOi008-A	myotonic dystrophy

- 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.
        ?cell rdfs:seeAlso ?so.
        filter contains(STR(?so),"ebisc").
}
GROUP by ?cell
ORDER by ?line
```

Result (excerpt):

line	disease
BIONi010-C-17	Alzheimer disease
BIONi010-C-2	Alzheimer's disease
BIONi010-C-25	Alzheimer's disease
BIONi010-C-3	Alzheimer disease
BIONi010-C-4	Alzheimer's disease

- 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.

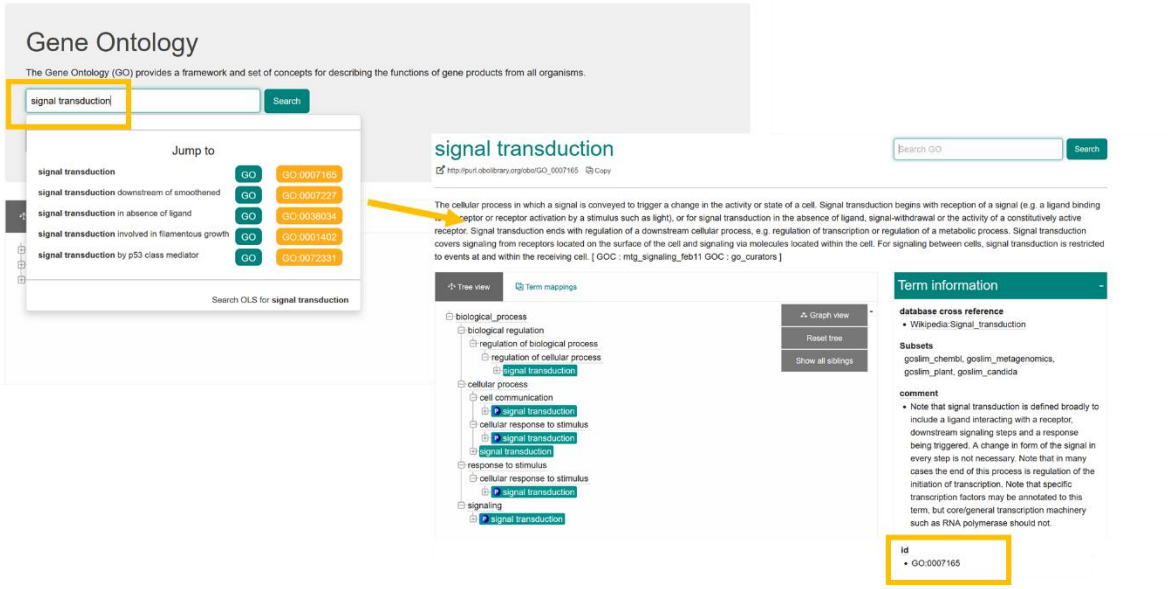


Figure 5: Get Id from Gene Ontology

```
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.
      ?class rdfs:seeAlso ?so.
      filter contains(STR(?so),"ebisc").
}
```

Result (excerpt):

line
BIONi010-C-9
BIONi010-C-5

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.

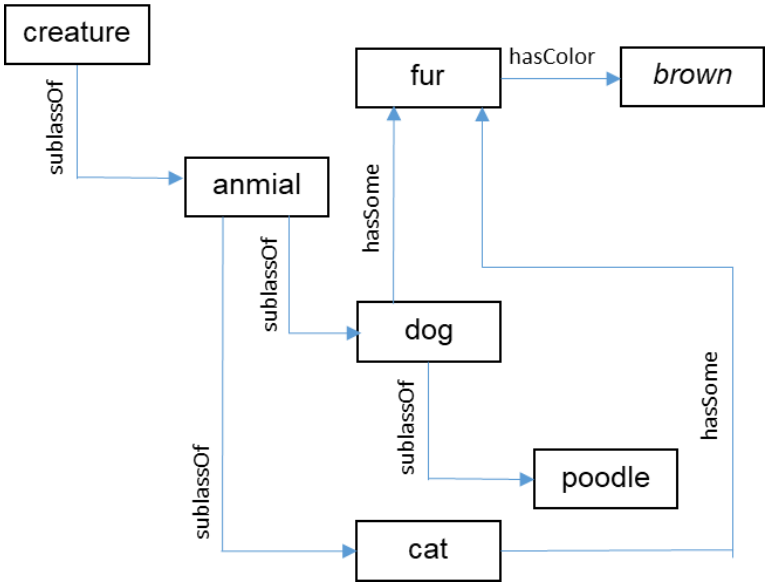


Figure 6: ontology example

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 EBiSC 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 standalone application.

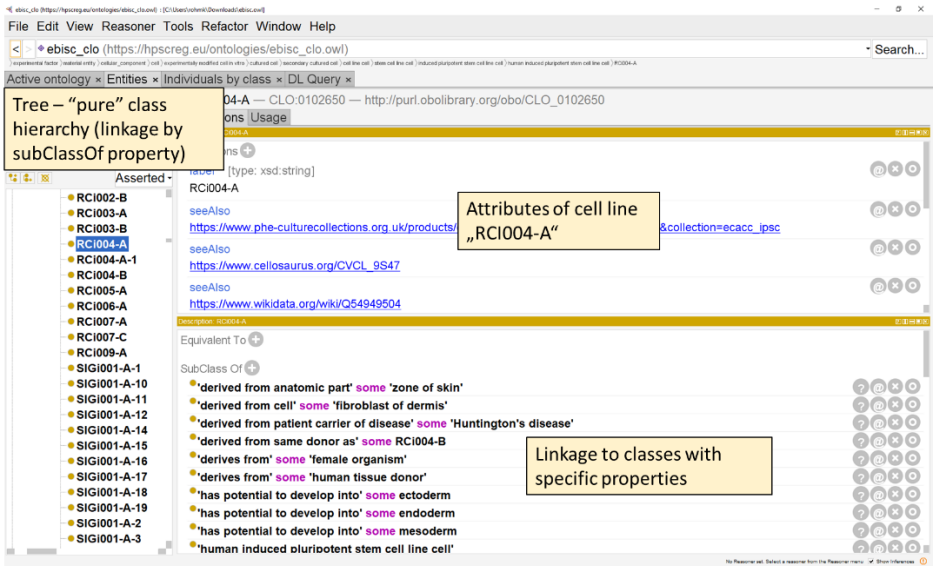


Figure 7: Details of cell line "RCI004-A" in Protégé.