

How to implement and verify FAIR principles ?

Alban Gaignard

CNRS, institut du thorax, Nantes, France

November 24, 2024

ANF "Workflow et reproductibilité" 2025, Paris

Who am I?

l'institut du thorax

Better understanding of
cardio-vascular and
metabolic diseases

Gene \longleftrightarrow **function**

associations

Translational medicine

university hospital \oplus reseach lab

Bioinformatics

- ▶ Massive production of genomic
sequence & health data
→ Workflows + HPC
- ▶ Integration of multi-modal and
multi-scale data
- ▶ Predictive models

French Institute for Bioinformatics



A national research infrastructure for Bioinformatics providing:

Compute & Storage

Tools & Workflows, Databases, Training, Open Sciences

Communities: health, agronomy, biodiversity, microbiology





Actively involved in

- ▶ **Open sciences & interoperability:** FAIR-Checker, metadata standards, ontologies (Bioschemas, EDAM)
- ▶ **Health community:** genomic data discoverability & sharing (Beacon, FEGA) + data integration (Knowledge Graphs)





A practical overview on FAIR assessment, in 2 hours ...

- ▶  **Quick intro on FAIR principles**

- ▶ **Knowledge graphs and semantic metadata**

-  What are Knowledge Graphs, Computational Ontologies ?
-  How to create Knowledge Graphs ? (team work, "pen & paper")
-  How to write machine-readable semantic metadata ? (demo, python code)
-  How to query knowledge graphs with SPARQL (demo, python code)

- ▶ **Tools and resources for increased FAIRness**

-  FAIR-Checker
-  Tools and workflow registries
-  Using FAIR-Checker and interpreting the FAIR assessment results
-  Using the FAIR-Checker API and processing multiple resources

FAIR principles

Being **findable** by both human and machines

A screenshot of a Google search for the word "pasteur". The search bar contains "pasteur". Below the search bar, there are tabs for "All", "Images", "Videos", "News", "Short videos", "Web", "Books", "More", and "Tools". There are also filters for "Vaccine", "Religion", "Meaning", "Pronunciation", "Voyage", "Experiment", "Institut Pasteur", and "Education". The search results show a sponsored link for "Institut Pasteur" with the text "Faire un Don | Soutenez le Pasteurdon | Défendre la Recherche". Below that is a link for "Institut Pasteur Site Officiel". The main result is from Wikipedia, titled "Louis Pasteur", with a small portrait of him. Below the Wikipedia result, there is a "People also ask" section with the question "Quelle différence entre pasteur et prêtre ?".

A screenshot of a Google search for "louis pasteur". The search bar contains "louis pasteur". Below the search bar, there are tabs for "All", "Images", "Videos", "News", "Short videos", "Web", "Books", "More", and "Tools". The search results show a knowledge panel for "Louis Pasteur" with the subtitle "French chemist". The panel includes a large portrait of Louis Pasteur, a smaller portrait, and a video thumbnail titled "TOUT SAVOIR SUR LOUIS PASTEUR". To the right of the portraits, there are two boxes: "Born 27 Dec 1822 Dole" and "Died 28 Sept 1895 Marnes-la-Coquette". Below these are two more boxes: "Académie de Lille" and "Académie de Lille Qui était Louis Pasteur". Below the knowledge panel, there is a Wikipedia result for "Louis Pasteur" with a small portrait. Below the Wikipedia result, there is a "People also ask" section with questions like "Qu'est-ce que Pasteur a inventé ?" and "Pourquoi Louis Pasteur est-il connu ?".

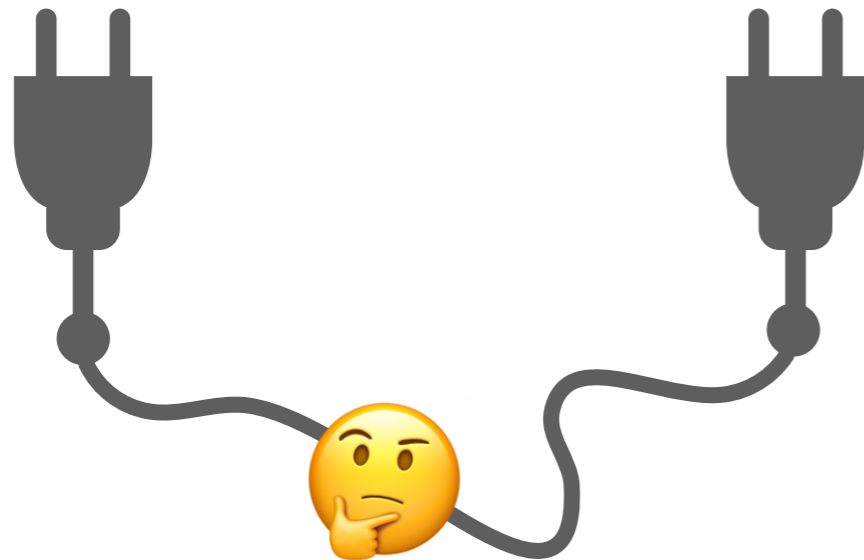
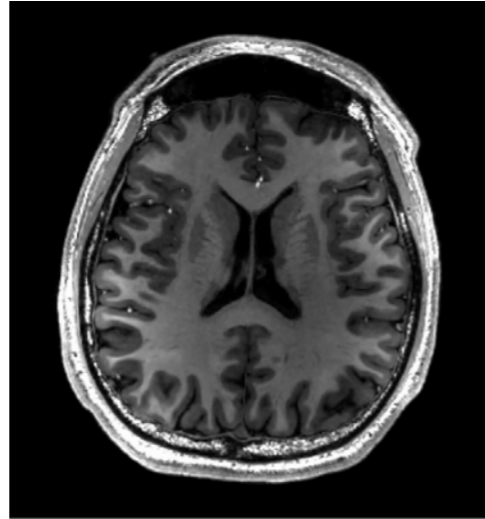
- ▶ "Pasteur" is not a good **identifier**
 - used to designate many "Humans"
 - used to designate many locations or institutions

We need **Persistent Globally Unique Identifiers**

- ▶ **Rich metadata** is key to identify **relevant** data
 - Entity kind/nature
 - Relationships with other entities

Being interoperable

```
@HWI-ST534_129:2:24:20503:16510:CGATGT
CTGAGAGCCGGGAAGCCGCGGAGCCGGGGACTGGCGAGCCGGAACAT
+
HHHHHHHHHHEFDDGDDFBFGG>7D4<9;<&?; ;<DC>CCDD@?= ?A###
@HWI-ST534_129:2:42:2118:9580:CGATGT
GGCGGAGCCGGGTGACTGGCGAGCCGGAACATCAGGCGCCGCCGAGAGA
+
GEECGGGBIDF6FFFFEF=IDFBEE8E8E?EEB@6=9B#####
@HWI-ST534_129:2:2:12654:80229:CGATGT
CGGAGCCGGGTGACTGGCGAGCCGGAACATCAGGCGCCGCCGAGAGAAG
+
GGEGFDCBBAEEEEGGFGFG;EGEEGFFBDEBDFGFCFF;DF2D<DD
@HWI-ST534_129:2:48:12356:179714:CGATGT
GAGCCGGGTGACTGGCGAGCCGGAACATCAGGCGCCGCCGAGAGAAGAA
+
E=GHFHEGHHBCGDDBEEBBCBDDDE@EGBD=ABDCB?EC;@@@EEB;E
@HWI-ST534_129:2:44:8225:39540:CGATGT
GGGTGACTGGCGAGCCGGAACATCAGGCGCCGCCGAGAGAAGAAGTATG
+
HHHEHHHHHHHHHGHHHHFHHHHHHHHDFDHHBHFDFFEFEFF>G<CCCE
```



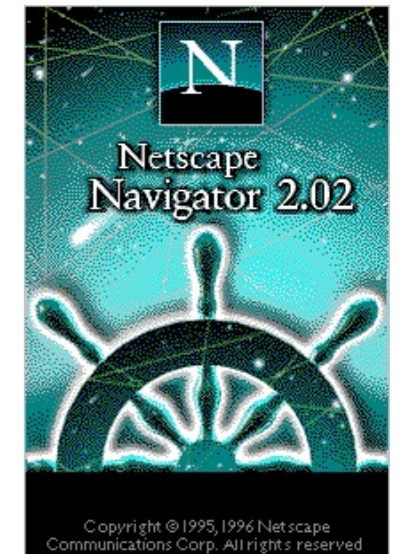
- ▶ **Technical**: exchange protocols compatible with different systems, e.g., HTTPS
- ▶ **Syntactic**: data and metadata structure and syntax, e.g., format, read/written by different systems
- ▶ **Semantic**: data and metadata can be understood/actioned upon by different systems
→ ontologies and vocabularies

Being **accessible** by both human and machines

- ▶ **Open (and free) communication protocols** (e.g. web standards) for everyone (human and software agents) ; CSV is a non-proprietary format

International Phonetic Alphabet	
[aɪ p^hiː eɪ]	
"IPA", <i>transcribed narrowly</i> according to Received Pronunciation as [aɪ p ^h iː eɪ]	
Script type	Alphabet – partially <i>featural</i>
Period	1888–present
Languages	Used for <i>phonetic</i> and <i>phonemic</i> transcription of any oral language

Hypertext Transfer Protocol	
	
Début d'une adresse web HTTP dans la barre d'adresse d'un navigateur web.	
Informations	
Fonction	Transmission d'hyper texte
Sigle	HTTP
Date de création	1990
Auteur(s) / Autrice(s)	Tim Berners-Lee
Port	80
RFC	1996 : RFC 1945 1997 : RFC 2068 1999 : RFC 2616 2014 : RFC 7230 à 7237 2015 : RFC 7540
modifier	



- ▶ Ability to specify **access control**



Being **re-usable**

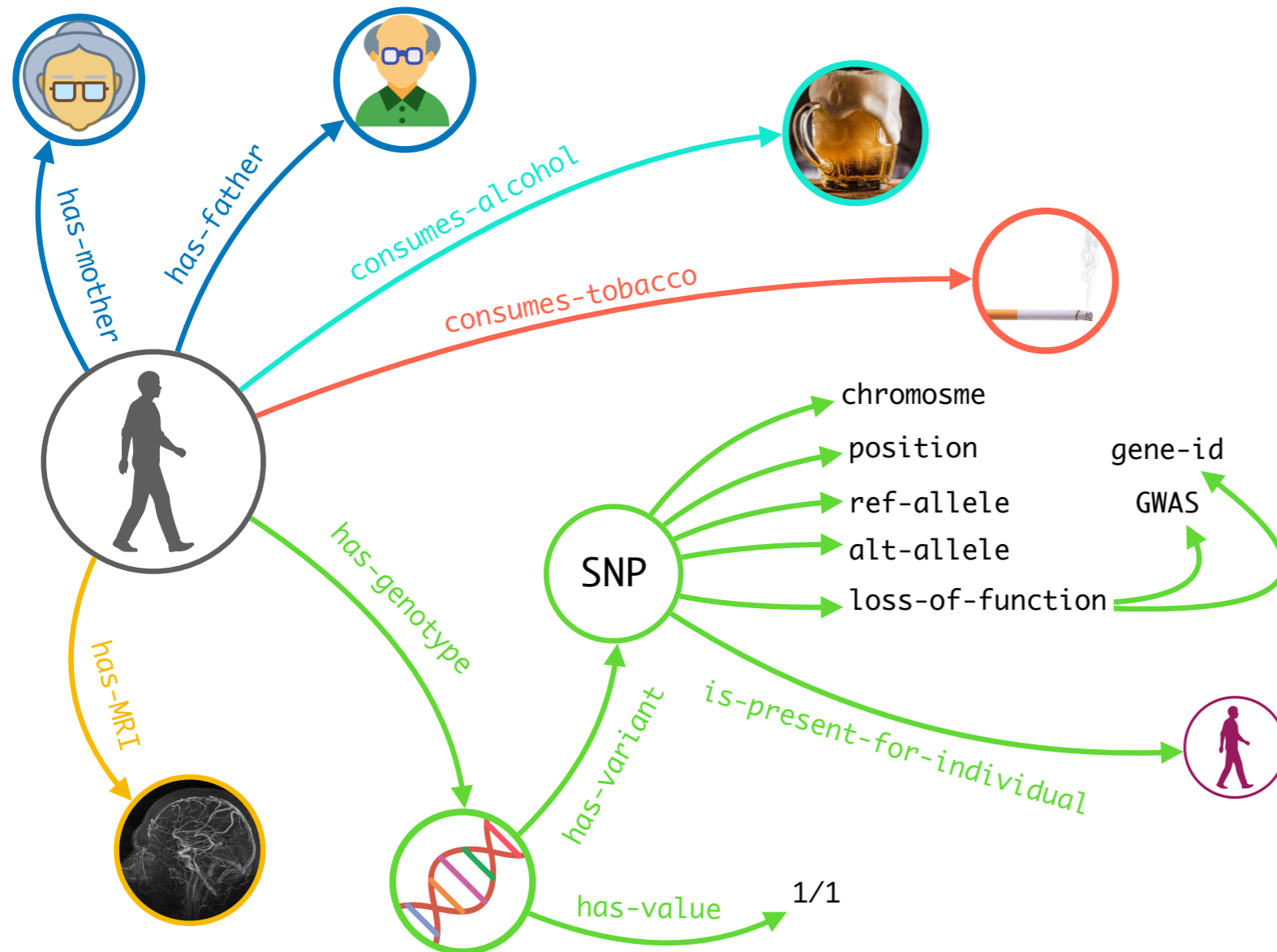
- ▶ Clearly define a **usage license**
→ legal condition for (re)use
- ▶ Cumulative sciences need **provenance**
 - crediting
 - funders
 - creators
 - fine-grained documentation of **data production**
fine-grained documentation of **software components**
- ▶ **Community standards**



Knowledge Graphs

1 Link data with Knowledge Graphs

« a collection of interlinked descriptions of things
(real-world objects, abstract concepts, events, etc.) »



- ✓ a **database** to store and retrieve information
- ✓ a **graph** to represent multiple relationships and to perform network / community analysis
- ✓ a **knowledge base** with formal semantics to perform logical reasoning, inferences ...

Wikipedia → DBpedia Knowledge Graph

RAC1

From Wikipedia, the free encyclopedia

"Rac1" redirects here. For the first game in the Ratchet & Clank series, see *Ratchet & Clank*.

Rac1, also known as **Ras-related C3 botulinum toxin substrate 1**, is a protein found in human cells. It is encoded by the *RAC1* gene.^{[6][7]} This gene can produce a variety of alternatively spliced versions of the Rac1 protein, which appear to carry out different functions.^[7]

Function [edit]

Rac1 is a small (~21 kDa) signaling G protein (more specifically a GTPase), and is a member of the Rac subfamily of the family Rho family of GTPases. Members superfamily appear to regulate a diverse array of cellular events, including the control of GLUT4^{[8][9]} translocation to glucose uptake, cell growth, cytoskeletal reorg, antimicrobial cytotoxicity,^[10] and the activation of protein kinases.^[11]

Rac1 is a pleiotropic regulator of many cellular processes, including the cell cycle, cell-cell adhesion, motility (through the actin network), and of epithelial differentiation (proposed to be necessary for maintaining epidermal stem cells).

Role in cancer [edit]

Along with other subfamily of Rac and Rho proteins, they exert an important regulatory role specifically in cell motility and cell growth. Rac1 has ubiquitous tissue and drives cell motility by formation of lamellipodia.^[12] In order for cancer cells to grow and invade local and distant tissues, deregulation of cell motility is one of the hallmarks in cancer cell invasion and metastasis.^[13] Overexpression of a constitutively active Rac1 V12 in mice caused a tumor that's phenotypically indistinguishable from sarcoma.^[14] Activating or gain-of-function mutations of Rac1 are shown to play active roles in promoting mesenchymal-type of cell movement assisted by NEDD9 protein complex.^[15] Such abnormal cell motility may result in epithelial mesenchymal transition (EMT) – a driving mechanism for tumor metastasis as well as drug relapse.^{[16][17]}

Role in glucose transport [edit]

Rac1 is expressed in significant amounts in insulin sensitive tissues, such as adipose tissue and skeletal muscle. Here Rac1 regulated the translocation of glucose GLUT4 vesicles from intracellular compartments to the plasma membrane.^{[9][18][19]} In response to insulin, this allows for blood glucose to enter the cell to lower blood

About: RAC1

An Entity of Type : Biomolecule, from Named Graph : <http://dbpedia.org>, within Data Space : dbpedia.org

Rac1, also known as Ras-related C3 botulinum toxin substrate 1, is a protein found in human cells. It is encoded by the RAC1 gene. This gene can produce a variety of alternatively spliced versions of the Rac1 protein, which appear to carry out different functions.

Property	Value
<code>dbo:abstract</code>	<ul style="list-style-type: none"> Rac1 (RAS-related C3 botulinus toxin substrate 1) は、ヒト細胞に存在するタンパク質であり、RAC1遺伝子によりコードされている。RAC1は選択的スプライシングにより異なる機能を持ったいくつかのタンパク質を生成しており、このうちの1つがRac1である。Rac1は、悪性黒色腫や肺非小細胞癌を含むさまざまな癌の発生において、重要な役割を果たしていると考えられている。そのため、現在これらの疾患に対する治療標的と考えられている。 (ja) Rac1 (англ. Ras-related C3 botulinum toxin substrate 1) — внутриклеточный белок из суперсемейства ГТФаз, относится к «малым» G-белкам. Находится в двух состояниях: активном ГТФ-связанном и неактивном ГДФ-связанном состоянии. В своей активной форме Rac1 связывается в клетке с целым рядом эффекторных белков и приводит к регулировке многих биологических процессов, таких как секреция, фагоцитоз апоптозных клеток, поляризация эпителиальных клеток и образование факторами роста образование мембранных складок и выростов (англ. membrane ruffles). (ru) Rac1, also known as Ras-related C3 botulinum toxin substrate 1, is a protein found in human cells. It is encoded by the RAC1 gene. This gene can produce a variety of alternatively spliced versions of the Rac1 protein, which appear to carry out different functions. (en)

<http://dbpedia.org/sparql>

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)
<http://dbpedia.org>

Query Text

```
SELECT DISTINCT ?gene ?entrez_id ?uniprot_id WHERE {
  ?gene dbo:abstract ?abstract .
  FILTER (regex(?abstract, "toxin")).
  ?gene dbo:entrezgene ?entrez_id .
  OPTIONAL {?gene dbo:uniprot ?uniprot_id} .
}
```

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

Results Format: HTML

Execution timeout: 30000 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](#))

Run Query Reset

gene	entrez_id	uniprot_id
http://dbpedia.org/resource/DsbA	"948353"	"P0AEG4"
http://dbpedia.org/resource/Cholinesterase	"590"	"P06276"
http://dbpedia.org/resource/Cholinesterase	"590"	"P22303"
http://dbpedia.org/resource/Cholinesterase	"43"	"P06276"
http://dbpedia.org/resource/Cholinesterase	"43"	"P22303"
http://dbpedia.org/resource/Clostridium_perfringens_alpha_toxin	"988262"	
http://dbpedia.org/resource/Lymphotoxin	"4049"	"P01374"
http://dbpedia.org/resource/Lymphotoxin	"4049"	"Q06643"
http://dbpedia.org/resource/Lymphotoxin	"4050"	"P01374"
http://dbpedia.org/resource/Lymphotoxin	"4050"	"Q06643"
http://dbpedia.org/resource/Casein_kinase_2	"1457"	"P19784"
http://dbpedia.org/resource/Casein_kinase_2	"1457"	"P67870"
http://dbpedia.org/resource/Casein_kinase_2	"1457"	"P68400"
http://dbpedia.org/resource/Casein_kinase_2	"1460"	"P19784"
http://dbpedia.org/resource/Casein_kinase_2	"1460"	"P67870"
http://dbpedia.org/resource/Casein_kinase_2	"1460"	"P68400"
http://dbpedia.org/resource/Casein_kinase_2	"1459"	"P19784"
http://dbpedia.org/resource/Casein_kinase_2	"1459"	"P67870"
http://dbpedia.org/resource/Casein_kinase_2	"1459"	"P68400"
http://dbpedia.org/resource/Collagenase	"4317"	"P03956"
http://dbpedia.org/resource/Collagenase	"4317"	"P22894"
http://dbpedia.org/resource/Collagenase	"4312"	"P03956"
http://dbpedia.org/resource/Collagenase	"4312"	"P22894"
http://dbpedia.org/resource/Guanylin	"2980"	"Q02747"
http://dbpedia.org/resource/Macrophage_inflammatory_protein	"6348"	"P10147"
http://dbpedia.org/resource/Macrophage_inflammatory_protein	"6348"	"P13236"
http://dbpedia.org/resource/Macrophage_inflammatory_protein	"6351"	"P10147"
http://dbpedia.org/resource/Macrophage_inflammatory_protein	"6351"	"P13236"

cmkb.cellmigration.org/report.cgi?report=orth_overview&gene_id=5879

www.cellmigration.org/index.shtml

705 (xsd:integer)

5305 (xsd:integer)

ng

:Q206229

:Q8054

molecule

tein

Uniprot Knowledge Graph

Namespaces
 up_core: <http://purl.uniprot.org/core/>
 uniprot: <http://purl.uniprot.org/uniprot/>
 up_citations: <http://purl.uniprot.org/citations/>
 up_taxonomy: <http://purl.uniprot.org/taxonomy/>
 up_annotations: <http://purl.uniprot.org/annotation/>
 up_keywords: <http://purl.uniprot.org/keywords/>
 up_isoforms: <http://purl.uniprot.org/isoforms/>
 ec: <http://purl.uniprot.org/enzyme/>
 go: <http://purl.uniprot.org/go/>
 rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
 rdfs: <http://www.w3.org/2000/01/rdf-schema#>
 owl: <http://www.w3.org/2002/07/owl#>
 skos: <http://www.w3.org/2004/02/skos/core#>

Functionⁱ
 Plasma membrane-associated small GTPase which cycle as secretory processes, phagocytosis of apoptotic cells, Rac1 p21/rho GDI heterodimer is the active component regulation of cell migration and adhesion assembly and

```

isoform:P06213-1 a up:Simple_Sequence ;
up:modified "2010-10-05"^^xsd:date ;
up:version 4 ;
up:precursor true ;
up:mass 156333 ;
    
```

There are 217,505,202,099 triples in this release. All triples are available in the default graph. There are 22 named graphs corresponding to specific datasets.

Graph	Documentation	Triples	Distinct subjects	Distinct predicates	Distinct classes	Distinct objects	License
uniparc	Documentation	160,189,731,20040,455,837,024	29	6	46,916,767,863	http://creativecommons.org/licenses/by/4.0/	
uniprot	Documentation	44,256,643,227	9,441,439,078	124	121	8,462,262,751	http://creativecommons.org/licenses/by/4.0/
uniref	Documentation	10,224,623,630	1,393,813,725	14	3	1,409,539,937	http://creativecommons.org/licenses/by/4.0/
obsolete	Documentation	2,102,255,458	277,358,373	10	3	286,609,935	http://creativecommons.org/licenses/by/4.0/
citationmapping	Documentation	625,262,380	123,810,071	12	4	29,448,749	http://creativecommons.org/licenses/by/4.0/
taxonomy	Documentation	60,041,721	26,918	21	4	4,698,602	http://creativecommons.org/licenses/by/4.0/
citations	Documentation	31,212,544	419,769	19	5	8,870,230	http://creativecommons.org/licenses/by/4.0/
proteomes	Documentation	8,984,258	1,999,807	33	11	3,777,324	http://creativecommons.org/licenses/by/4.0/
chebi	Documentation	3,419,539	221,830	24	6	1,828,527	http://creativecommons.org/licenses/by/4.0/
rhea	Documentation	1,962,186	138,720	67	3	540,446	http://creativecommons.org/licenses/by/4.0/

Uniprot Knowledge Graph



SPARQL

Downloads

Documentation/Help



Your SPARQL query

Add common prefixes

```
1 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
3 PREFIX taxon: <http://purl.uniprot.org/taxonomy/>
4 PREFIX up: <http://purl.uniprot.org/core/>
5 SELECT ?protein ?organism ?isoform ?sequence
6 WHERE
7 {
8   ?protein a up:Protein .
9   ?protein up:organism ?organism .
10  # Taxon subclasses are materialized, do not use rdfs:subClassOf+
11  ?organism rdfs:subClassOf taxon:83333 .|
12  ?protein up:sequence ?isoform .
13  ?isoform rdf:value ?sequence .
14 }
```

Submit Query

Examples

1. Select all taxa from the [UniProt taxonomy](#) **Use**
2. Select all bacterial taxa and their scientific name from the [UniProt taxonomy](#) **Use**
3. Select all UniProtKB entries, and their organism and amino acid sequences (including isoforms), for *E. coli K12* and all its strains **Use**
4. Select the UniProtKB entry with the mnemonic 'A4_HUMAN' **Use**
5. Select a mapping of UniProtKB to PDB entries using the UniProtKB cross-references to the [PDB](#) database **Use**
6. Select all cross-references to external databases of the category '3D structure databases' of UniProtKB entries that are classified with the keyword 'Acetoin biosynthesis (KW-0005)' **Use**
7. Select reviewed UniProtKB entries (Swiss-Prot), and their recommended protein name, that have a preferred gene name that contains the text 'DNA' **Use**
8. Select the preferred gene name and disease annotation of all human UniProtKB entries that are known to be involved in a disease **Use**
9. Select all human UniProtKB entries with a sequence variant that leads to a 'loss of function' **Use**
10. Select all human UniProtKB entries with a sequence variant that leads to a tyrosine to phenylalanine substitution **Use**
11. Select all UniProtKB entries with annotated transmembrane regions and the regions' begin and end coordinates on the canonical sequence **Use**
12. Select all UniProtKB entries that were integrated on the 30th of November 2010 **Use**
13. Was any UniProtKB entry integrated on the 9th of January 2013 **Use**
14. Construct new triples of the type 'HumanProtein' from all human UniProtKB entries **Use**
15. Select the average number of cross-references to the [PDB](#) database of UniProtKB entries that have at least one cross-reference to the [PDB](#) database **Use**
16. [More examples](#)

All protein sequences associated to Escherichia coli K-12 ?

About

This SPARQL endpoint contains all UniProt data. It is free to access and supports the [SPARQL 1.1 Standard](#).

There are 217,505,202,099 triples in this release (2025_04). The query timeout is 45 minutes. All triples are available in the default graph. There are 22 named graphs.

Documentation

1. [Classes and predicates defined by the UniProt consortium](#)
2. [Tutorial on using SPARQL with UniProt](#)
3. [Statistics and diagrams](#)
4. [Example queries](#)

News



Forthcoming changes

****Table of contents**** * [Reorganizing the protein space in ...]

[UniProt release 2025_04](#)

The (RPs) provided by UniProt aim to ...

[UniProt release 2025_03](#)

Cross-references have been added to the CARD database, The Comprehensive Antibiotic Resistance Database. CARD is ...

[News archive](#)

Massive (and diverse) data already available
in the form of **knowledge graphs** ...

... how can we ensure machines/humans
speak the **same language** ?

What about **semantic interoperability** ?

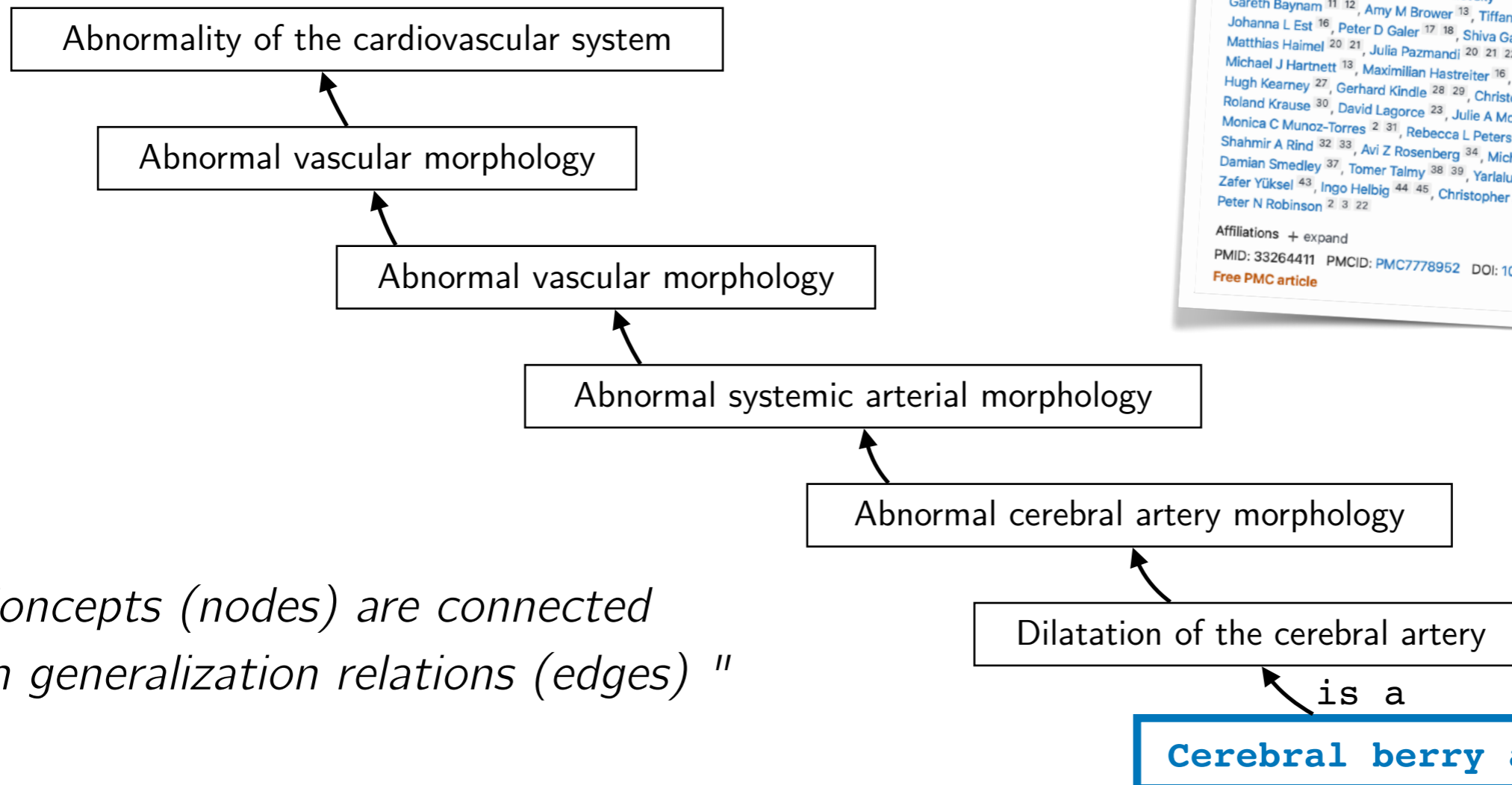
② Uniformly describe **what** is observed with data ?

Computational ontology

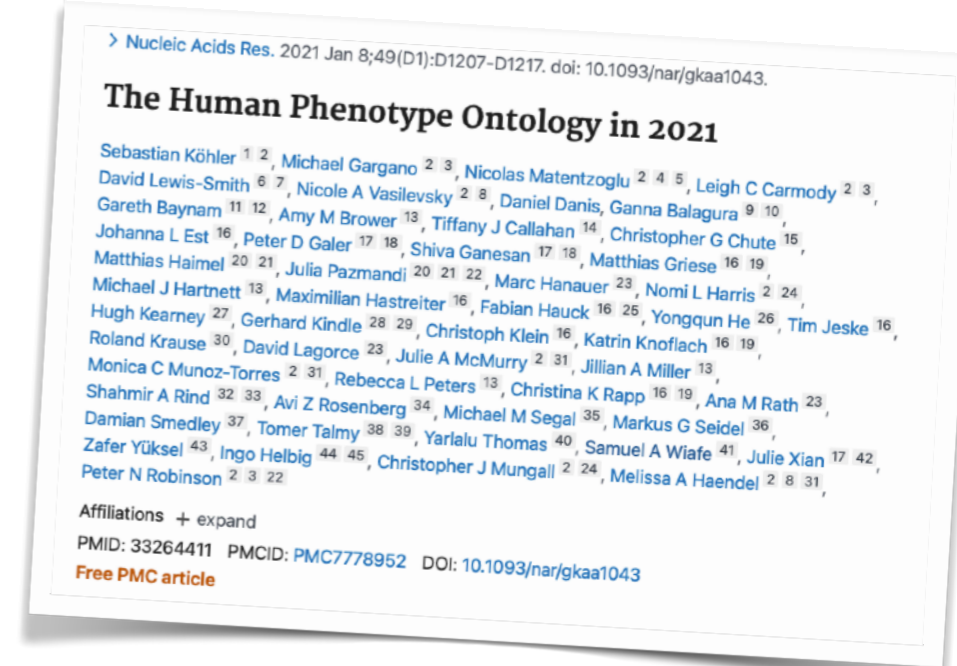
« a **formal specification** of a **shared conceptualization** » (Borst, 1997)

→ 1,049 life science ontologies registered in BioPortal (2023)

Human Phenotype Ontology



" Concepts (nodes) are connected with generalization relations (edges) "



Welcome to BioPortal, the world's most comprehensive repository of biomedical ontologies

Search for a class

Enter a class, e.g. Melanoma



[Advanced search](#)

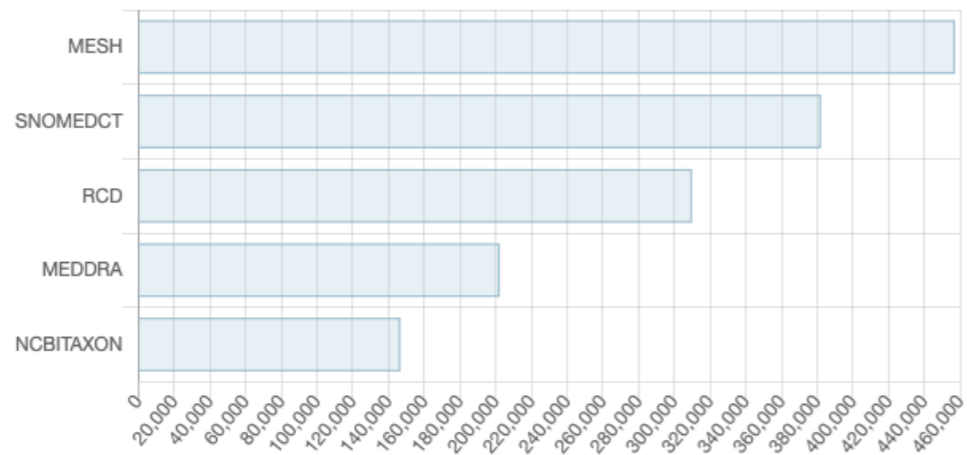
Find an ontology

Start typing ontology name, then choose from list



[Browse ontologies](#)

Ontology visits (October 2025)



[More](#)

Statistics

Ontologies	1,233
Classes	17,558,240
Properties	36,286
Mappings	92,868,606

Environment Ontology

Last uploaded: October 22, 2025



Summary **Classes** Properties Notes Mappings Widgets

Jump to

- biological_process
 - entity
 - continuant
 - generically dependent continuant
 - independent continuant
 - anatomical entity
 - immaterial entity
 - material entity
 - anthropogenic litter
 - astronomical body part
 - abyssal clay
 - acid dune sand
 - alpine tree line ecotone
 - aquatic ecosystem
 - aquatic natural environment
 - area of attached mussel assemblages
 - area of drift ice
 - area of open water
 - area of pack ice
 - area of perennial ice or snow
 - area of sea ice
 - beach sand
 - biome
 - alpine biome
 - alpine tundra biome**
 - aquatic biome

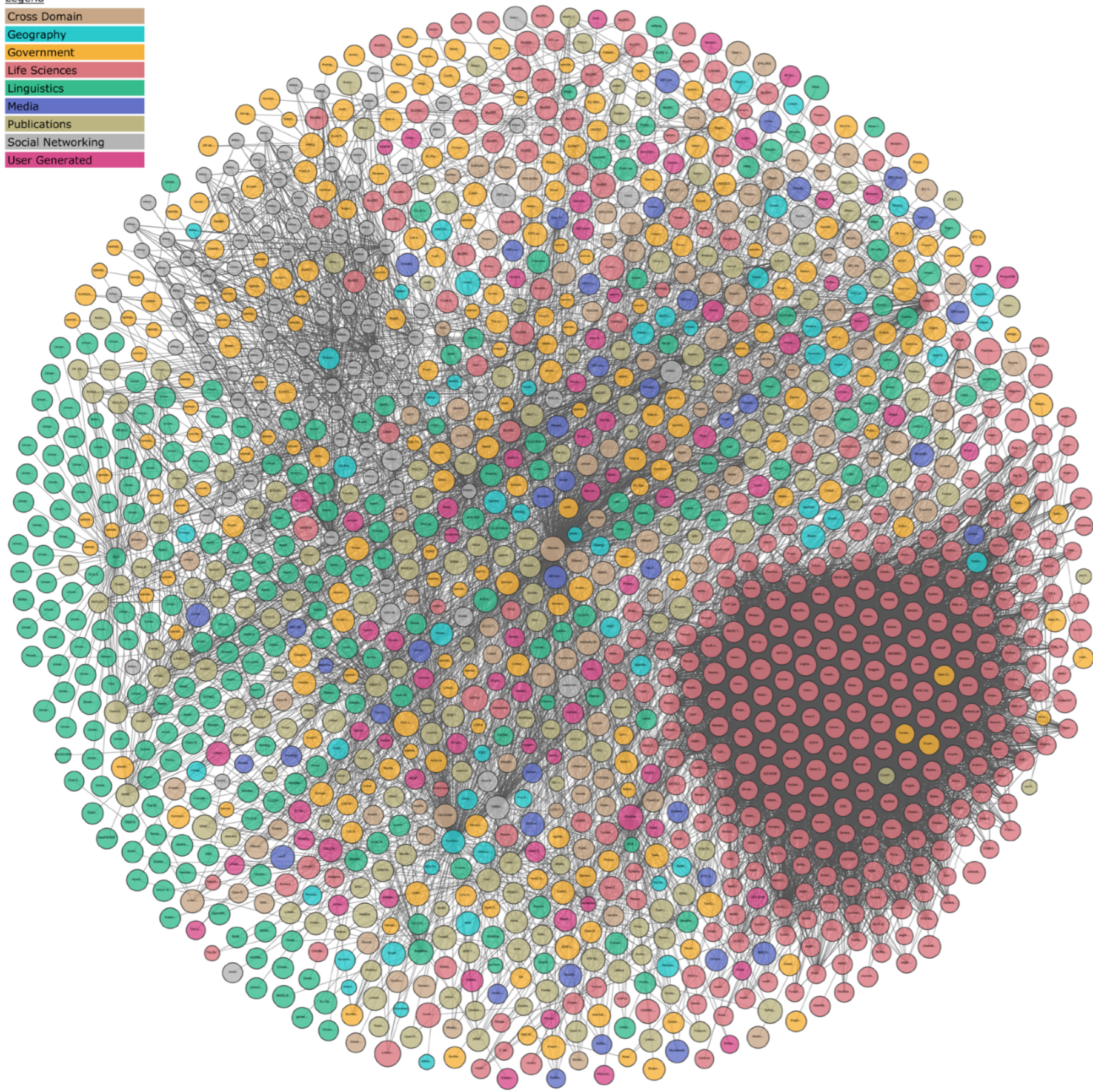
Details Visualization Notes (0) Mappings (8)



Add a proposal

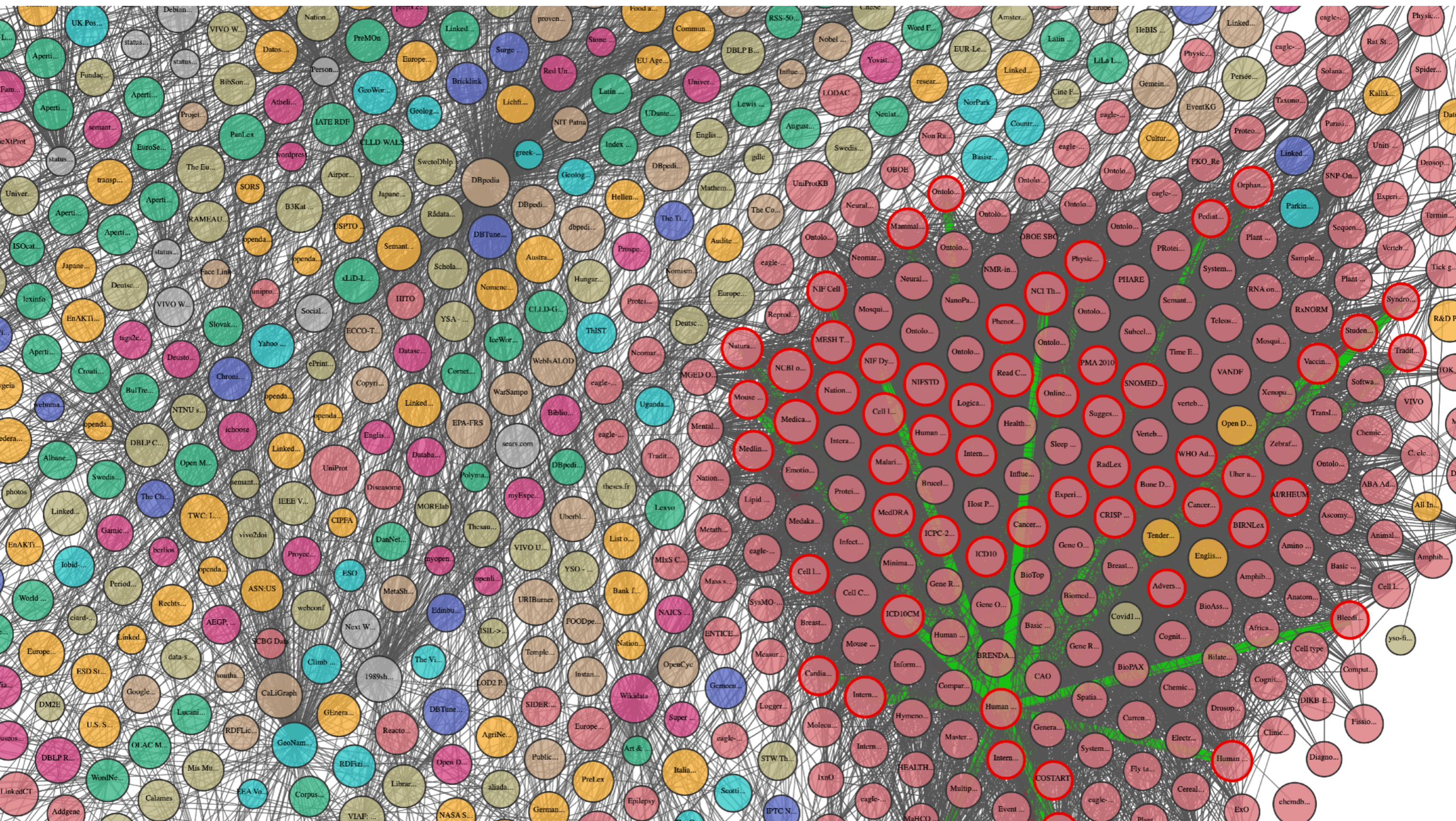
Id	http://purl.obolibrary.org/obo/ENVO_01001505
Preferred Name	alpine tundra biome
Definitions	A tundra biome which exists at high altitudes and where vegetation - dominated by a few species of dwarf shrubs, a few grasses, sedges, lichens, and mosses - is stunted due to low temperatures and high winds. The absence of trees in this biome is primarily due to high altitude rather than high latitude. On Earth, it lies roughly between the summer isotherm of 10 degrees Centigrade and the snow line. Primary productivity is low in this biome because of the extremes of climate.
Synonyms	mountain tundra
Type	http://www.w3.org/2002/07/owl#Class

All Properties	
definition	A tundra biome which exists at high altitudes and where vegetation - dominated by a few species of dwarf shrubs, a few grasses, sedges, lichens, and mosses - is stunted due to low temperatures and high winds.
label	alpine tundra biome
comment	The absence of trees in this biome is primarily due to high altitude rather than high latitude. On Earth, it lies roughly between the summer isotherm of 10 degrees Centigrade and the snow line. Primary productivity is low in this biome because of the extremes of climate.
prefLabel	alpine tundra biome
database_cross_referenc e	SPIRE:Tundra http://sweetontology.net/realmCryo/AlpineTundra
in_subset	envoPolar









Linked Open
Data Cloud
→ 1357
interlinked
RDF datasets
in 2025

Life-science semantic resources are massively interconnected



Knowledge Graphs are instrumental for FAIR principles

- ▶ By-design, built for being both **human and machine-readable**
 - Interoperability 
- ▶ Semantic Web technologies provide open and **standard protocols**: URLs / HTTP / RDF format / SPARQL query language (W3C standards)
 - Findability  (URIs for identifying things on the web)
 - Accessibility 
 - Interoperability 
- ▶ Ontologies are **community**-agreed controlled vocabularies
 - Interoperability 
 - Reuse 

Creating machine-readable
metadata with
"mini Knowledge Graphs"

RDF to link data

Definitions

- (1) An RDF **statement** expresses a **relationship** between two resources (things)
- (2) The **subject** and the **object** represent the two resources being related ; the **predicate** represents the nature of their relationship
- (3) The relationship is phrased in a **directional** way (from subject to object) and is called in RDF a **property**

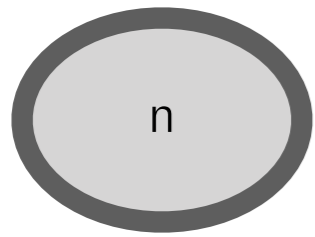
RDF triples are simple "subject | verb | object" sentences:

```
<RAC1> <is a> <human gene> .
```

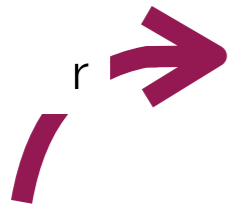
```
<RAC1> <has_label> "Rac Family Small GTPase 1" .
```

```
<seq1> <is a variant of> <RAC1> .
```


Graphical syntax



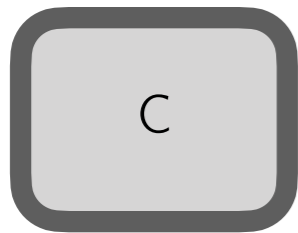
: a node in the knowledge graph



: a property/relation/edge in the knowledge graph

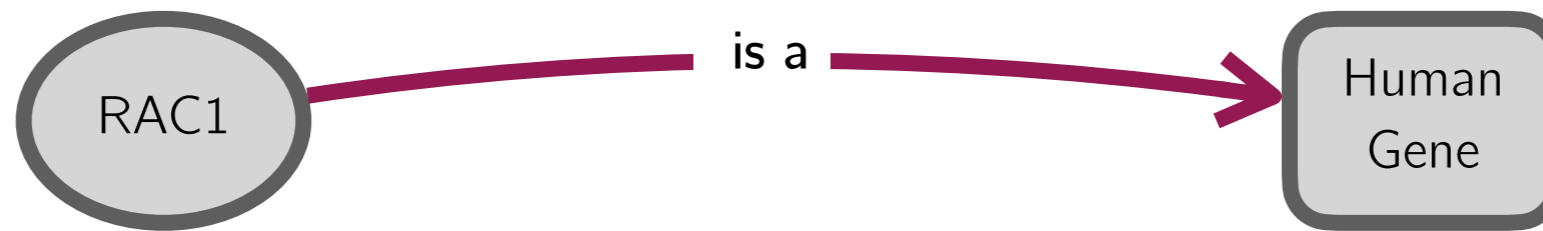
"..."

: a literal → simple textual, numerical, boolean, date value

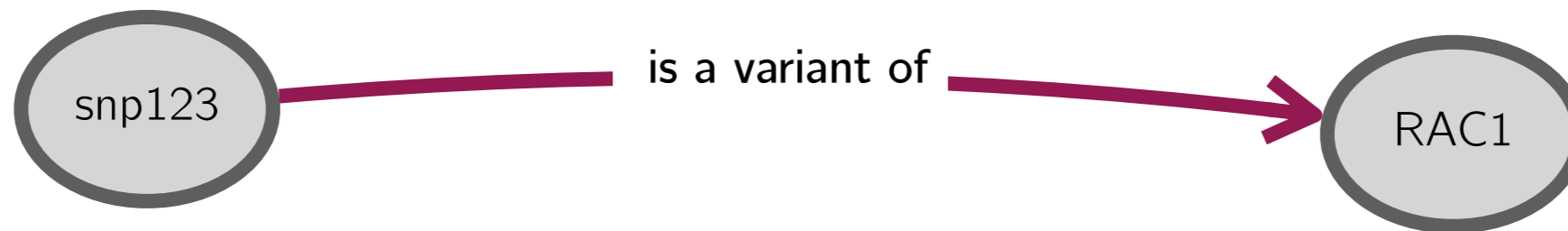


: a special node, denoting an ontology concept/class

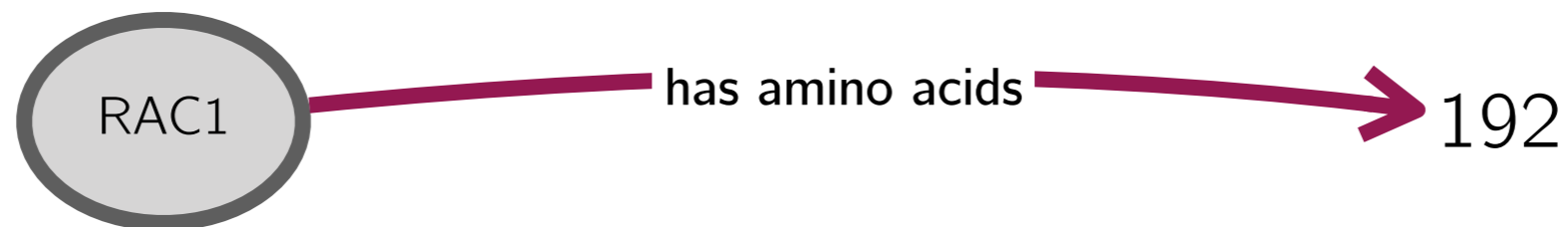
Examples



`<http://RAC1> <http://is_a> <http://Human_Gene> .`



`<http://snp123> <http://is_a_variant_of> <http://RAC1> .`

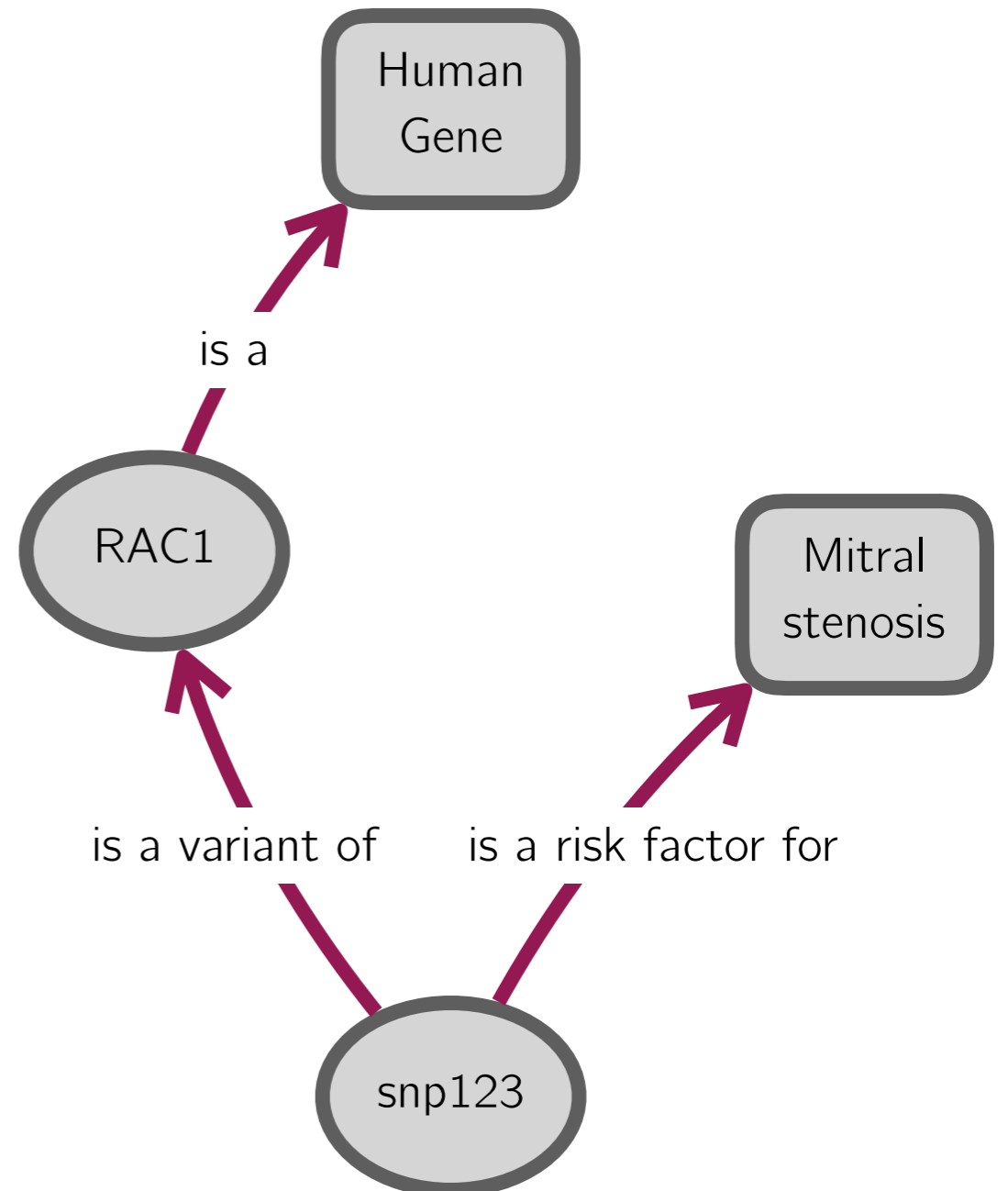


`<http://RAC1> <http://has_amino_acids> 192 .`

RDF graphs

Definitions

- (1) A **graph** structure is formed with a set of **nodes** (resources) and **edges** (relationships between resources)
- (2) A set of RDF triples is called an RDF graph. RDF is a **directed, labeled graph** data format for representing information/ knowledge on the Web.



Writing RDF graphs with the Turtle syntax

Definitions

(1) One line per triple, each element separated by **space**, each triple ends with a **.**

S P O .

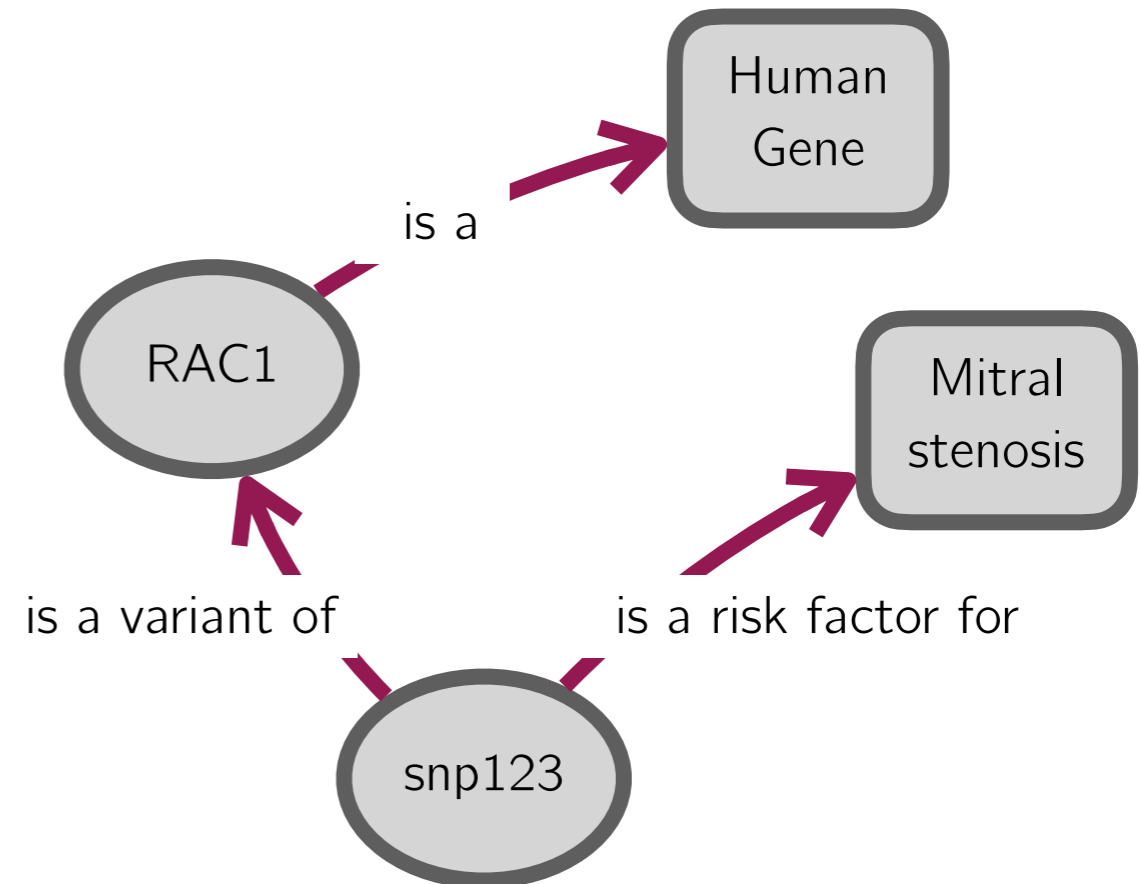
(2) If two triples describe the same subject, you can reuse it:

S P₁ O₁ ;

P₂ O₂ .

(3) If two triples describe the same subject and predicate, you can reuse it:

S P O₁ , O₂ .



```
@prefix ns: <http://my/namespace/> .
```

```
ns:RAC1    rdf:type ns:Human_gene .
```

```
ns:snp123 ns:is_a_variant_of ns:RAC1 ;
```

```
ns:is_a_risk_factor_for ns:Mitral_stenosis .
```

Hands-on session: from text to KG

Question #1

From wikipedia : “*The insulin receptor (IR) is a [transmembrane receptor](#) that is activated by [insulin](#), [IGF-I](#), [IGF-II](#) and belongs to the large class of [receptor tyrosine kinase](#).*”

Draft a **graphical representation** of the associated knowledge graph.

- ✓ Identify verbs → RDF predicates
- ✓ Identify linked entities,
 - who is a subject of a relation ?
 - who is the object of a relation ?

"Pen & paper" team work

The screenshot shows a collaborative workspace with a top toolbar containing icons for lock, hand, mouse cursor, square, diamond, circle, arrow, minus, eraser, text, image, and link. A text instruction reads: "Pour déplacer le canevas, maintenez **Clic molette** ou **Espace** enfoncé tout en faisant glisser, ou utilisez l'outil main".

On the left, a task is given: "Traduire le texte suivant sous la forme de graphe de connaissances : 'The insulin receptor (IR) is a transmembrane receptor that is activated by insulin, IGF-I, IGF-II and belongs to the large class of receptor tyrosine kinase.' Vous vous appuierez sur la syntaxe graphique suivante :".

The syntax diagram shows a box labeled "classe 'parent'" with an arrow labeled "is a" pointing to a box labeled "classe 'enfant'". It also shows a circle labeled "sujet" with an arrow labeled "predicat" pointing to a circle labeled "objet".

Four workspace panels are visible, labeled "Groupe #1", "Groupe #2", "Groupe #3", and "Groupe #4".

- Groupe #1** contains a legend with:
 - A box labeled "Classe/Concept"
 - A circle labeled "objet"
 - The word "Literal"
 - An arrow labeled "relation X"
 - An arrow labeled "is a"
- Groupe #3** contains a legend with:
 - A box labeled "Classe/Concept"
 - A circle labeled "objet"
 - The word "Literal"
 - An arrow labeled "relation X"
 - An arrow labeled "is a"

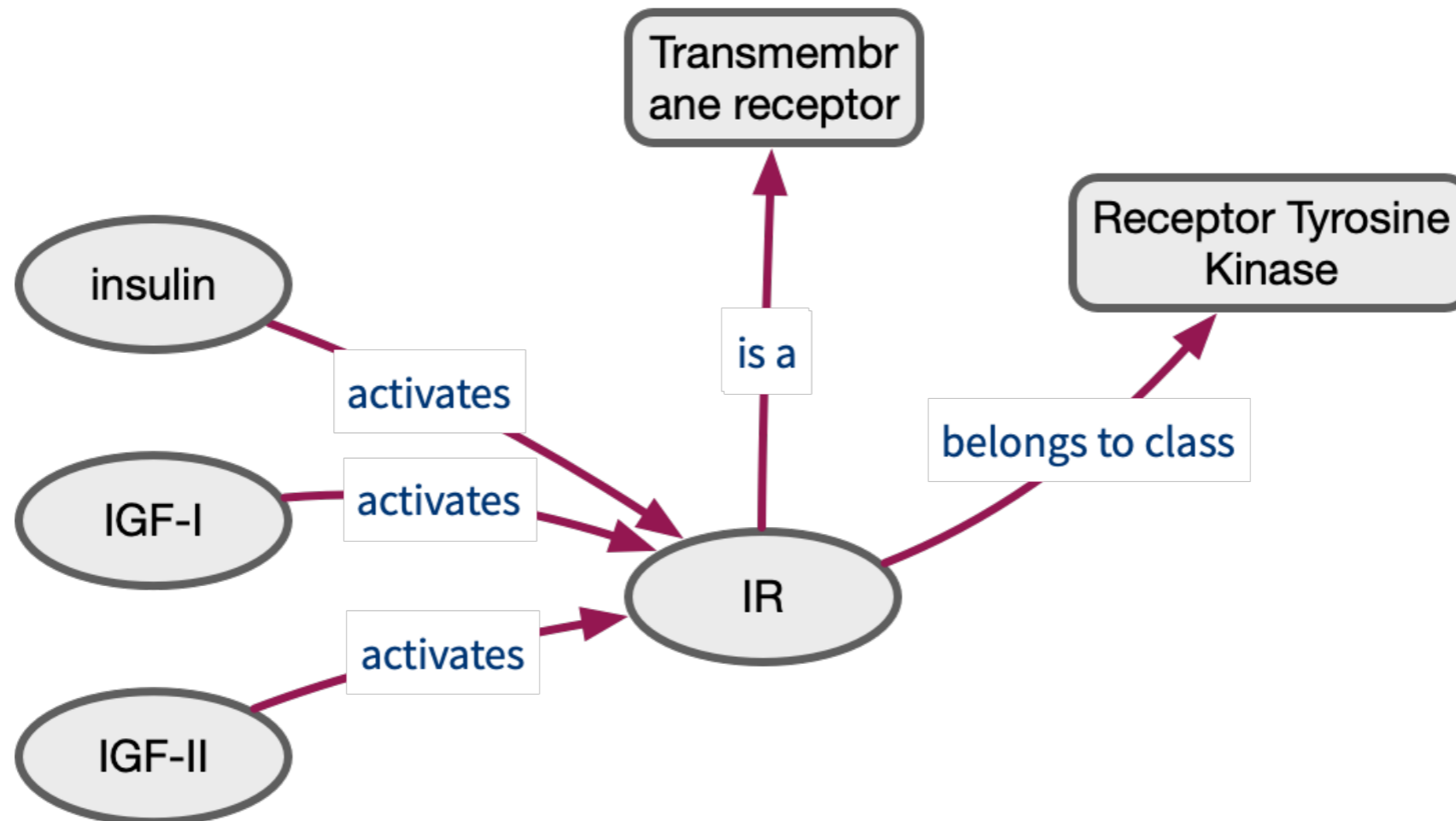
Groupe #2 and Groupe #4 are currently empty.

<https://tinyurl.com/4xf83nxd>



One of the many solutions

“The insulin receptor (IR) is a transmembrane receptor that is activated by insulin, IGF-I, IGF-II and belongs to the large class of receptor tyrosine kinase.”



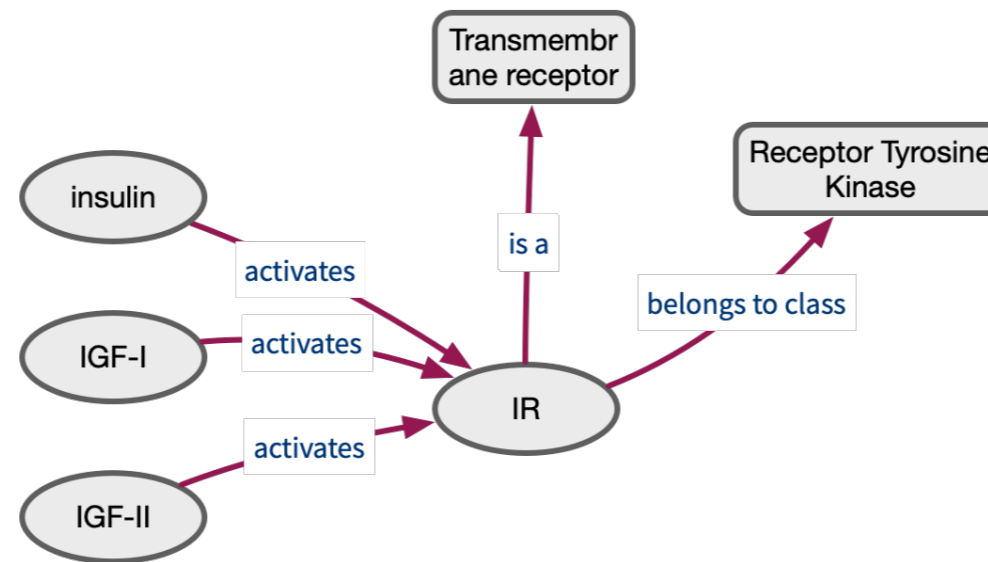
Hands-on session: from text to KG

Question #2

From wikipedia : “*The insulin receptor (IR) is a [transmembrane receptor](#) that is activated by [insulin](#), [IGF-I](#), [IGF-II](#) and belongs to the large class of [receptor tyrosine kinase](#).*”

Translate your KG into **RDF triples**.

In practice ...



```
@prefix ns: <http://my/namespace/> .
```

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
```

```
ns:insulin ns:activates ns:IR .
```

```
ns:IGF_I ns:activates ns:IR .
```

```
ns:IGF_II ns:activates ns:IR .
```

```
ns:IR rdf:type ns:TransmembraneReceptor ;
```

```
ns:belongs_to_class ns:ReceptorTyrosineKinase .
```

Hands-on session: from text to KG

<https://rest.uniprot.org/uniprotkb/P06213.ttl>

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix up: <http://purl.uniprot.org/core/> .
@prefix annotation: <http://purl.uniprot.org/annotation/> .
@prefix citation: <http://purl.uniprot.org/citations/> .
@prefix range: <http://purl.uniprot.org/range/> .
@prefix faldo: <http://biohackathon.org/resource/faldo#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix pubmed: <http://purl.uniprot.org/pubmed/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix position: <http://purl.uniprot.org/position/> .
```

```
<P06213> rdf:type up:Protein ;
up:citation citation:2859121 ,
              citation:2983222 ;
up:annotation annotation:PRO_0000016687 ,
                annotation:PRO_0000016689 ,
                annotation:VAR_015924 .
```

```
citation:2859121 rdf:type up:Journal_Citation ;
up:title "The human insulin receptor cDNA: the structural basis for
hormone-activated transmembrane signalling." ;
up:author "Ebina Y." , "Ellis L." ;
skos:exactMatch pubmed:2859121 .
```

```
annotation:PRO_0000016687 rdf:type up:Chain_Annotation ;
rdfs:comment "Insulin receptor subunit alpha" ;
up:mass 83642 ;
up:range range:22571007465304878tt28tt758 .
```

```
range:22571007465304878tt28tt758 rdf:type faldo:Region ;
faldo:begin position:22571007465304878tt28 ;
faldo:end position:22571007465304878tt758 .
```

Question #3

Draft the knowledge graph associated to some of the RDF triples representing the P06213 Uniprot entity.

"Pen & paper" team work

Traduire les triplets RDF suivants sous la forme de graphe de connaissances en utilisant la syntaxe graphique :

```
graph LR; C1[Classe] -- is a --> C2[Classe]; S((sujet)) -- predicat --> O((objet))
```

```
<P06213> rdf:type up:Protein ;  
up:citation citation:2859121 ,  
citation:2983222 ;  
up:annotation annotation:PRO_0000016687 ,  
annotation:PRO_0000016689 ,  
annotation:VAR_015924 .  
  
citation:2859121 rdf:type up:Journal_Citation ;  
up:title "The human insulin receptor cDNA: the structural basis for  
hormone-activated transmembrane signalling." ;  
up:author "Ebina Y." , "Ellis L." ;  
skos:exactMatch pubmed:2859121 .  
  
annotation:PRO_0000016687 rdf:type up:Chain_Annotation ;  
rdfs:comment "Insulin receptor subunit alpha" ;  
up:mass 83642 ;  
up:range range:22571007465304878tt28tt758 .  
  
range:22571007465304878tt28tt758 rdf:type faldo:Region ;  
faldo:begin position:22571007465304878tt28 ;  
faldo:end position:22571007465304878tt758 .
```

Legend for graph symbols:

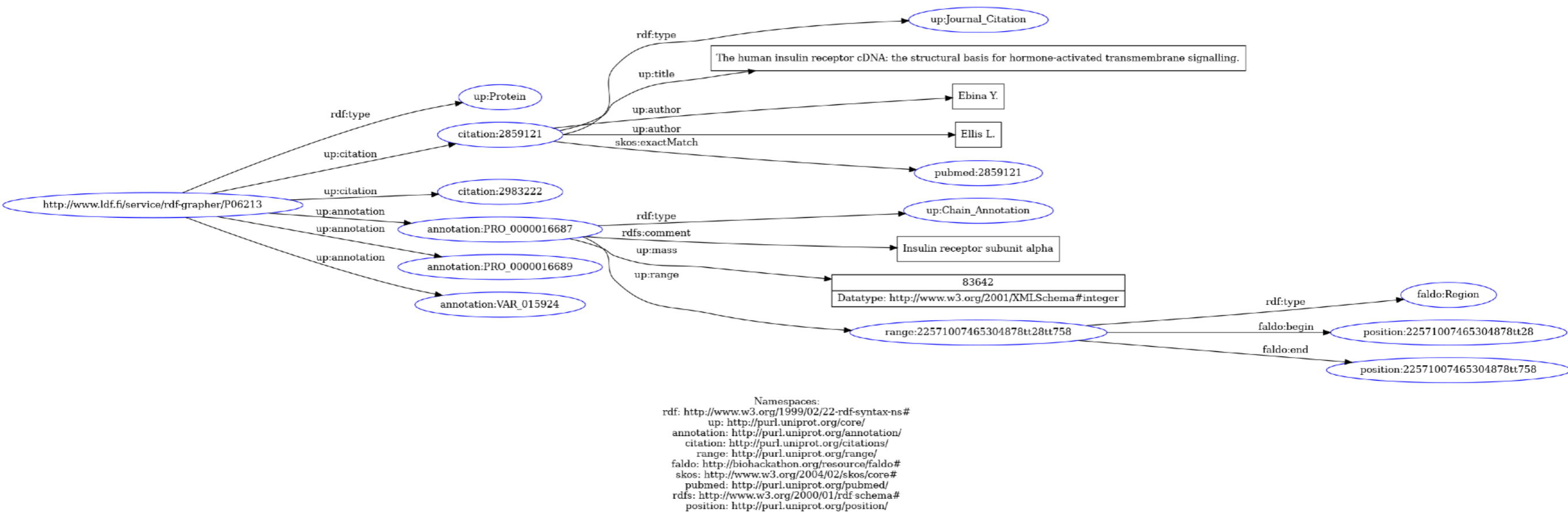
- Classe/Concept (rectangle)
- objet (circle)
- Literal (text)
- relation X (arrow)
- is a (arrow)

<https://tinyurl.com/4xf83nxd>



Practice ... from KG to text

<https://www.ldf.fi/service/rdf-grapher>



Ontologies for FAIR bioinformatics tools and workflows

schema.org



Full Hierarchy

Schema.org is defined as two hierarchies: one for textual property values, and one for the things that they describe.

This is the main schema.org hierarchy: a collection of types (or "classes"), each of which has one or more parent types. Although a type may have more than one super-type, here we show each type in one branch of the tree only. There is also a parallel hierarchy for **data types**.

Types:

Close hierarchy / Open hierarchy

Thing

- ▶ Action +
- ▶ BioChemEntity +
- ▶ CreativeWork +
- ▶ Event +
- ▶ Intangible +
- ▶ MedicalEntity +
- ▶ Organization +
- ▶ Person +
- ▶ Place +
- Product
 - DietarySupplement
 - Drug
 - IndividualProduct
 - ProductCollection
 - ProductGroup

- ▶ General purpose **lightweight** ontology
- ▶ Aimed at **annotating web pages**
- ▶ Targetting **FINDABILITY**
- ▶ Originating from major search engines



Dublin Core (aka DC-Terms)



```
dcterms:Box
  dcterms:issued "2000-07-11"^^<http://www.w3.org/2001/XMLSchema#date> ;
  a rdfs:Datatype ;
  rdfs:comment "The set of regions in space defined by their geographic coordinates according to the DCMI Box Encoding Scheme."@en ;
  rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
  rdfs:label "DCMI Box"@en ;
  rdfs:seeAlso <https://www.dublincore.org/specifications/dublin-core/dcmi-box/> .

dcterms:DCMIType
  dcterms:issued "2000-07-11"^^<http://www.w3.org/2001/XMLSchema#date> ;
  a dcam:VocabularyEncodingScheme ;
  rdfs:comment "The set of classes specified by the DCMI Type Vocabulary, used to categorize the nature or genre of the resource."@en ;
  rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
  rdfs:label "DCMI Type Vocabulary"@en ;
  rdfs:seeAlso <http://purl.org/dc/dcmitype/> .

dcterms:DDC
  dcterms:issued "2000-07-11"^^<http://www.w3.org/2001/XMLSchema#date> ;
  a dcam:VocabularyEncodingScheme ;
  rdfs:comment "The set of conceptual resources specified by the Dewey Decimal Classification."@en ;
  rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
  rdfs:label "DDC"@en ;
  rdfs:seeAlso <http://www.oclc.org/dewey/> .

dcterms:FileFormat
  dcterms:issued "2008-01-14"^^<http://www.w3.org/2001/XMLSchema#date> ;
  a rdfs:Class ;
  rdfs:comment "A digital resource format."@en ;
  rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
  rdfs:label "File Format"@en ;
  rdfs:subClassOf dcterms:MediaType .

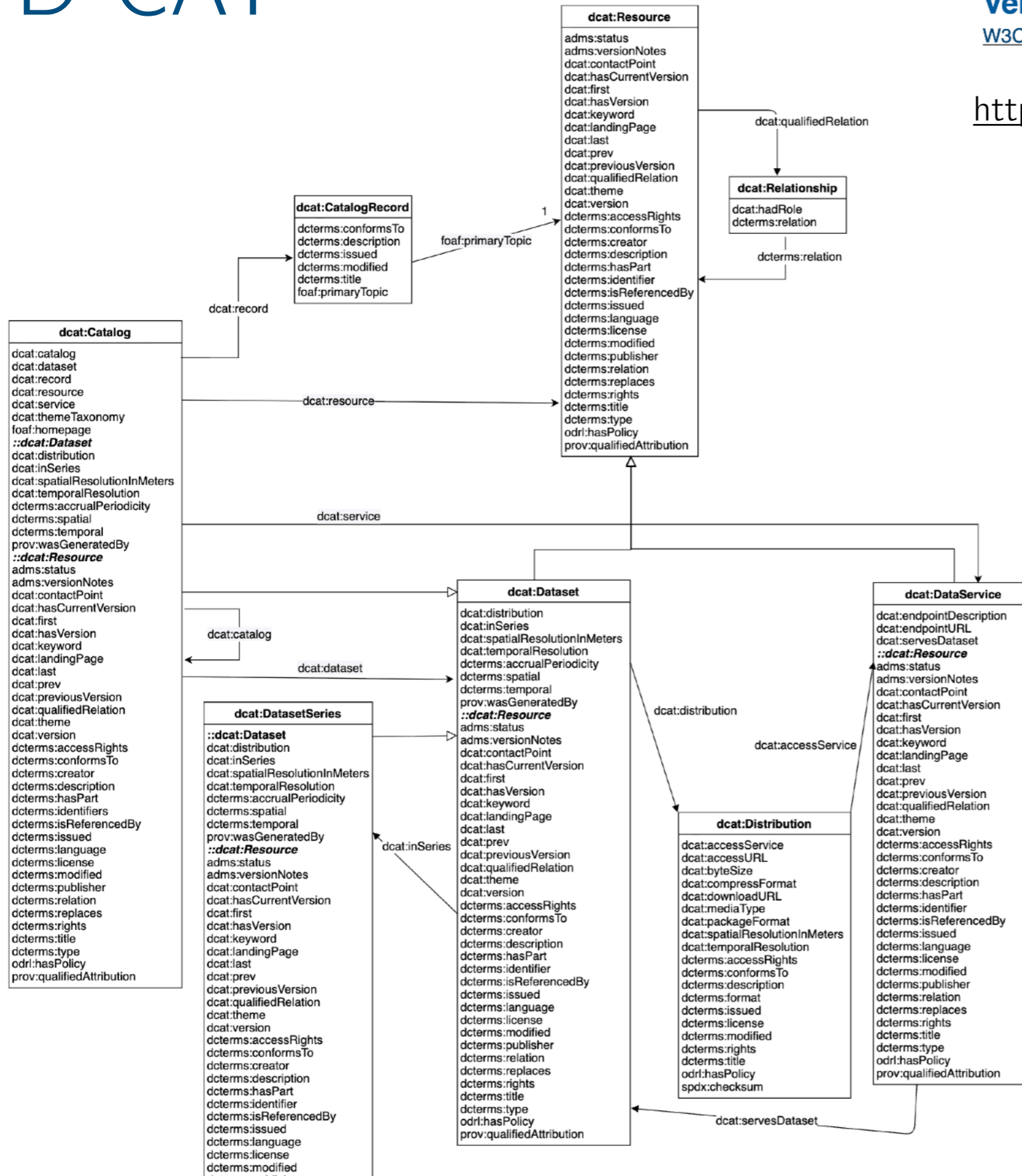
dcterms:Frequency
  dcterms:issued "2008-01-14"^^<http://www.w3.org/2001/XMLSchema#date> ;
  a rdfs:Class ;
  rdfs:comment "A rate at which something recurs."@en ;
  rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
  rdfs:label "Frequency"@en .

dcterms:IMT
  dcterms:issued "2000-07-11"^^<http://www.w3.org/2001/XMLSchema#date> ;
  a dcam:VocabularyEncodingScheme ;
  rdfs:comment "The set of media types specified by the Internet Assigned Numbers Authority."@en ;
  rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
  rdfs:label "IMT"@en ;
  rdfs:seeAlso <http://www.iana.org/assignments/media-types/> .

dcterms:ISO3166
  dcterms:issued "2000-07-11"^^<http://www.w3.org/2001/XMLSchema#date> ;
  a rdfs:Datatype ;
  rdfs:comment "The set of codes listed in ISO 3166-1 for the representation of names of countries."@en ;
  rdfs:isDefinedBy <http://purl.org/dc/terms/> ;
  rdfs:label "ISO 3166"@en ;
  rdfs:seeAlso <https://www.iso.org/obp/ui/#search> .
```

- ▶ **lightweight** ontology
- ▶ Generic metadata
- ▶ 22 classes
- ▶ 55 properties

<https://www.w3.org/TR/vocab-dcat-3/>



European Union

EU Vocabularies

Publications Office | EU law | European data | EU tenders | EU research results | EU Whoiswh

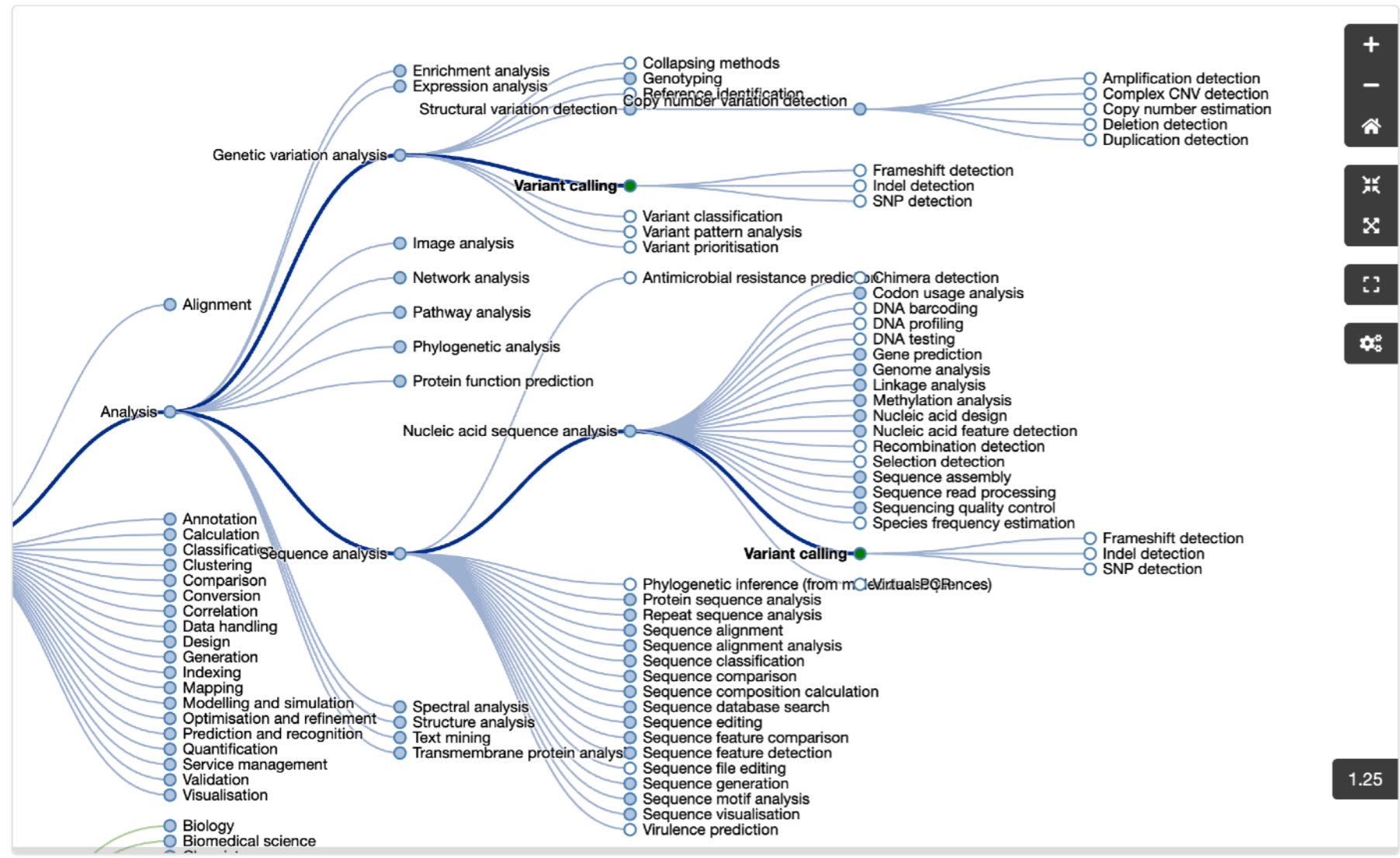
Publications Office > EU Vocabularies > Business collections > DCAT-AP for data portals in Europe

Home | Controlled vocabularies | Models | Business collections | Online tools

DCAT Application profile for data portals in Europe (DCAT-AP)

FAIR Data Point

Working Draft, 20 October 2023



Details of term "Variant calling"

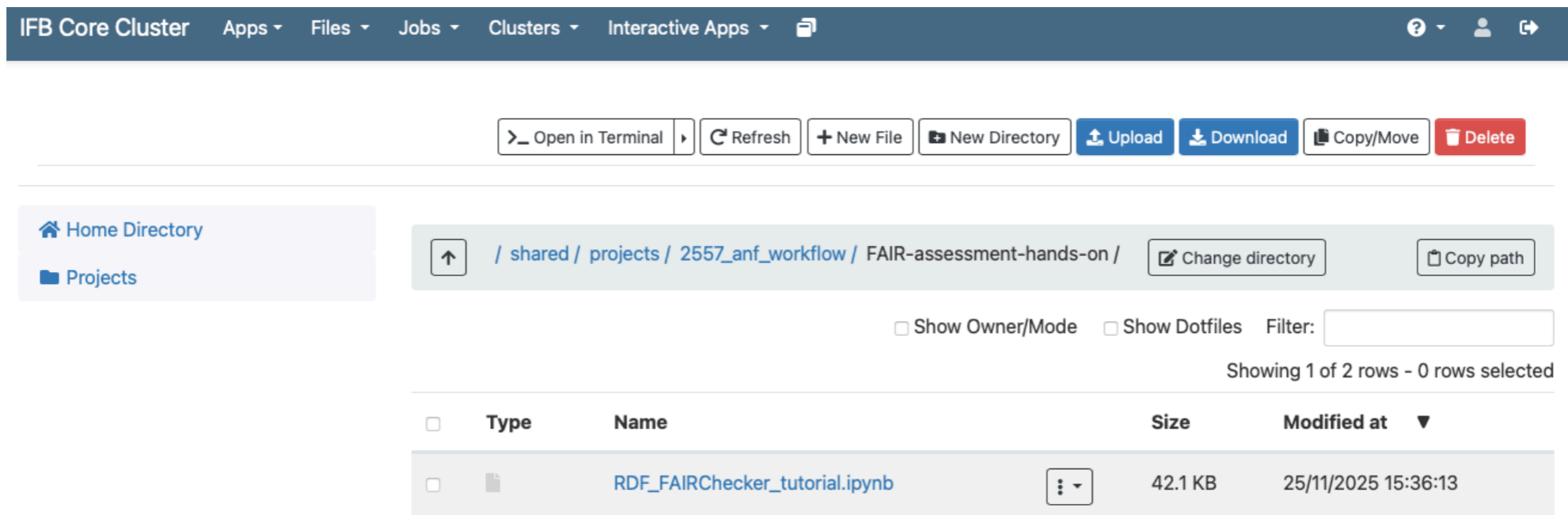
Term	Variant calling
Definition	Detect, identify and map mutations, such as single nucleotide polymorphisms, short indels and structural variants, in multiple DNA sequences. Typically the alignment and comparison of the fluorescent traces produced by DNA sequencing hardware, to study genomic alterations.
Comment	<p>Methods often utilise a database of aligned reads.</p> <p>Somatic variant calling is the detection of variations established in somatic cells and hence not inherited as a germ line variant.</p> <p>Variant detection</p>
Exact synonyms	Variant mapping
Narrow synonyms	<p>Allele calling</p> <p>Exome variant detection</p> <p>Genome variant detection</p> <p>Germ line variant calling</p> <p>Mutation detection</p> <p>Somatic variant calling</p> <p>de novo mutation detection</p>
URI	operation_3227
Parents	<p>Nucleic acid sequence analysis</p> <p>Genetic variation analysis</p>

- ▶ A domain-specific ontology for Bioinformatics for annotating
 - **usage context** of bioinformatics tools (*EDAM Topics*)
 - **what does** the tool (*EDAM Operations*)
 - **input and output data** (*EDAM Data, EDAM Formats*)

Hands-on session: RDF graphs in Python

Question #4

Run step-by-step the PART1 section of the Jupyter notebook available on the IFB cluster :



The screenshot shows the IFB Core Cluster file manager interface. The top navigation bar includes "IFB Core Cluster", "Apps", "Files", "Jobs", "Clusters", and "Interactive Apps". Below this is a toolbar with buttons for "Open in Terminal", "Refresh", "New File", "New Directory", "Upload", "Download", "Copy/Move", and "Delete". The main area shows the current directory path: "/ shared / projects / 2557_anf_workflow / FAIR-assessment-hands-on /". A sidebar on the left shows "Home Directory" and "Projects". Below the path bar, there are checkboxes for "Show Owner/Mode" and "Show Dotfiles", and a "Filter:" input field. The status bar indicates "Showing 1 of 2 rows - 0 rows selected". A table lists the files in the directory:





Type	Name	Size	Modified at
<input type="checkbox"/>	RDF_FAIRChecker_tutorial.ipynb	42.1 KB	25/11/2025 15:36:13

Recap'





A practical overview on FAIR assessment, in 2 hours ...

- ▶  **Quick intro on FAIR principles**

- ▶ **Knowledge graphs and semantic metadata**

-  What are Knowledge Graphs, Computational Ontologies ?
-  How to create Knowledge Graphs ? (team work, "pen & paper")
-  How to write machine-readable semantic metadata ? (demo, python code)
-  How to query knowledge graphs with SPARQL (demo, python code)



- ▶ **Tools and resources for increased FAIRness**

-  FAIR-Checker
-  Tools and workflow registries
-  Using FAIR-Checker and interpreting the FAIR assessment results
-  Using the FAIR-Checker API and processing multiple resources





A practical overview on FAIR assessment, in 2 hours ...

- ▶  Quick intro on FAIR principles

- ▶ Knowledge graphs and semantic metadata

-  What are Knowledge Graphs, Computational Ontologies ?
-  How to create Knowledge Graphs ? (team work, "pen & paper")
-  How to write machine-readable semantic metadata ? (demo, python code)
-  How to query knowledge graphs with SPARQL (demo, python code)

- ▶ Tools and resources for increased FAIRness

-  FAIR-Checker
-  Tools and workflow registries
-  Using FAIR-Checker and interpreting the FAIR assessment results
-  Using the FAIR-Checker API and processing multiple resources

Coffee break

