# Knowledge graphs for life science data integration

Olivier Dameron, Alban Gaignard, Pierre Larmande

April 01, 2026

ETBII 2026



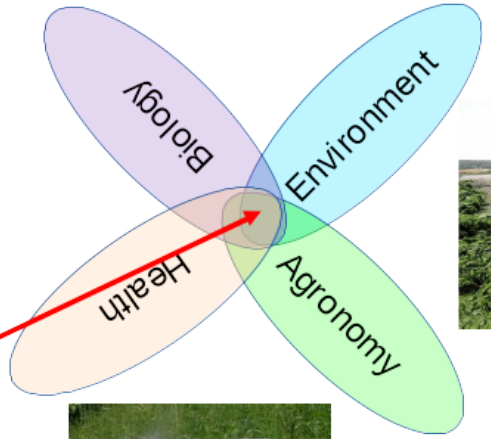





# Life science data

Life science : a large, complex, inter-dependent domain...

Data!






... that stands out among other experimental sciences

"Biology has become an information science" [T. Lenoir, 1998 Stanford]

# What to expect for 2025 ?

Our estimation is that genomics is a "four-headed beast" – it is either on par with or the most demanding domain [...] in terms of

- data acquisition
- data storage
- data distribution
- data analysis

## Big Data: Astronomical or Genomical?

**Zachary D. Stephens[1], Skylar Y. Lee[1], Faraz Faghri[2], Roy H. Campbell[2], Chengxiang Zhai[3], Miles J. Efron[4], Ravishankar Iyer[1], Michael C. Schatz[5]*, Saurabh Sinha[3]*, Gene E. Robinson[6]***

PLOS Biology | DOI:10.1371/journal.pbio.1002195 July 7, 2015

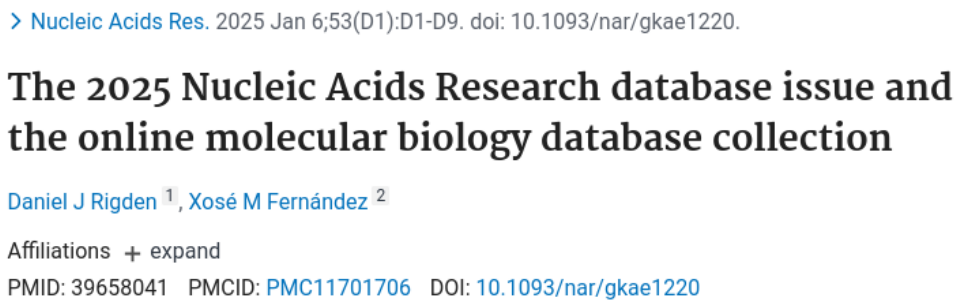
**Table 1. Four domains of Big Data in 2025.** In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

| Data Phase | Astronomy | Twitter | YouTube | Genomics |
|---|---|---|---|---|
| **Acquisition** | 25 zetta-bytes/year | 0.5–15 billion tweets/year | 500–900 million hours/year | 1 zetta-bases/year |
| **Storage** | 1 EB/year | 1–17 PB/year | 1–2 EB/year | 2–40 EB/year |
| **Analysis** | In situ data reduction | Topic and sentiment mining | Limited requirements | Heterogeneous data and analysis |
| | Real-time processing | Metadata analysis | | Variant calling, ~2 trillion central processing unit (CPU) hours |
| | Massive volumes | | | All-pairs genome alignments, ~10,000 trillion CPU hours |
| **Distribution** | Dedicated lines from antennae to server (600 TB/s) | Small units of distribution | Major component of modern user's bandwidth (10 MB/s) | Many small (10 MB/s) and fewer massive (10 TB/s) data movement |

doi:10.1371/journal.pbio.1002195.t001

# Life science : beyond "data science" ...

- we have accumulated a trove of data
- we store and share these data
  (in 2.000+ reference databases [Rigden2025])



> Nucleic Acids Res. 2025 Jan 6;53(D1):D1-D9. doi: 10.1093/nar/gkae1220.

**The 2025 Nucleic Acids Research database issue and the online molecular biology database collection**

Daniel J Rigden [1], Xosé M Fernández [2]

Affiliations + expand

PMID: 39658041 PMCID: PMC11701706 DOI: 10.1093/nar/gkae1220

## Life science : a domain that stands out

- by its complexity
- by the scarcity of its unifying laws
- by its long history of knowledge description and formalization

## ... and toward "knowledge(-based) science" ?

We needs a framework to support this transition

# Attempt at defining some underlying notions

## Database (e.g. Gene Expression Omnibus)

structured description of sets of homogeneous (as in "of the same class") instances and the relations between them.

## Ontology (e.g. GeneOntology, ChEBI, HPO)

formal description of the general concepts (as in "the classes of things") of a domain and the relations between them.
(support general inferences).

## Knowledge base (e.g. UniProt, Reactome, Rhea)

combines elements of databases and ontologies
(supports domain-based inferences about instances).

## Knowledge graph (e.g. Reactome)

knowledge base that emphasizes graph-based capabilities.

# Requirements for coping with life science data complexity

- **Requirement 1 :** **identify** resources with interoperable identifiers
- **Requirement 2 :** **describe** resources
  - their characteristics (e.g. start and end position of a gene,...)
  - their relations to other entities (e.g. the transcripts associated to a gene, the transcription factors that regulate it,...)
  - the categories they belong to
- **Requirement 3 :** **combine** descriptions from different origins, different points of view, different granularity levels
- **Requirement 4 :** **query** these descriptions
- **Requirement 5 :** support **semantically-rich** querying and reasoning (because of the inner complexity) using domain knowledge (this is required for capturing *expertise*)
- **Requirement 6 :** cover the whole **data life cycle**
- **Requirement 7 :** enforce **reproducibility**

If only the solutions to all these requirements were compatible !

# Outline

- Life science data require
  - integration
  - knowledge-based reasoning
- the Semantic Web provides a relavant framework
  - RDF to represent knowledge graphs
  - RDFS and OWL to represent knowledge and perform reasoning
  - SPARQL to query knowledge graphs

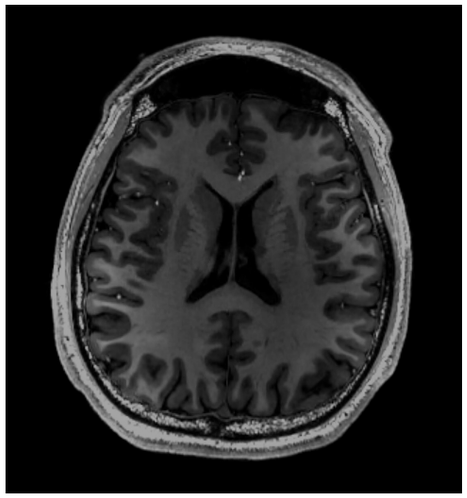
## What you will learn (hopefully)

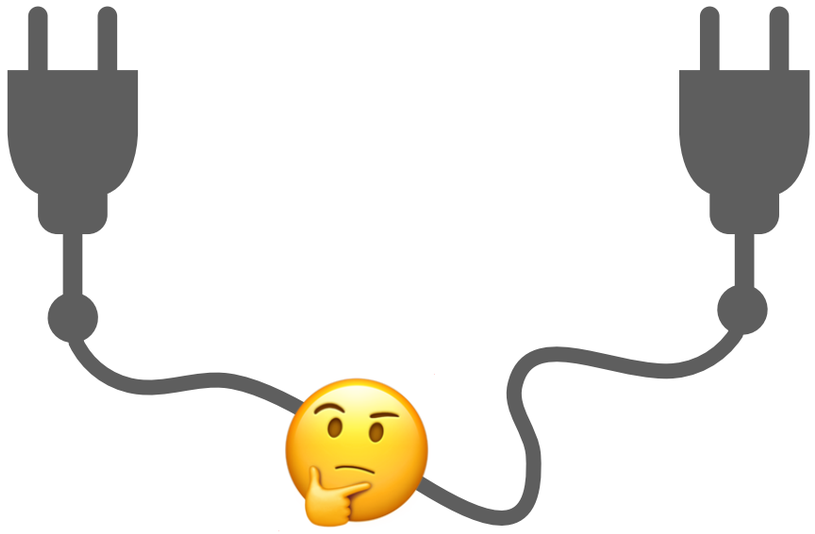
- a general understanding of symbolic knowledge...
- ... that relies on surprisingly simple principles

# Introduction to Knowledge Graphs

# Data integration needs **interoperability**

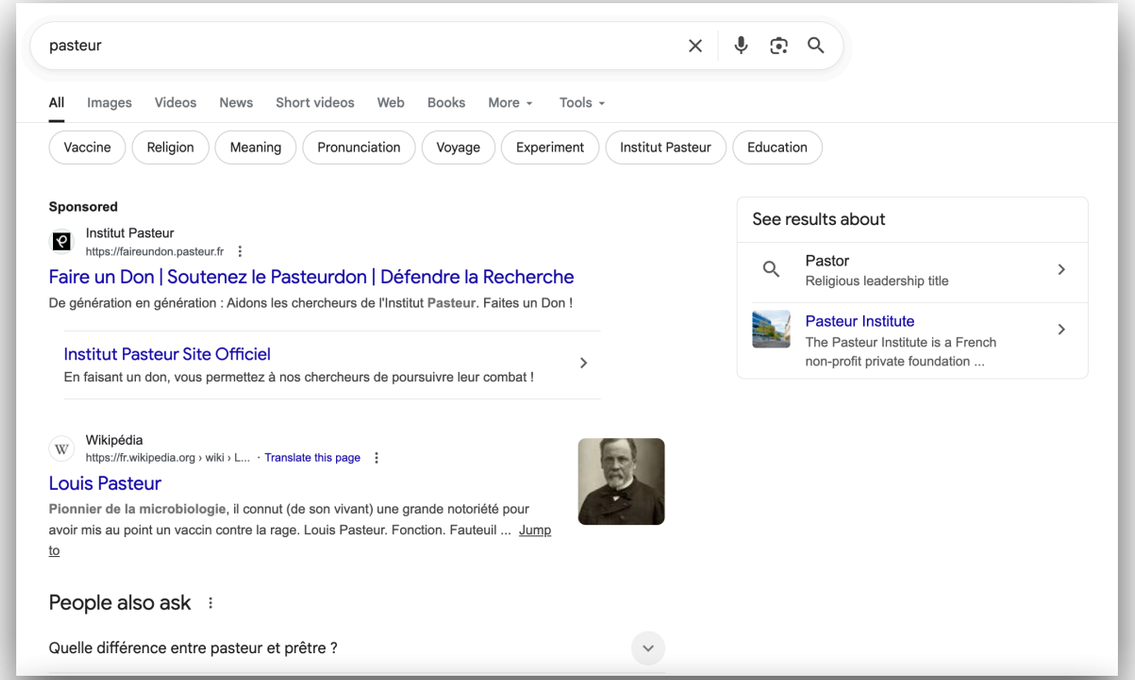
```
@HWI-ST534_129:2:24:20503:16510:CGATGT
CTGAGAGCCGGGGAAGCCGGCGGAGCCGGGGGACTGGCGAGCCGGAACAT
+
HHHHHHHHHHEFDDGDDFBFGG>7D4<9;<&?:;<DC>CCDD@?=?A###
@HWI-ST534_129:2:42:2118:9580:CGATGT
GGCGGAGCCGGGTGACTGGCGAGCCGGAACATCAGGCGCCGCCGCAGAGA
+
GEECGGGBGIDF6FFFFEF=IDEFBEE8E8E?EEB@6=9B##########
@HWI-ST534_129:2:2:12654:80229:CGATGT
CGGAGCCGGGTGACTGGCGAGCCGGAACATCAGGCGCCGCCGCAGAGAAG
+
GGEGFCDCBBAEEEEGGFGFG;EGEEGFFBBEDEBDFGFCFF;DF2D<DD
@HWI-ST534_129:2:48:12356:179714:CGATGT
GAGCCGGGTGACTGGCGAGCCGGAACATCAGGCGCCGCCGCAGAGAAGAA
+
E=GHFHEGHHBCGDDBEEBBCBDDDE@EGBD=ABDCB?EC;@@8@EEB;E
@HWI-ST534_129:2:44:8225:39540:CGATGT
GGGTGACTGGCGAGCCGGAACATCAGGCGCCGCCGCAGAGAAGAACTATG
+
HHHEHHHHHHHHHGHHHHFHHHHHHHHDFDHHBHFDDFEFEFF>G<CCCE
```

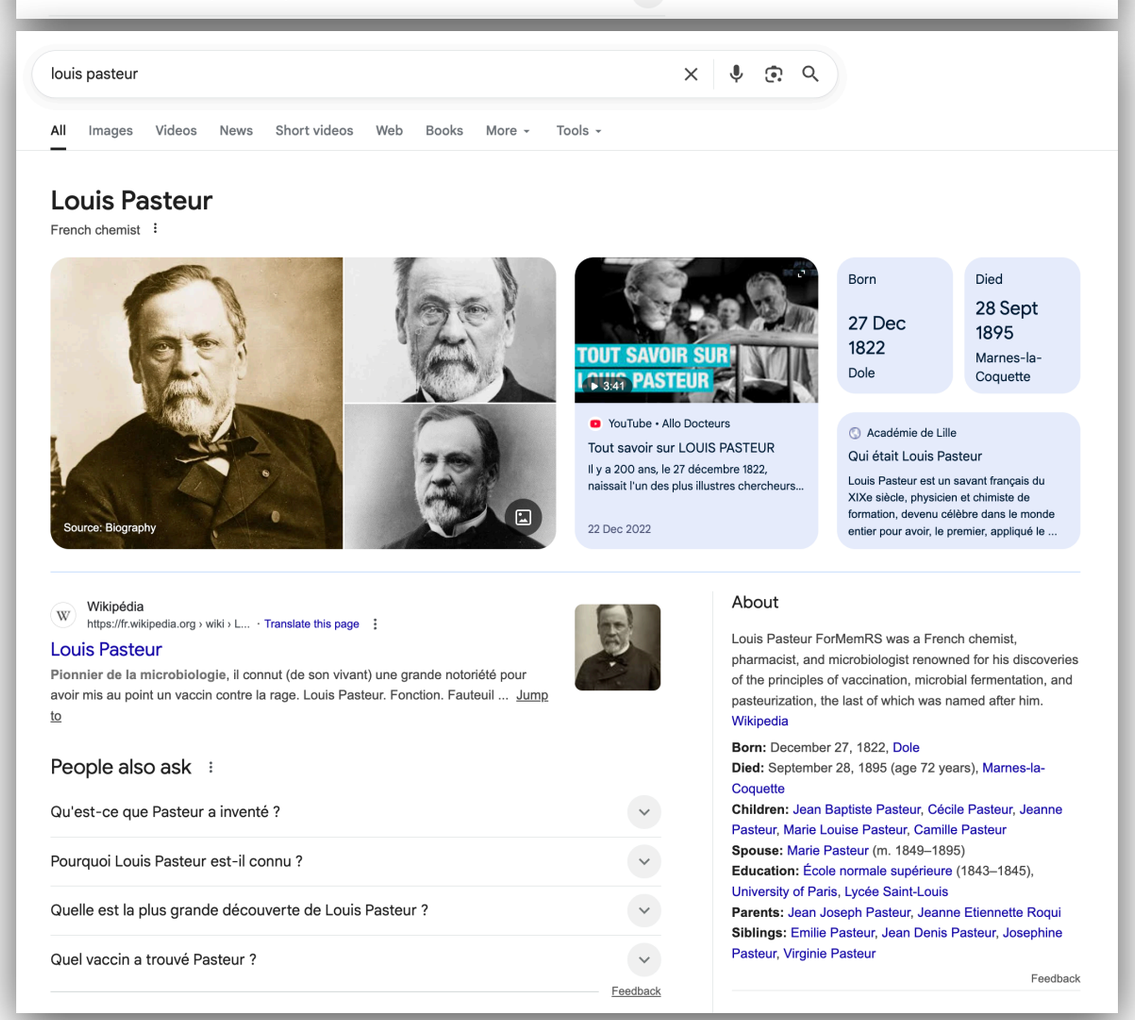




- **Technical**: exchange protocols compatible with different systems, e.g., HTTPS
- **Syntactic**: data and metadata structure can be **read/written** by different systems
  → formats
- **Semantic**: data and metadata can be **understood/actioned** upon by different systems
  → controlled vocabularies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

# Being **findable** by both human and machines ?



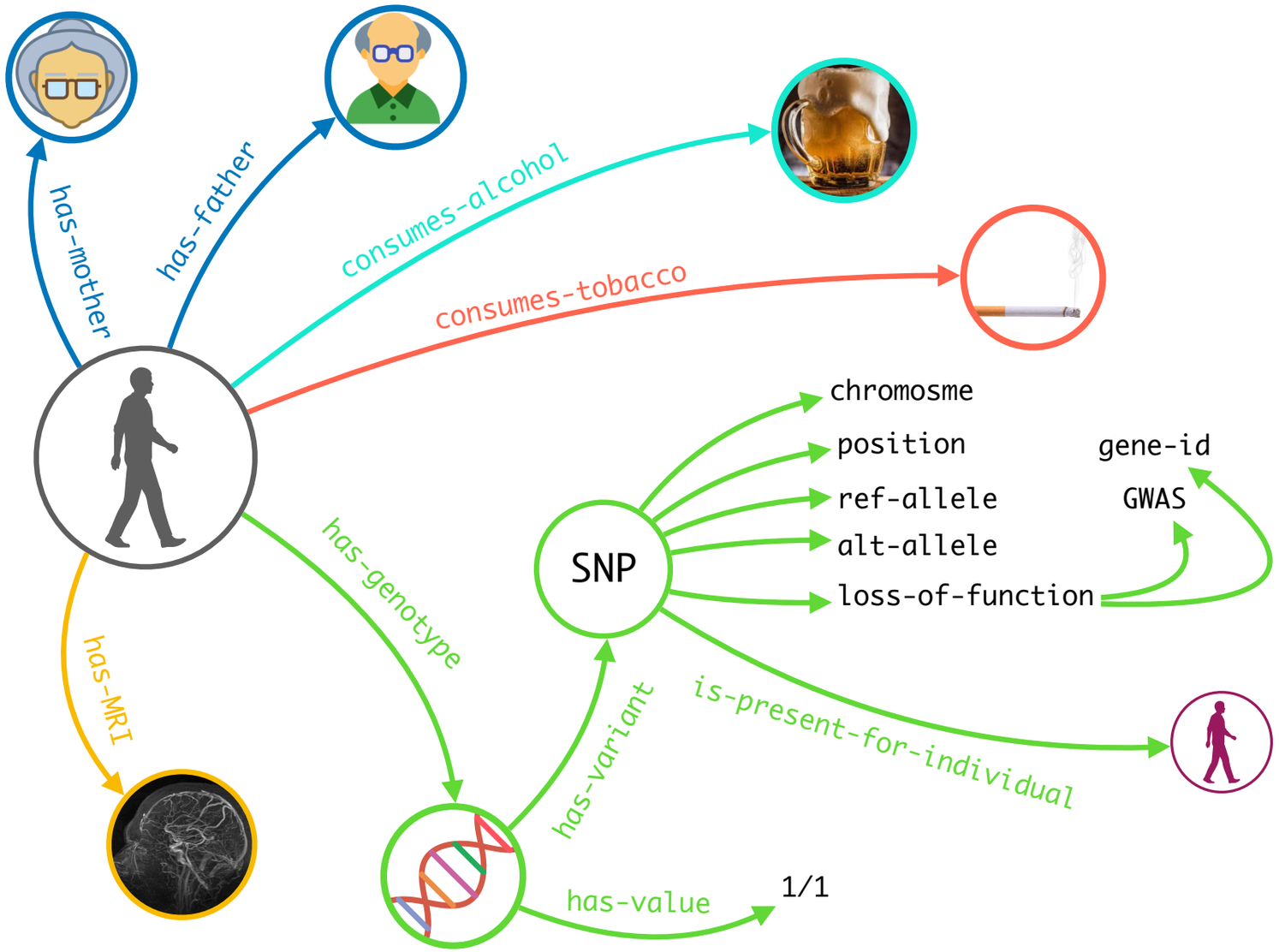




- "Pasteur" is not a good **identifier**
  → used to designate many "Humans"
  → used to designate many locations or institutions

- **Knowledge representation** is key:
  - Entity kind/nature
    - Person, Research Institute
  - Relationships with other entities
    - Children, Education,

# ❶ inter-linking data with knowledge graphs

« a collection of interlinked descriptions of things
(real-word 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

WIKIPEDIA
The Free Encyclopedia

Article | Talk | Read | Edit | View history | Search Wikipedia

## 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.[5][6] This gene can produce a variety of alternatively spliced versions of the Rac1 protein, which appear to carry out different functions.[7]


Contents [hide]
1 Function
2 Role in cancer
3 Role in glucose transport
4 Clinical significance
5 Interactions
6 References
7 Further reading
8 External links

### 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 differenti (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 e 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 ha in cancer cell invasion and metastasis.[13] Overexpression of a constitutively active Rac1 V12 in mice caused a tumor that's phenotypically indistinguishable form 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 bl

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Languages
日本語
Русский
Українська
Edit links

DBpedia | Browse using | Formats | Faceted Browser | Sparql Endpoint

## 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 |
|---|---|
| dbo:abstract | ▪ 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。 (ja)<br>▪ 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 ных процессов, таких как секреция, фагоцитоз апоптозных клеток, поляризация эпителиальных клеток и рованное факторами роста образование мембранных складок и выростов (англ. membrane ruffles). (ru)<br>also known as Ras-related C3 botulinum toxin substrate 1, is a protein found in human cells. It is encoded by the RAC1 gene. ene can produce a variety of alternatively spliced versions of the Rac1 protein, which appear to carry out different functions. (en) |
| | cmkb.cellmigration.org/report.cgi?report=orth_overview&gene_id=5879<br>www.cellmigration.org/index.shtml |
| | 705 (xsd:integer) |
| | 5305 (xsd:integer) |
| | ng<br>:Q206229<br>:Q8054<br>molecule<br>tein |

## http://dbpedia.org/sparql

Virtuoso SPARQL Query Editor

About | Namespace Prefixes | Inference rules | RDF views | iSPARQL

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.)*

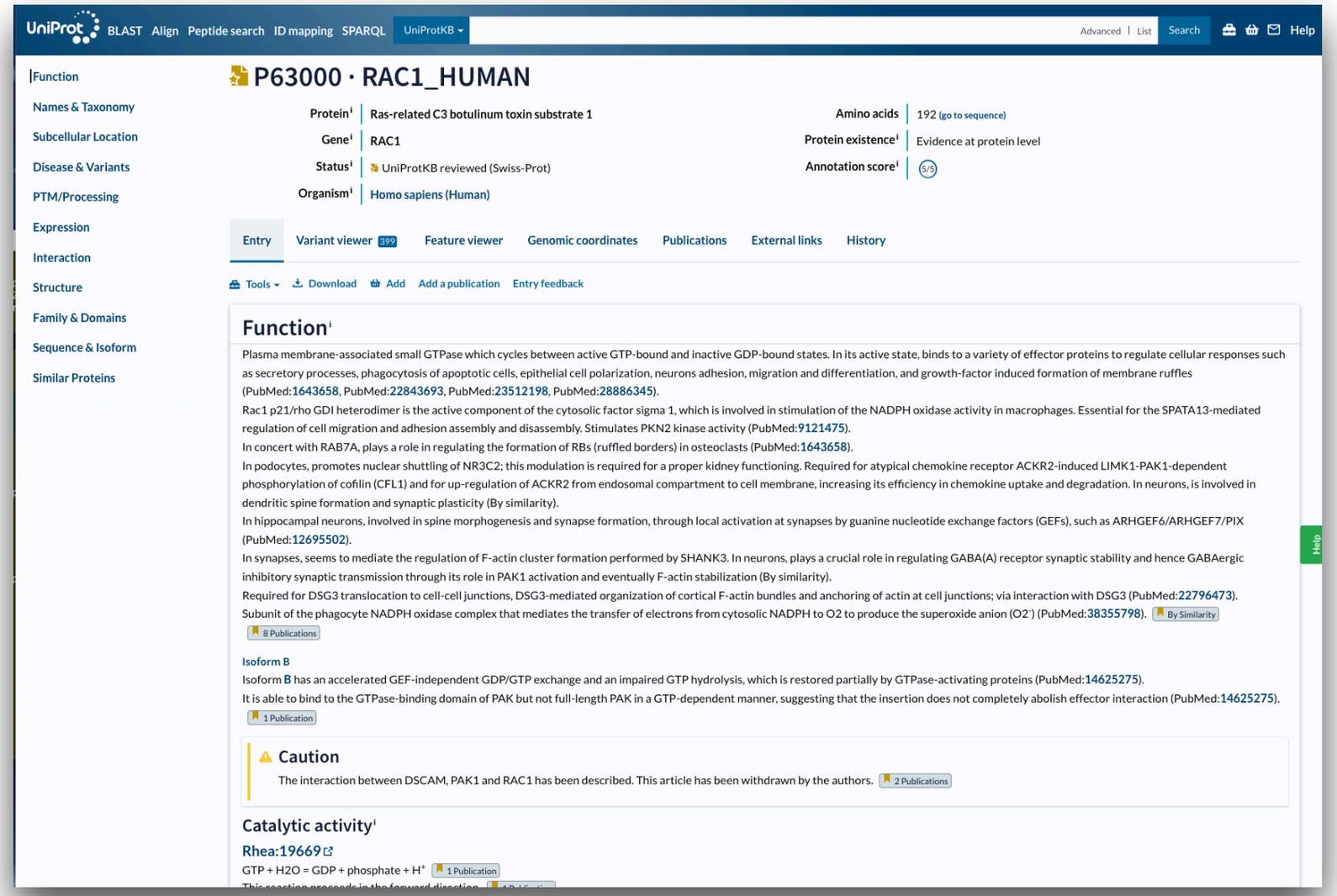
Results Format: HTML
Execution timeout: 30000 milliseconds *(values less than 1000 are ignored)*
Options:
- [x] Strict checking of void variables
- [ ] Log debug info at the end of output (has no effect on some queries and output formats)
- [ ] Generate SPARQL compilation report (instead of executing the query)

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

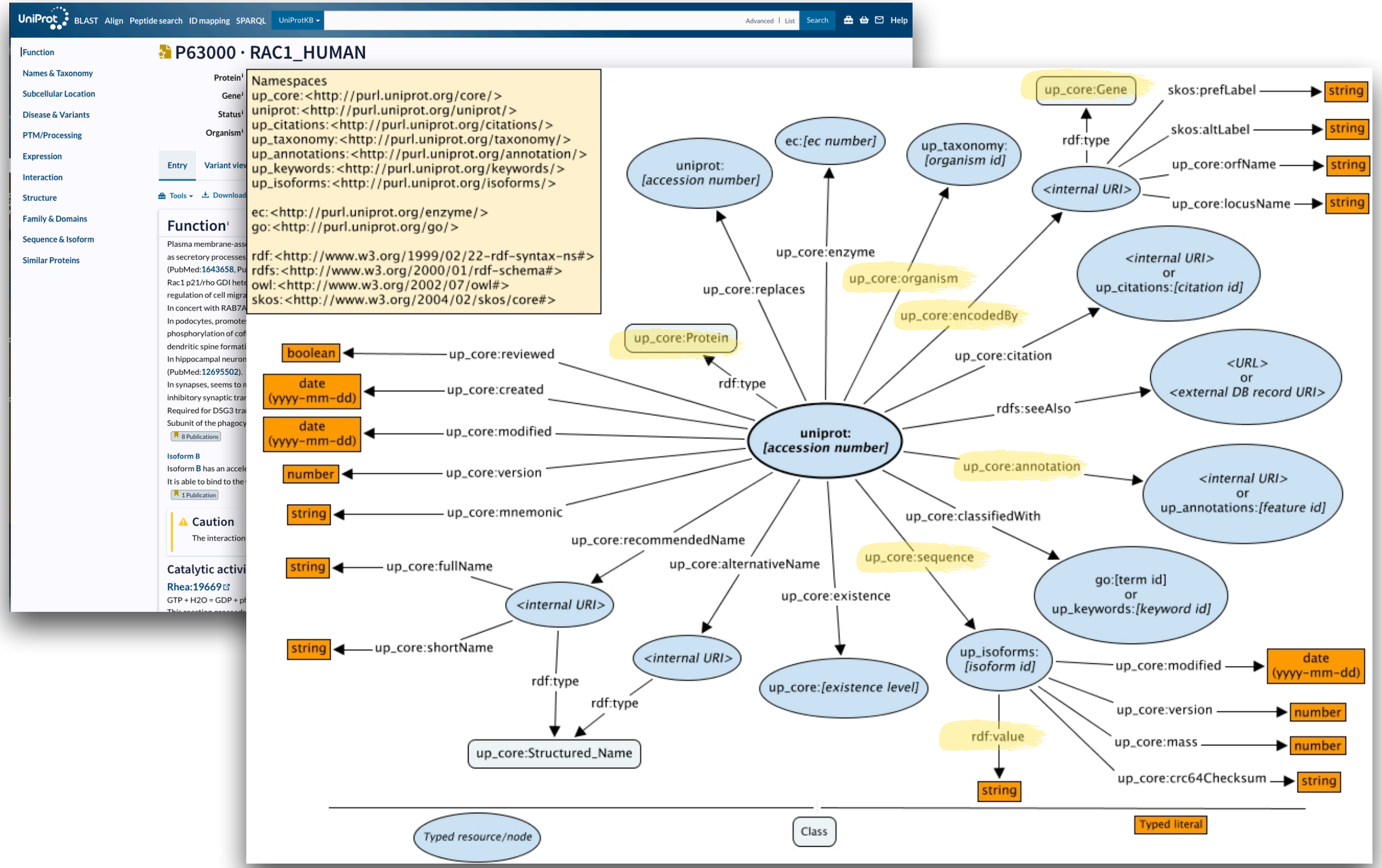
Run Query | Reset

| 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" |

# Uniprot Knowledge Graph



# Uniprot Knowledge Graph



# Uniprot Knowledge Graph

```
isoform:P06213-1 a up:Simple_Sequence ;
  up:modified "2010-10-05"^^xsd:date ;
  up:version 4 ;
  up:precursor true ;
  up:mass 156333 ;
  up:md5Checksum "8eb04104be33f0a0cb376ed145e684ff" ;
  skos:prefLabel "Long" ;
  skos:altLabel "HIR-B" ;
  rdf:value
"MATGGRRGAAAAPLLVAVAALLLGAAGHLYPGEVCPGMDIRNNLTRLHELENCSVIEGHLQILLMFKTRPEDFRDLSFPKLIMITDYLLLFRVYGLESLKDLFPNLTVIRGS
RLFFNYALVIFEMVHLKELGLYNLMNITRGSVRIEKNNELCYLATIDWSRILDSVEDNYIVLNKDDNEECGDICPGTAKGKTNCPATVINGQFVERCWTHSHCQKVCPTICKS
HGCTAEGLCCHSECLGNCSQPDDPTKCVACRNFYLDGRCVETCPPPYYHFQDWRCVNFSFCQDLHHKCKNSRRQGCHQYVIHNNKCIPECPSGYTMNSSNLLCTPCLGPCPKV
CHLLEGEKTIDSVTSAQELRGCTVINGSLIINIRGGNNLAAELEANLGLIEEISGYLKIRRSYALVSLSFFRKLRLIRGETLEIGNYSFYALDNQNLRQLWDWSKHNLTITQG
KLFFHYNPKLCLSEIHKMEEVSGTKGRQERNDIALKTNGDQASCENELLKFSYIRTSFDKILLRWEPYWPPDFRDLLGFMLFYKEAPYQNVTEFDGQDACGSNSWTVVDIDPP
LRSNDPKSQNHPGWLMRGLKPWTQYAIFVKTLVTFSDERRTYGAKSDIIYVQTDATNPSVPLDPISVSNSSSQIILKWKPPSDPNGNITHYLVFWERQAEDSELFELDYCLKG
LKLPSRTWSPPFESEDSQKHNQSEYEDSAGECCSCPKTDSQILKELEESSFRKTFEDYLHNVVFVPRKTSSGTGAEDPRPSRKRRSLGDVGNVTVAVPTVAAFPNTSSTSVPT
SPEEHRPFEKVVNKESLVISGLRHFTGYRIELQACNQDTPEERCSVAAYVSARTMPEAKADDIVGPVTHEIFENNVVHLMWQEPKEPNGLIVLYEVSYRRYGDEELHLCVSRK
HFALERGCRLRGLSPGNYSVRIRATSLAGNGSWTEPTYFYVTDYLDVPSNIAKIIIGPLIFVFLFSVVIGSIYLFLRKRQPDGPLGPLYASSNPEYLSASDVFPCSVYVPDEW
EVSREKITLLRELGQGSFGMVYEGNARDIIKGEAETRVAVKTVNESASLRERIEFLNEASVMKGFTCHHVVRLLGVVSKGQPTLVVMELMAHGDLKSYLRSLRPEAENNPGRP
PPTLQEMIQMAAEIADGMAYLNAKKFVHRDLAARNCMVAHDFTVKIGDFGMTRDIYETDYYRKGGKGLLPVRWMAPESLKDGVFTTSSDMWSFGVVLWEITSLAEQPYQGLSN
EQVLKFVMDGGYLDQPDNCPERVTDLMRMCWQFNPKMRPTFLEIVNLLKDDLHPSFPEVSFFHSEENKAPESEELEMEFEDMENVPLDRSSHCQREEAGGRDGGSSLGFKRSY
EEHIPYTHMNGGKKNGRILTLPRSNPS" .
```

There are 217,505,202,099 triples is 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 | 160,189,731,200 | 40,455,837,024 | 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

All protein sequences associated to Escherichia coli K-12 ?

UniProt SPARQL Downloads Documentation/Help

## Your SPARQL query

Add common prefixes

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX taxon: <http://purl.uniprot.org/taxonomy/>
PREFIX up: <http://purl.uniprot.org/core/>
SELECT ?protein ?organism ?isoform ?sequence
WHERE
{
    ?protein a up:Protein .
    ?protein up:organism ?organism .
    # Taxon subclasses are materialized, do not use rdfs:subClassOf+
    ?organism rdfs:subClassOf taxon:83333 .
    ?protein up:sequence ?isoform .
    ?isoform rdf:value ?sequence .
}
```

Submit Query

## 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

## 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

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

... how can we ensure machines/humans "speak" the **same language** ? (semantic interoperability)

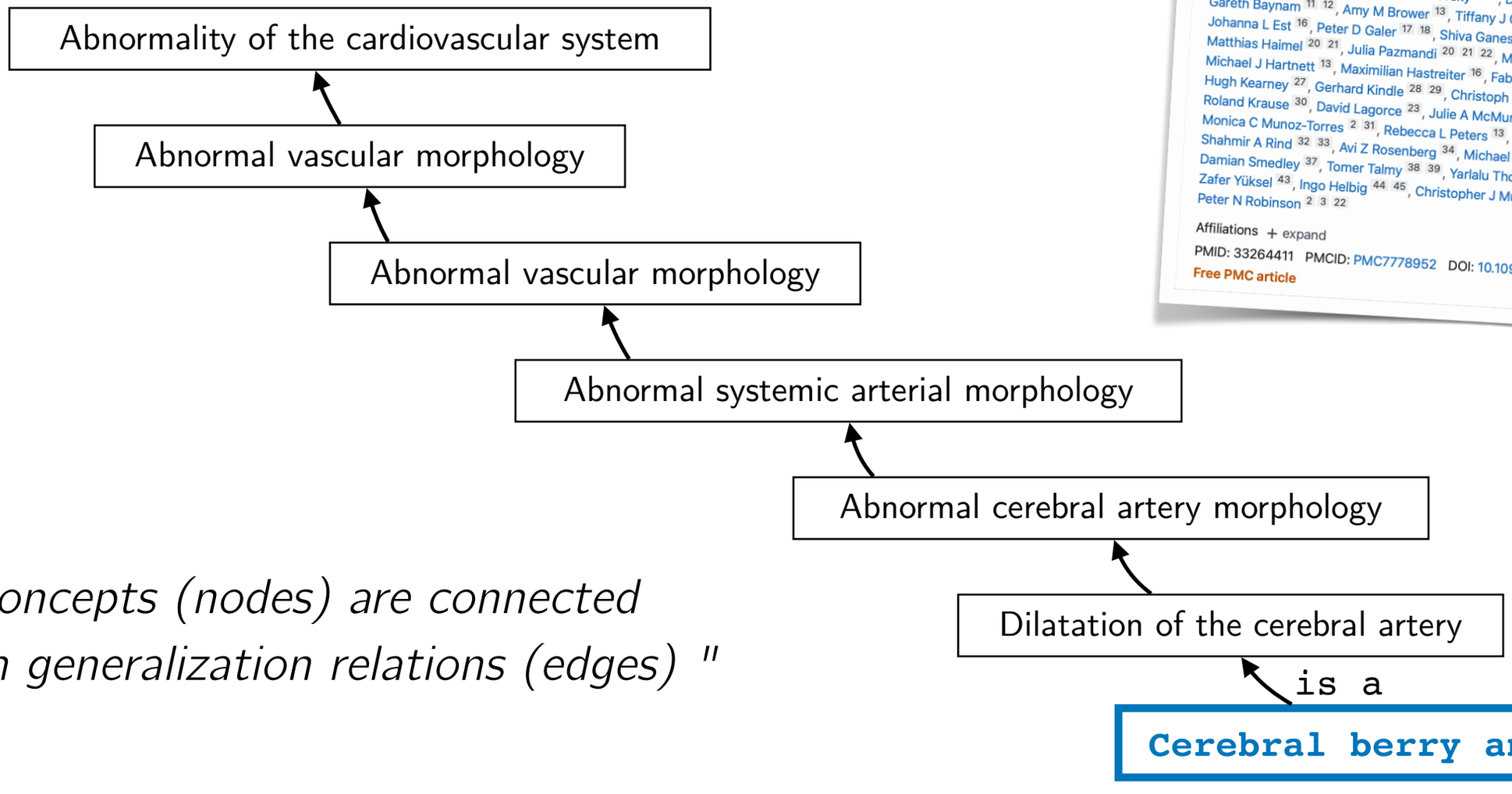
# ❷ Uniformly describe **what** is observed with data ?

## Computational ontology

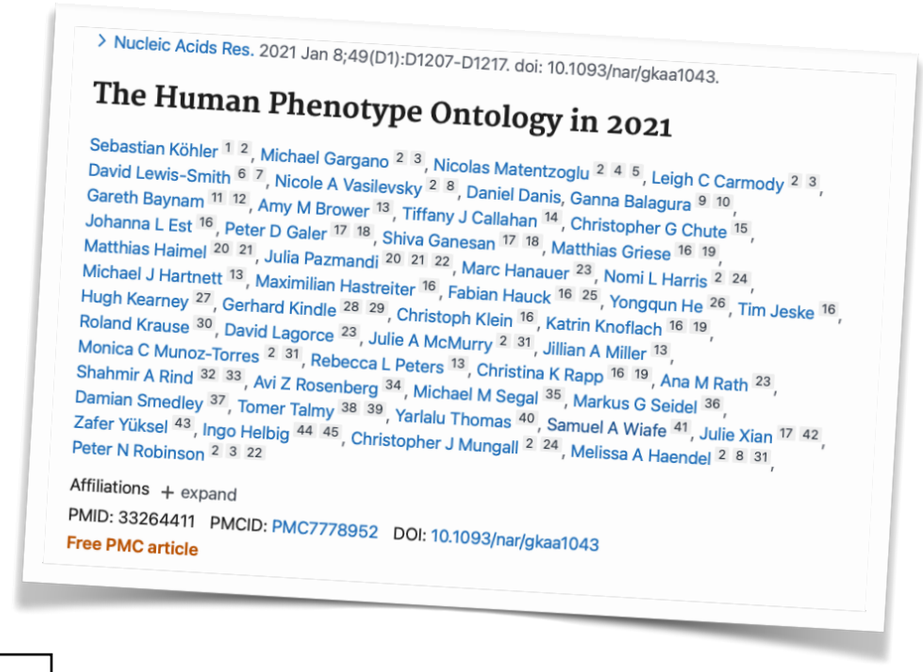
« a formal specification of a shared conceptualization » (Borst, 1997)

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

## Human Phenotype Ontology




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



> Nucleic Acids Res. 2021 Jan 8;49(D1):D1207-D1217. doi: 10.1093/nar/gkaa1043.

**The Human Phenotype Ontology in 2021**

Sebastian Köhler [1 2], Michael Gargano [2 3], Nicolas Matentzoglu [2 4 5], Leigh C Carmody [2 3], David Lewis-Smith [6 7], Nicole A Vasilevsky [2 8], Daniel Danis, Ganna Balagura [9 10], Gareth Baynam [11 12], Amy M Brower [13], Tiffany J Callahan [14], Christopher G Chute [15], Johanna L Est [16], Peter D Galer [17 18], Shiva Ganesan [17 18], Matthias Griese [16 19], Matthias Haimel [20 21], Julia Pazmandi [20 21 22], Marc Hanauer [23], Nomi L Harris [2 24], Michael J Hartnett [13], Maximilian Hastreiter [16], Fabian Hauck [16 25], Yongqun He [26], Tim Jeske [16], Hugh Kearney [27], Gerhard Kindle [28 29], Christoph Klein [16], Katrin Knoflach [16 19], Roland Krause [30], David Lagorce [23], Julie A McMurry [2 31], Jillian A Miller [13], Monica C Munoz-Torres [2 31], Rebecca L Peters [13], Christina K Rapp [16 19], Ana M Rath [23], Shahmir A Rind [32 33], Avi Z Rosenberg [34], Michael M Segal [35], Markus G Seidel [36], Damian Smedley [37], Tomer Talmy [38 39], Yarlalu Thomas [40], Samuel A Wiafe [41], Julie Xian [17 42], Zafer Yüksel [43], Ingo Helbig [44 45], Christopher J Mungall [2 24], Melissa A Haendel [2 8 31], Peter N Robinson [2 3 22]

Affiliations + expand

PMID: 33264411 PMCID: PMC7778952 DOI: 10.1093/nar/gkaa1043

Free PMC article

# BioPortal

BioPortal Ontologies Search Annotator Recommender Mappings Login

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

### Search for a class

Enter a class, e.g. Melanoma

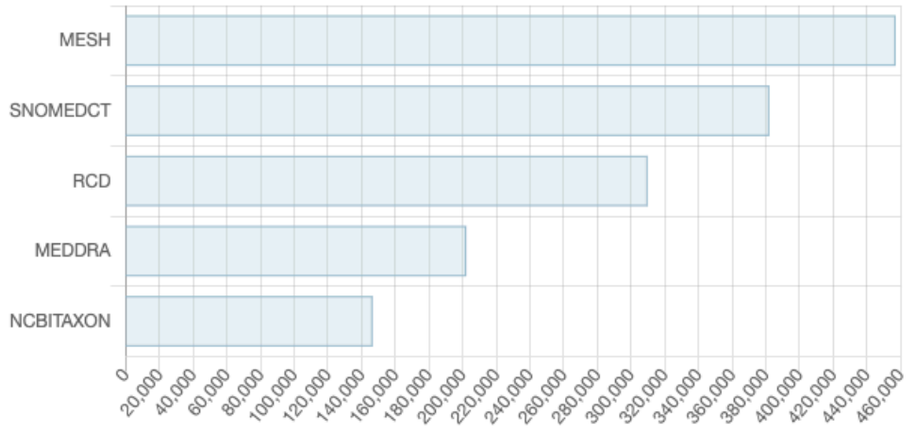
Advanced search

### 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



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

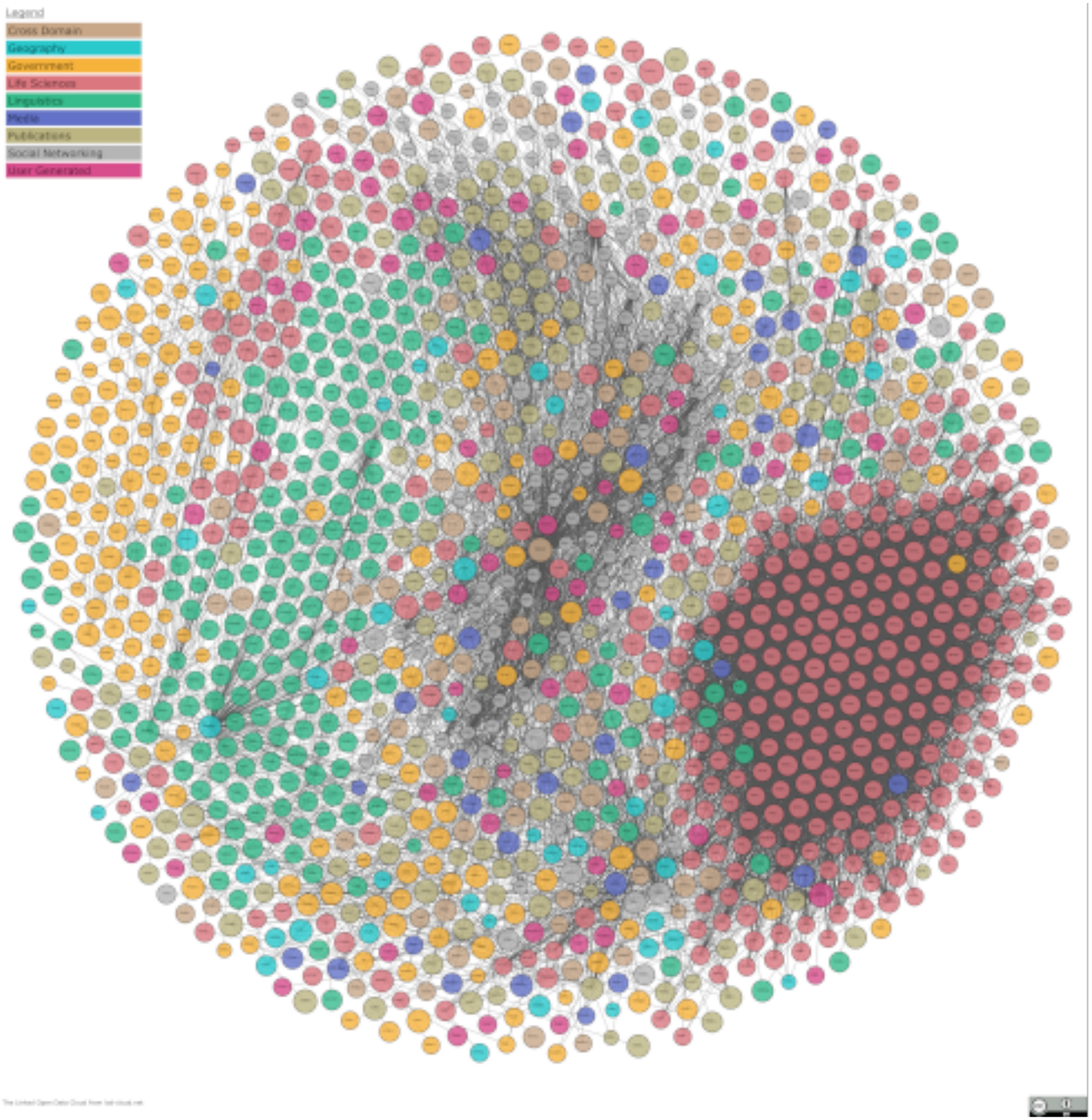
Details | Visualization | Notes (0) | Mappings ( 8 )

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_reference | SPIRE:Tundra http://sweetontology.net/realmCryo/AlpineTundra |
| in_subset | envoPolar |

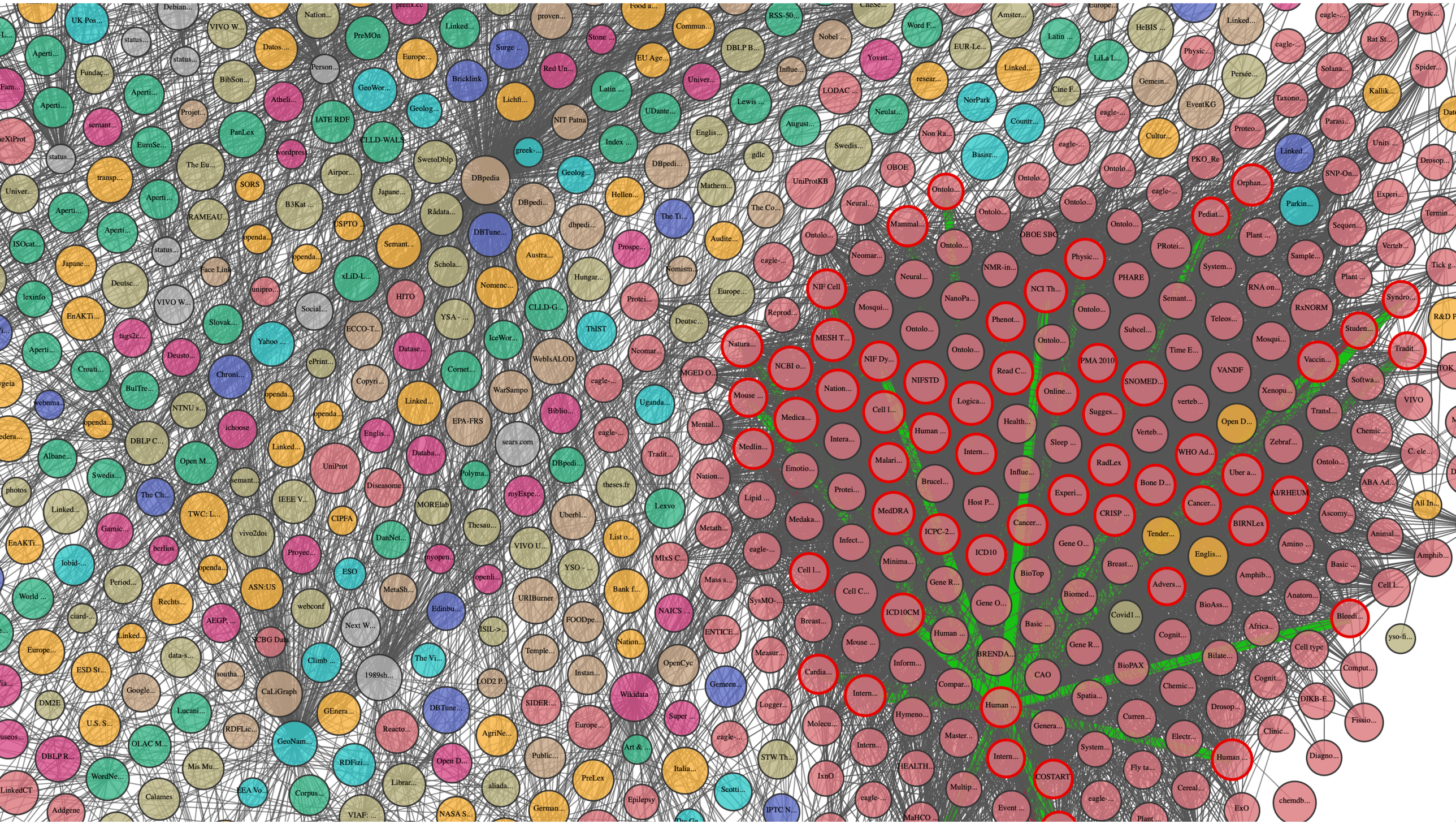



Linked Open Data Cloud → 1357 interlinked RDF datasets in 2025



https://lod-cloud.net/versions/2026-02-05/lod-cloud.svg

# 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

  → (semantic) Interoperability ✅

  → Reuse ✅

# Transforming your data

# → Knowledge graphs

# RDF to link data

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> .
```

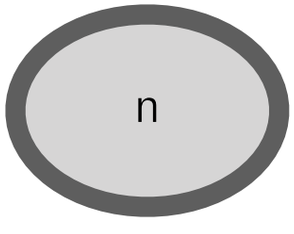
**Definitions**

(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**

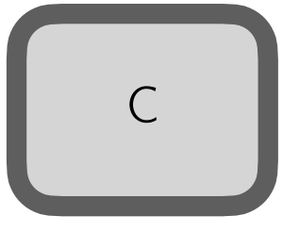
# Graphical syntax




: a node in the knowledge graph

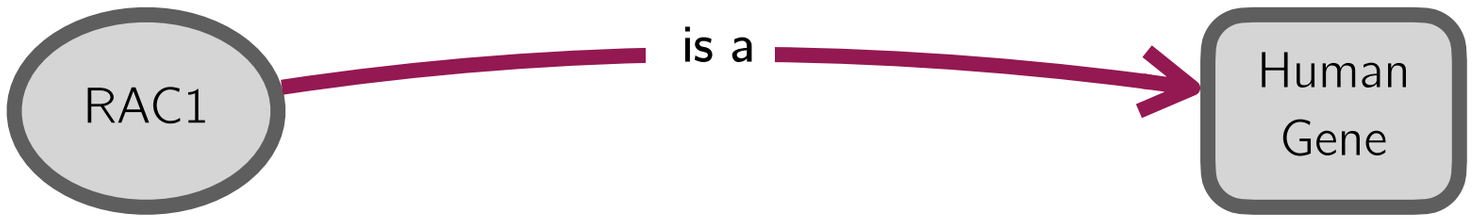

: a property/relation/edge in the knowledge graph

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

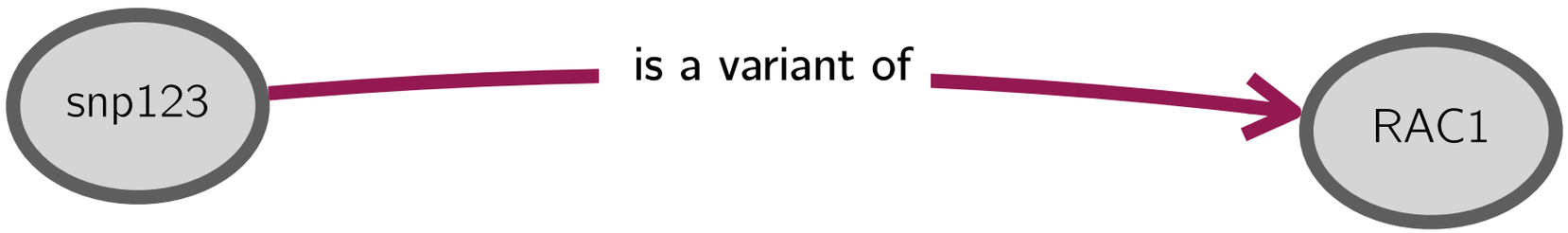



: a special node, denoting an ontology concept/class

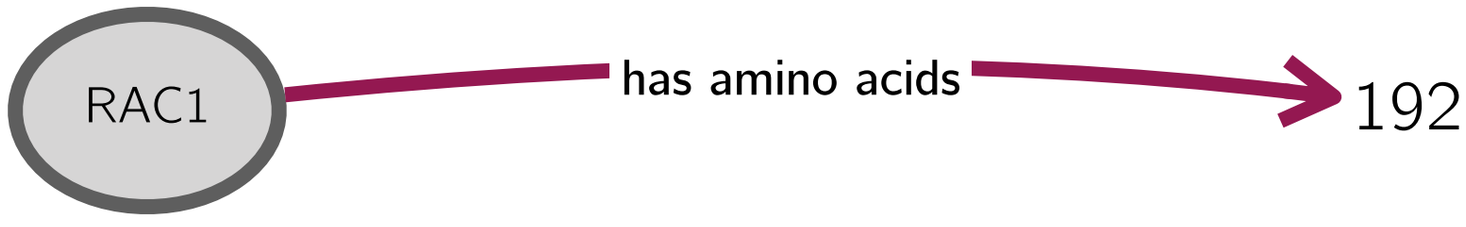
# Examples




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




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



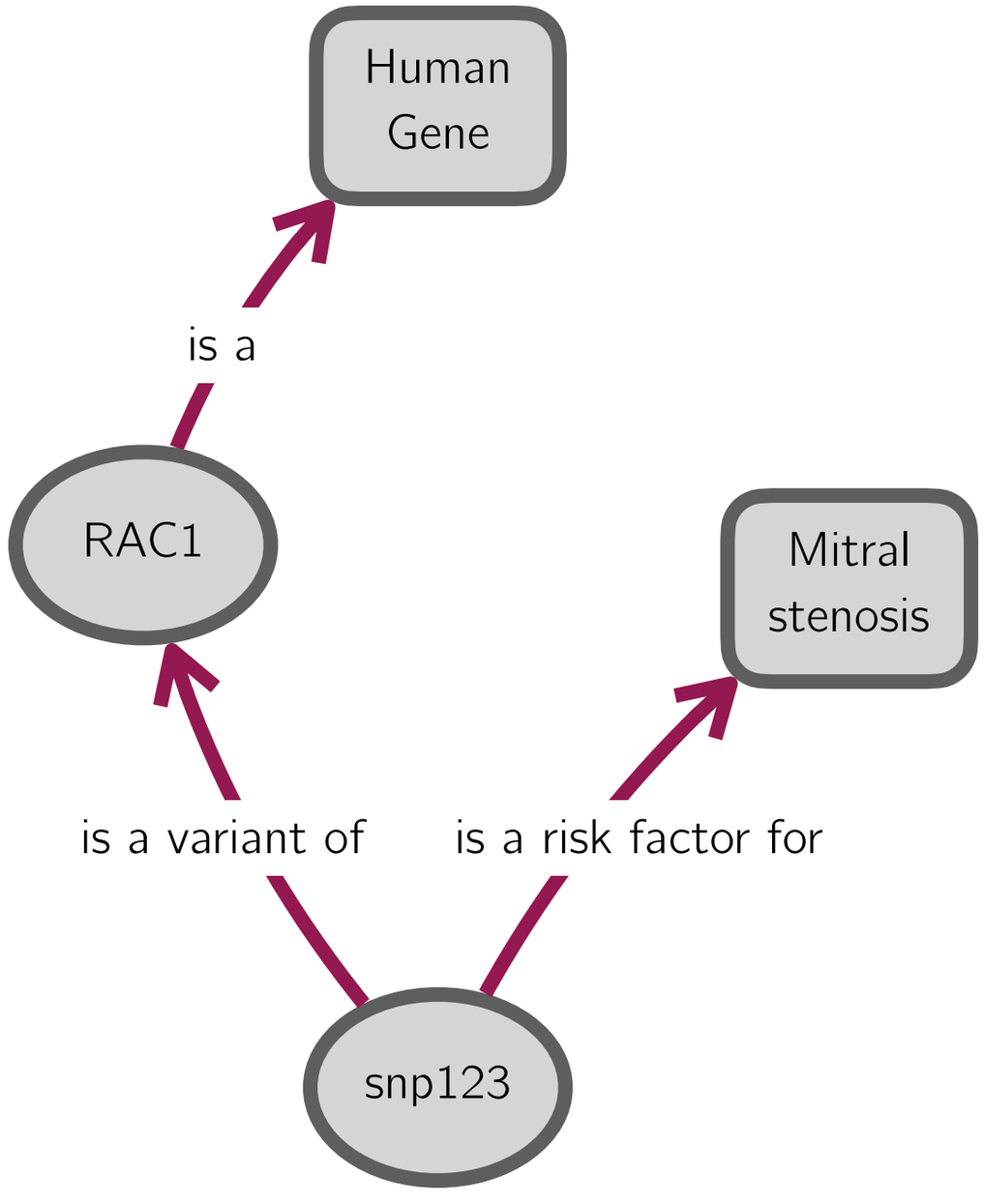

```
<http://RAC1>  <http://has_amino_acids>  192 .
```

# 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 endswith a **.**

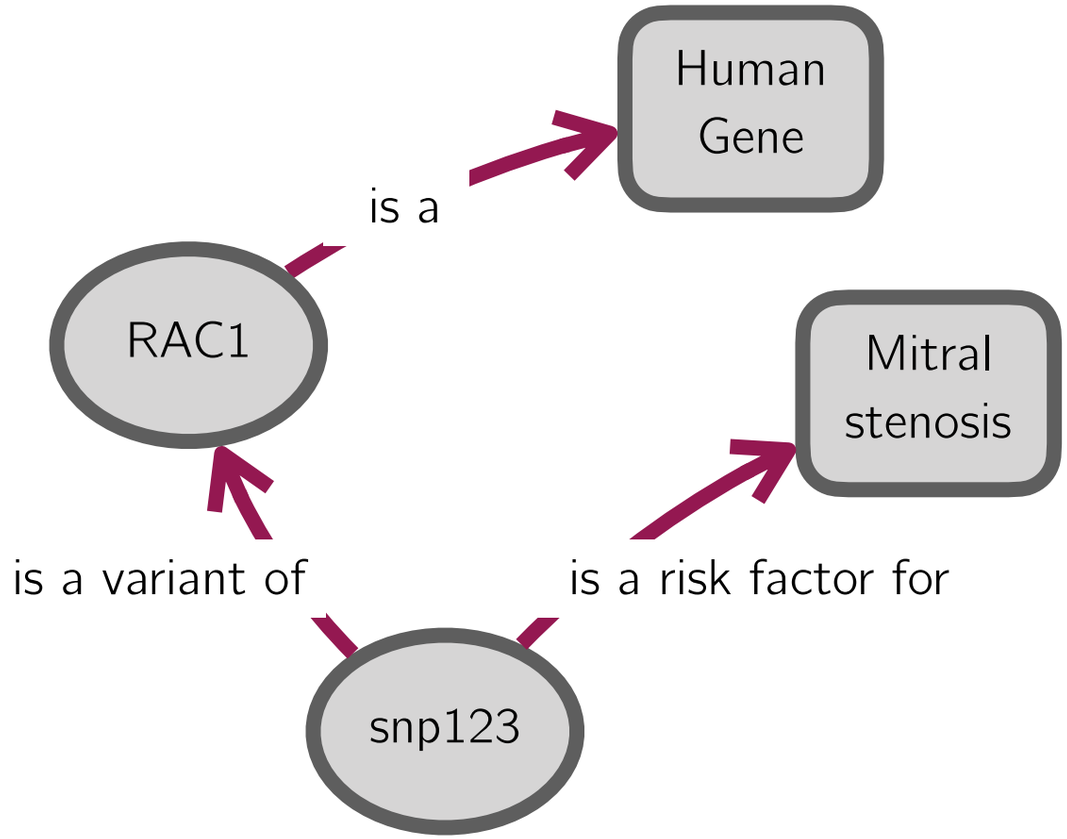
S P O .

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

S $P_1$ $O_1$ ;
$P_2$ $O_2$ .

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

S P $O_1$ , $O_2$ .




```
@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 ✏️

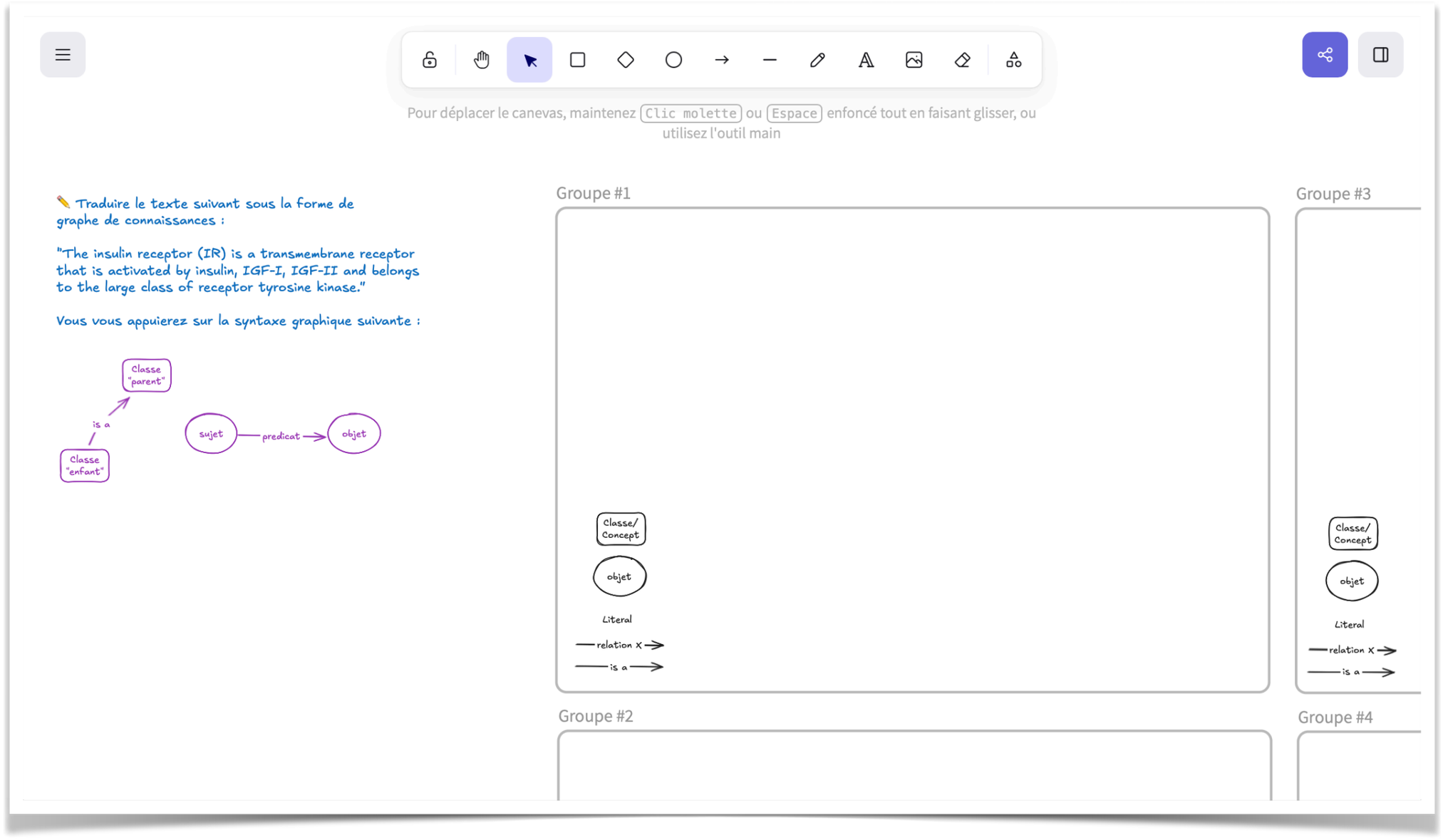
Question #1

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 $\rightarrow$ RDF predicates
- ✓ Identify linked entities,
  - who is a subject of a relation ?
  - who is the object of a relation ?

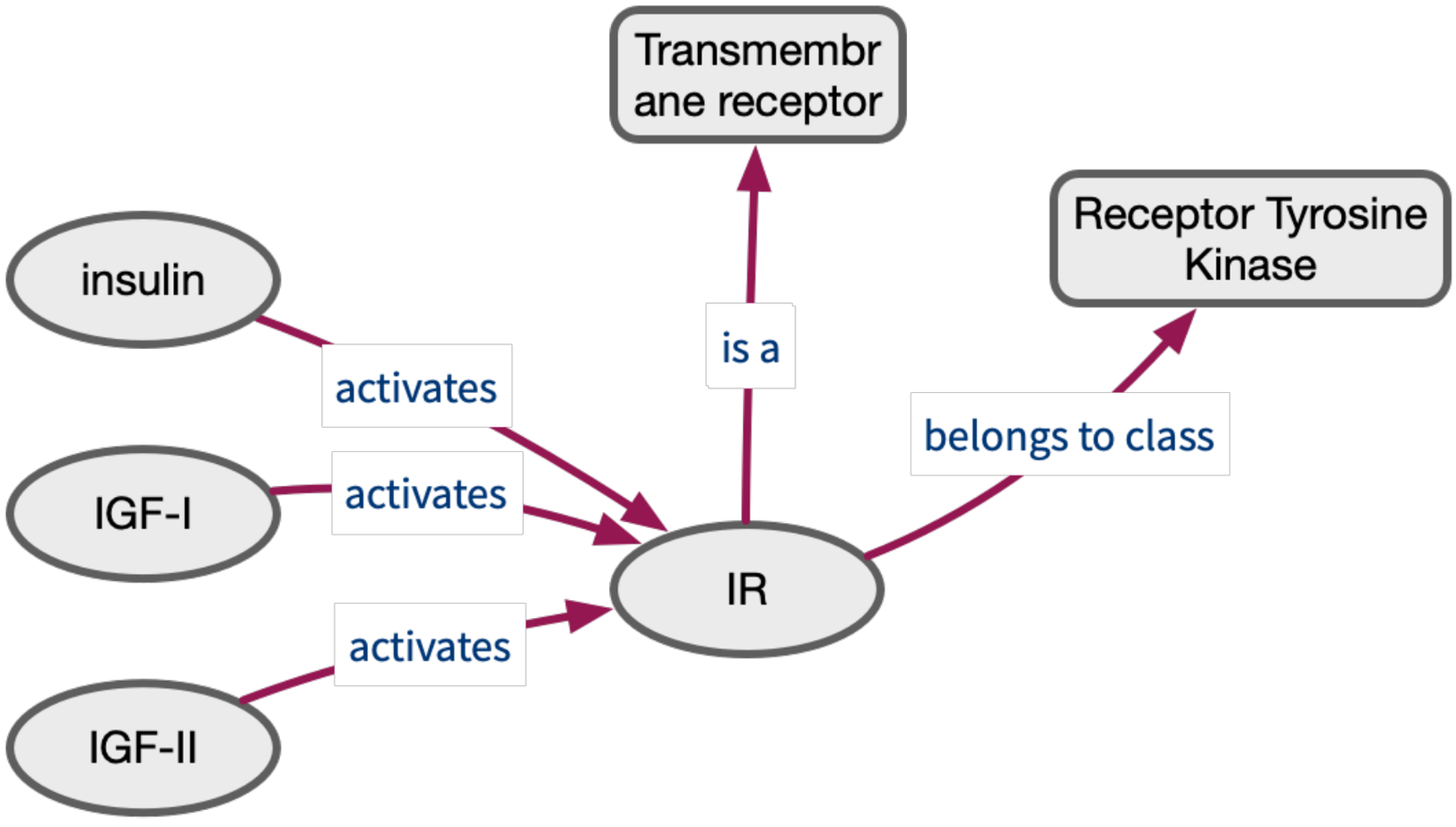
# "Pen & paper" team work




https://tinyurl.com/tvyeufcr

# 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."*



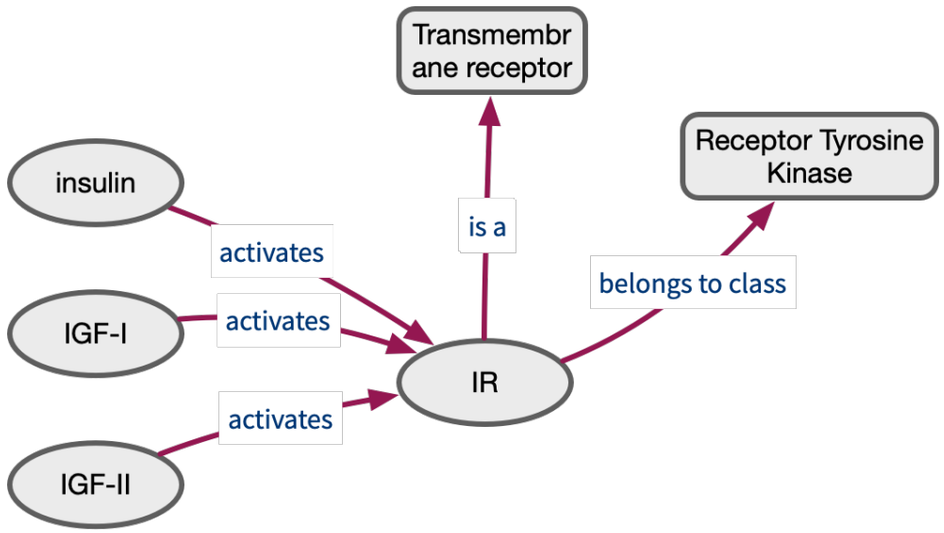
# Hands-on session: from text to KG ✏️

**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 ✏️

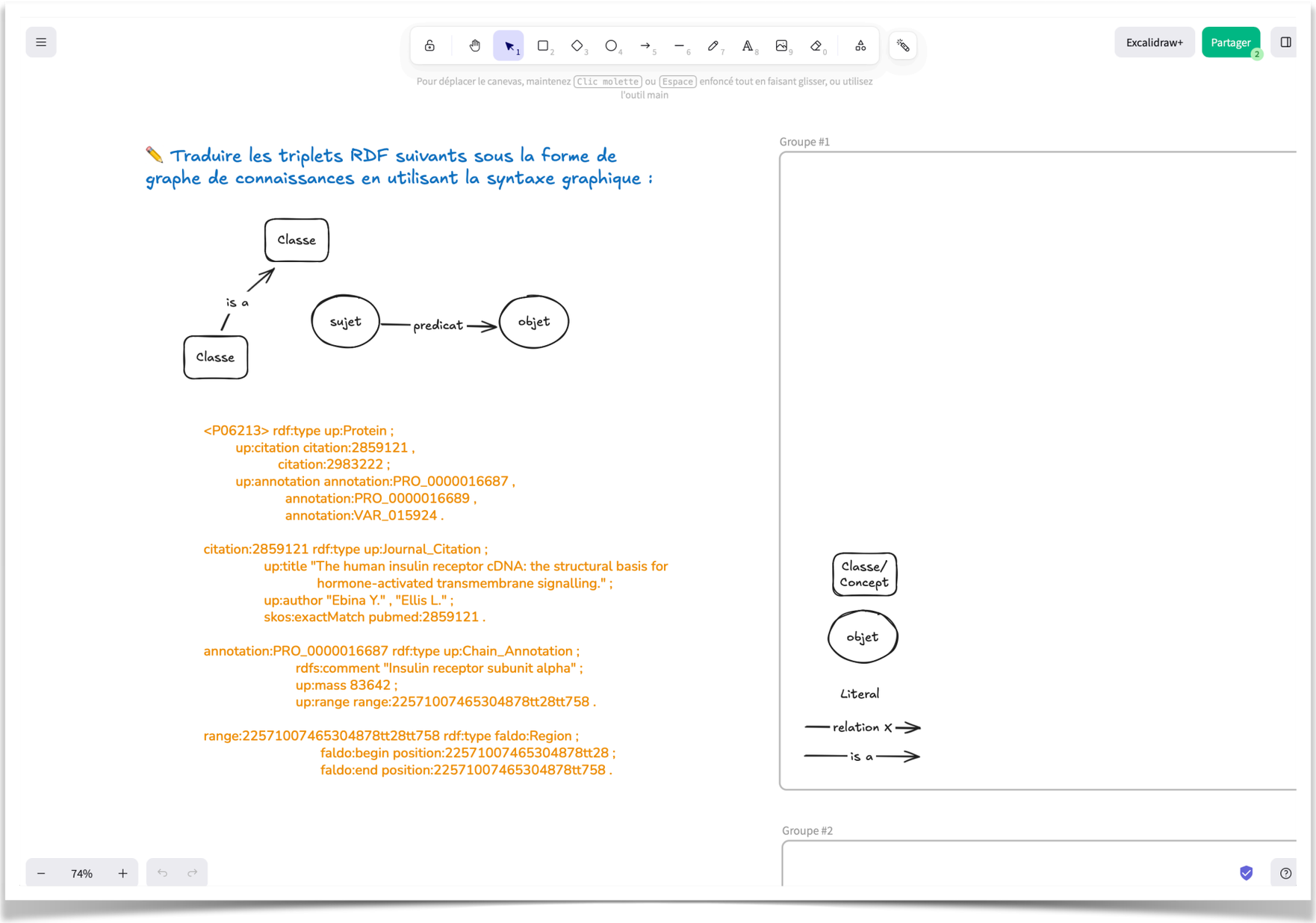
https://rest.uniprot.org/uniprotkb/P06213.ttl

```
@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:225710074653048 78tt28tt758 .

range:225710074653048 78tt28tt758 rdf:type faldo:Region ;
                                  faldo:begin position:225710074653048 78tt28 ;
                                  faldo:end position:225710074653048 78tt758 .
```

## Question #3

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

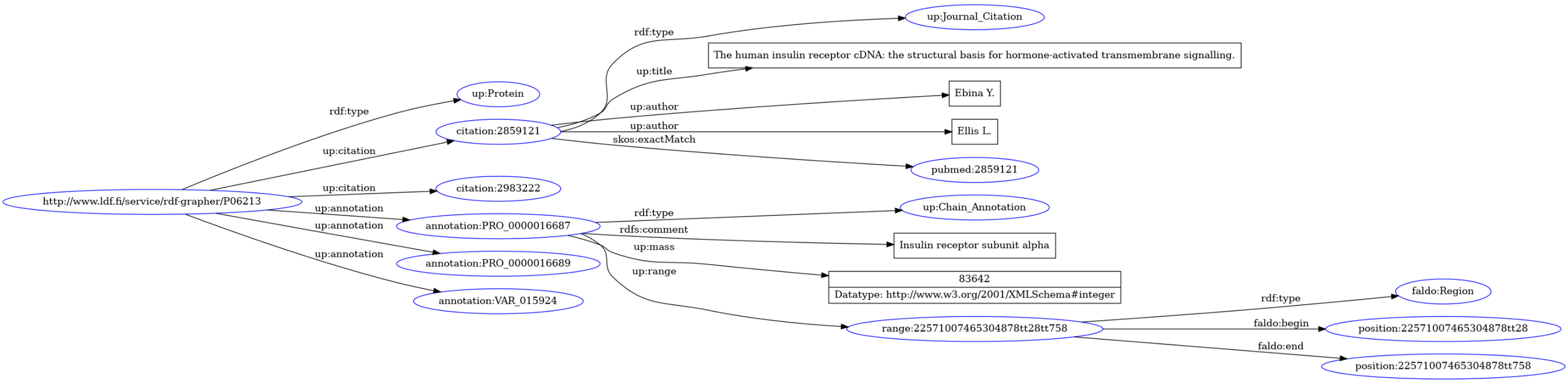
# "Pen & paper" team work




https://tinyurl.com/4xf83nxd

# Practice ... from KG to text

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



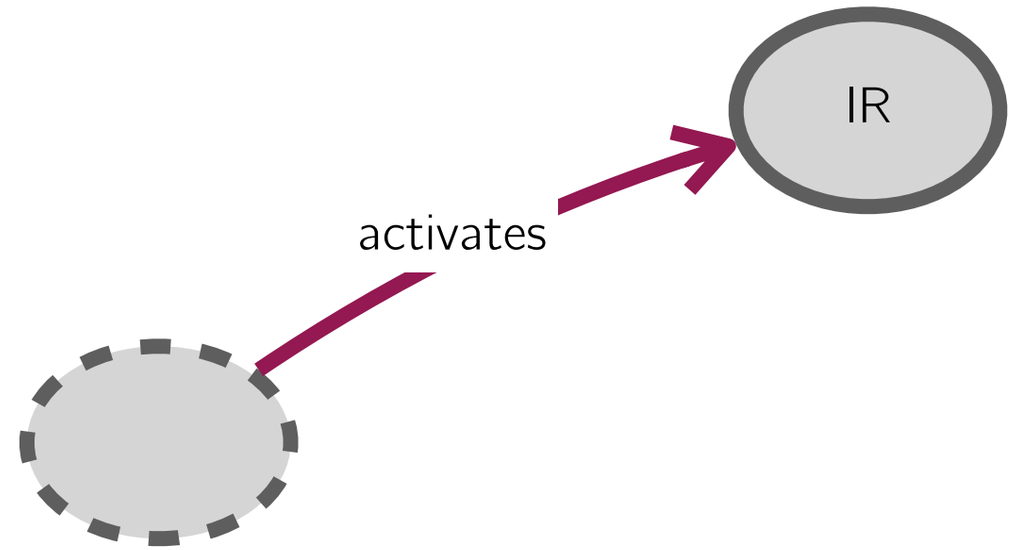

Namespaces:
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
up: http://purl.uniprot.org/core/
annotation: http://purl.uniprot.org/annotation/
citation: http://purl.uniprot.org/citations/
range: http://purl.uniprot.org/range/
faldo: http://biohackathon.org/resource/faldo#
skos: http://www.w3.org/2004/02/skos/core#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

# Querying with graph patterns

# Triple patterns

SPARQL is the W3C language to query multiple data sources expressed in RDF.

The principle consists in defining **graph patterns** to be **matched** against an RDF graph.




```
?x   ns:activates   ns:IR .
```

**Definition**

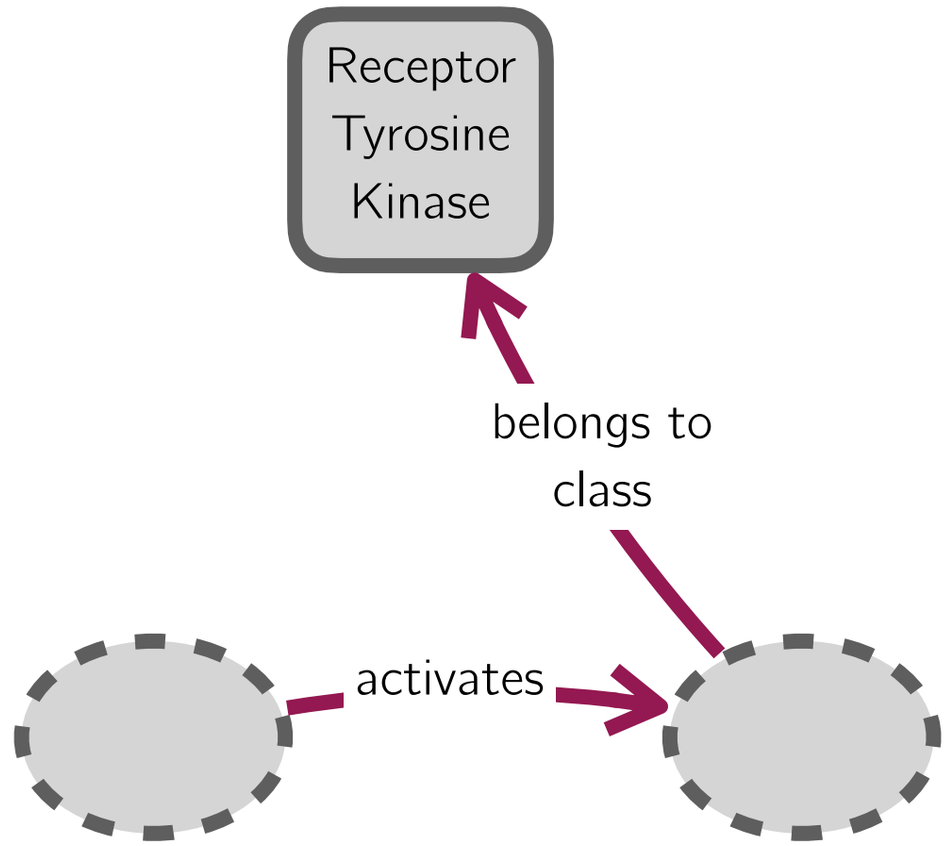
**Triple Patterns** (TP) are similar to RDF triples except that each of the *subject*, *predicate* or *object* may be a **variable**.

Variables are prefixed with a **?** .

# Basic graph patterns

**Definition**

**Basic Graph Patterns** (BGPs) consist in **a set of triple patterns** to be matched on an RDF graph.



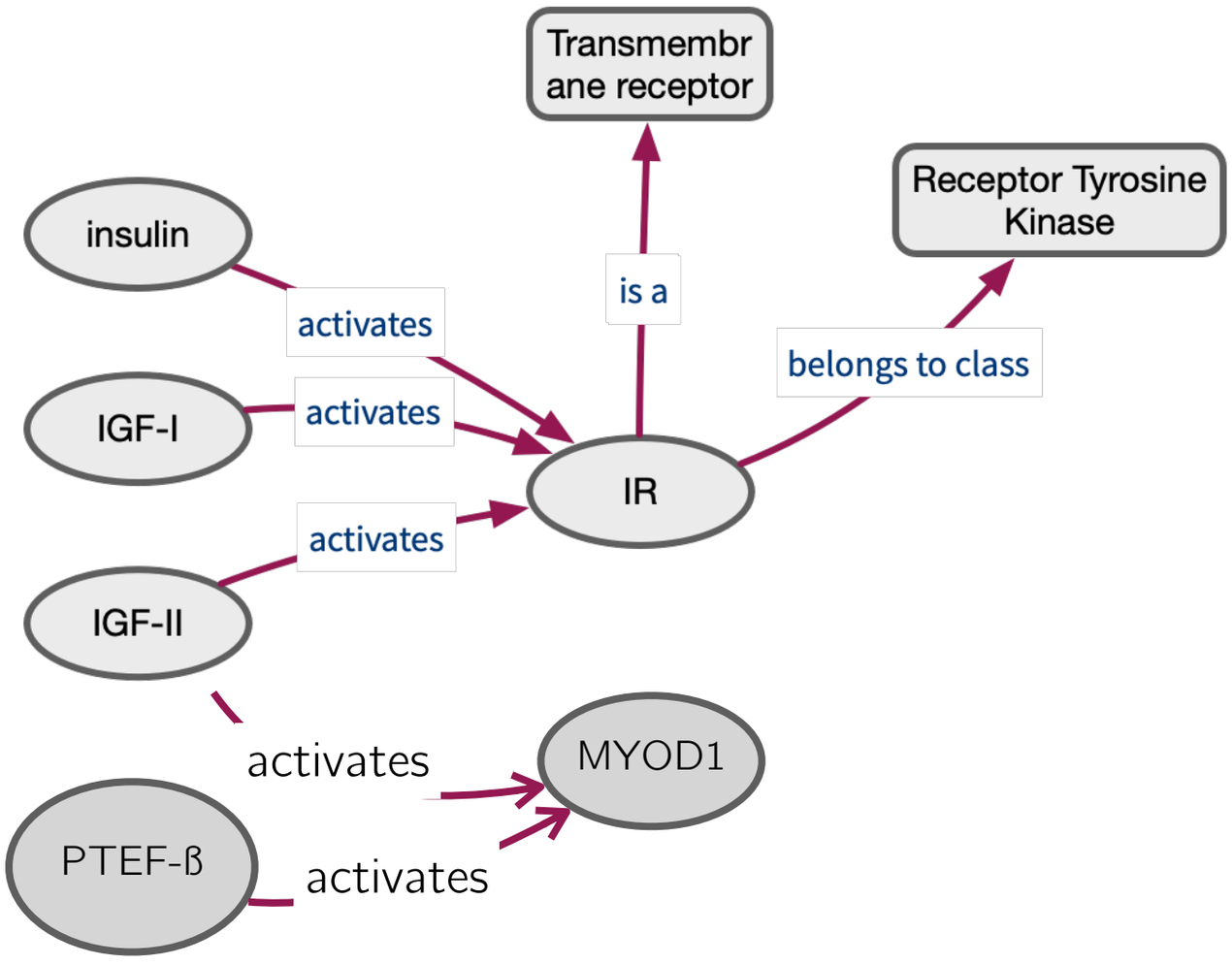

*Give me all known activators of any Receptor Tyrosine Kinase ?*

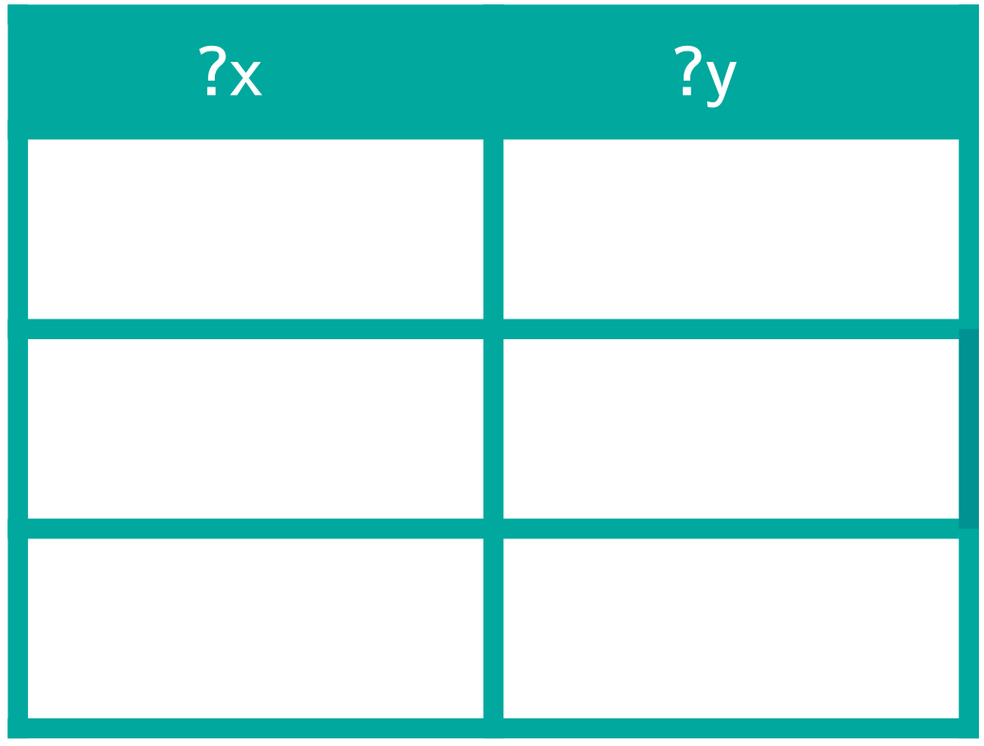
→ **all entities** that **activate** **something** that **belongs to class** "Receptor Tyrosine Kinase"

```
?x   ns:activates    ?y .
?y   ns:belongs_to_class
                    ns:ReceptorTyrosineKinase .
```

# SPARQL query evaluation

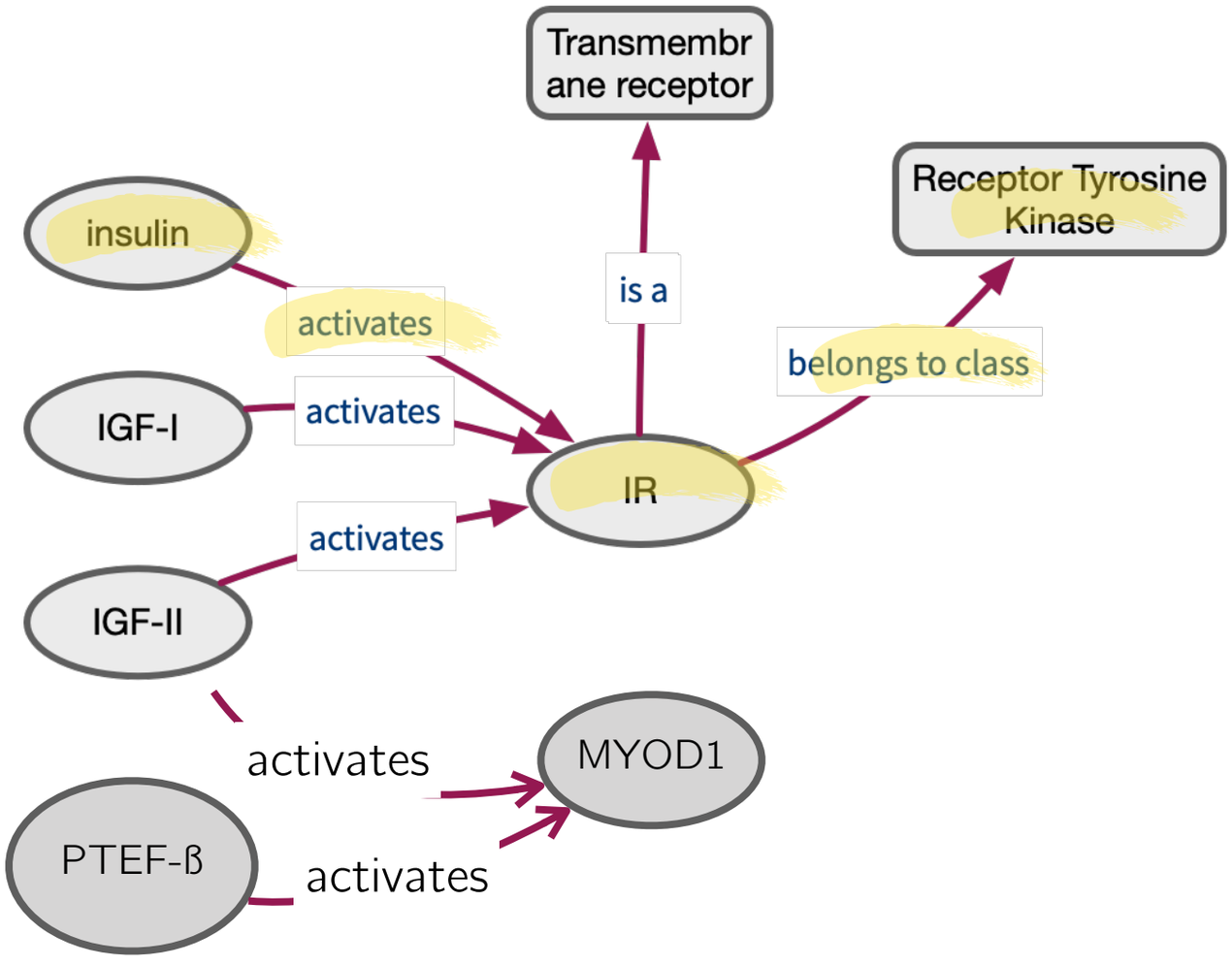
```
?x    ns:activates     ?y .
?y    ns:belongs_to_class
                  ns:ReceptorTyrosineKinase .
```



# SPARQL query evaluation

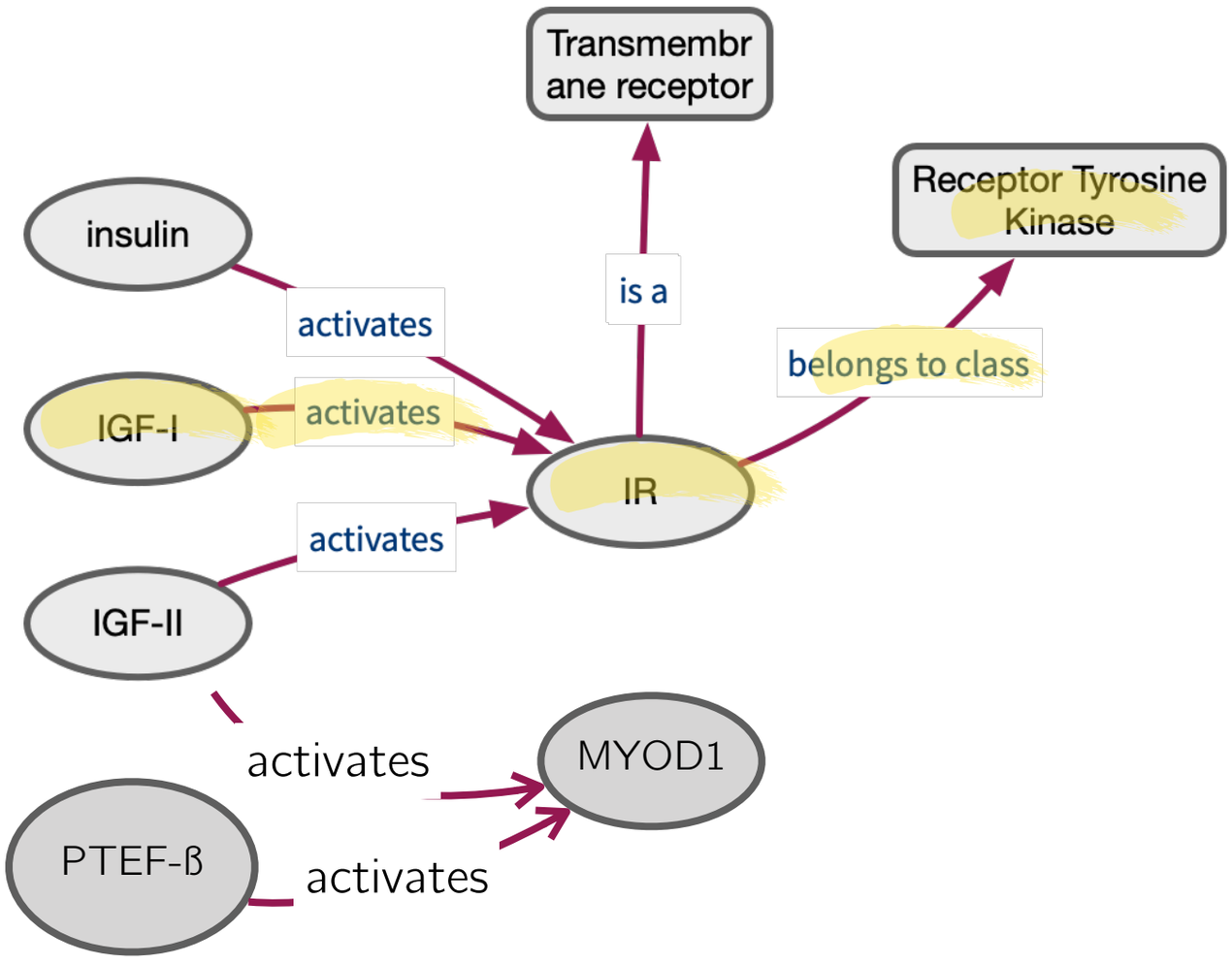
```
?x   ns:activates    ?y .
?y   ns:belongs_to_class
                  ns:ReceptorTyrosineKinase .
```




| ?x | ?y |
|---|---|
| insulin | IR |
| | |
| | |

# SPARQL query evaluation

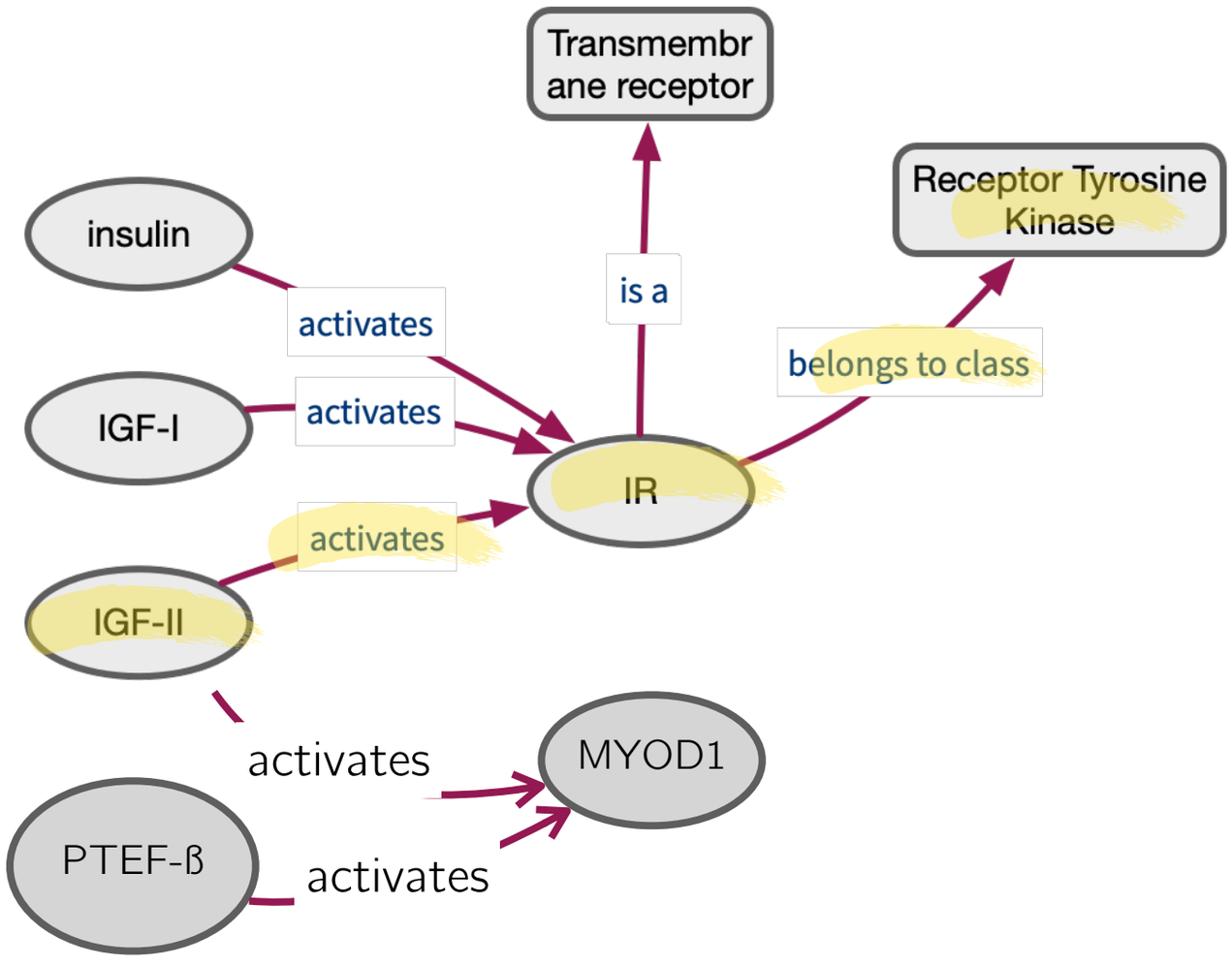
```
?x    ns:activates     ?y .
?y    ns:belongs_to_class
                  ns:ReceptorTyrosineKinase .
```




| ?x | ?y |
|---|---|
| IGF-I | IR |
| insulin | IR |
| | |

# SPARQL query evaluation

```
?x    ns:activates     ?y .
?y    ns:belongs_to_class
                  ns:ReceptorTyrosineKinase .
```




| ?x | ?y |
|---|---|
| IGF-II | IR |
| IGF-I | IR |
| insulin | IR |

# Typical SPARQL query

Shortcuts definition

```
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 dc:       <http://purl.org/dc/elements/1.1/>
4  PREFIX wp:       <http://vocabularies.wikipathways.org/wp#>
5  PREFIX dcterms:  <http://purl.org/dc/terms/>
6  PREFIX identifiers:<http://identifiers.org/ensembl/>
7  PREFIX atlas: <http://rdf.ebi.ac.uk/resource/atlas/>
8  PREFIX atlasterms: <http://rdf.ebi.ac.uk/terms/atlas/>
9  PREFIX efo: <http://www.ebi.ac.uk/efo/>
10
11 SELECT DISTINCT ?wpURL ?pwTitle ?expressionValue ?pvalue where {
12
13 SERVICE <https://www.ebi.ac.uk/rdf/services/atlas/sparql> {
14       ?factor rdf:type efo:EFO_0000270 .
15       ?value atlasterms:hasFactorValue ?factor .
16       ?value atlasterms:isMeasurementOf ?probe .
17       ?value atlasterms:pValue ?pvalue .
18       ?value rdfs:label ?expressionValue .
19       ?probe atlasterms:dbXref ?dbXref .
20 }
21       ?pwElement dcterms:isPartOf ?pathway .
22       ?pathway dc:title ?pwTitle .
23       ?pathway dc:identifier ?wpURL .
24       ?pwElement wp:bdbEnsembl ?dbXref .
25 }
26 ORDER BY ASC(?pvalue)
```

Query clause

BGP

BGP

Query pattern

modifier

# Reasoning with ontologies

## **Handle synonyms** (from PubMed https://pubmed.ncbi.nlm.nih.gov/)

- Look for articles about “vitamin c” in full text search
- Look at the MeSH annotations
- Look for the MeSH term vitamin C and the articles it annotates
- Look for the MeSH term ascorbic acid and the articles it annotates

## **Handle taxonomy** (from the MeSH https://www.nlm.nih.gov/mesh/)

- Look for cardiovascular disease
- Select the relevant MeSH term (https://meshb.nlm.nih.gov/record/ui?ui=D002318)
- Look at its synonyms and its descendants
- Add it to the search builder
- Search on PubMed

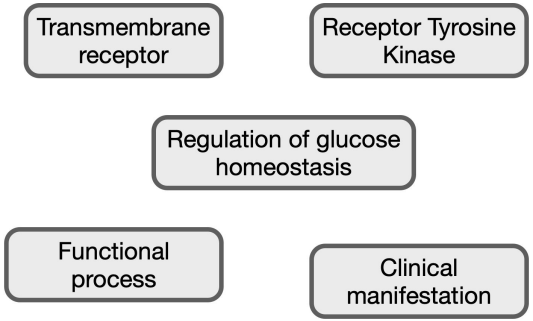
Synonyms and taxonomy are handled transparently

In the GO website (http://geneontology.org/)

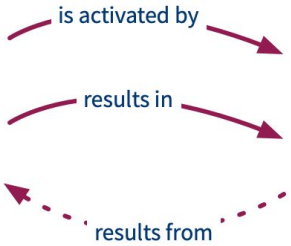
- Look for “glucose metabolic process”
- Select “ontology” in the radio box
- Select the relevant GO term (http://amigo.geneontology.org/amigo/term/GO:0006006)
- Select either the “graph view” or the “inferred tree view”
  - Visualise the GO term ancestors
  - Visualize the GO term descendants
- For Homo sapiens, how may proteins, miRNA, etc are annotated by this GO term (or one of its descendants)?

# Toy example

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.[5] Metabolically, the insulin receptor plays a key role in the regulation of glucose homeostasis, a functional process that under degenerate conditions may result in a range of clinical manifestations including diabetes and cancer.




How these concepts are related together ?

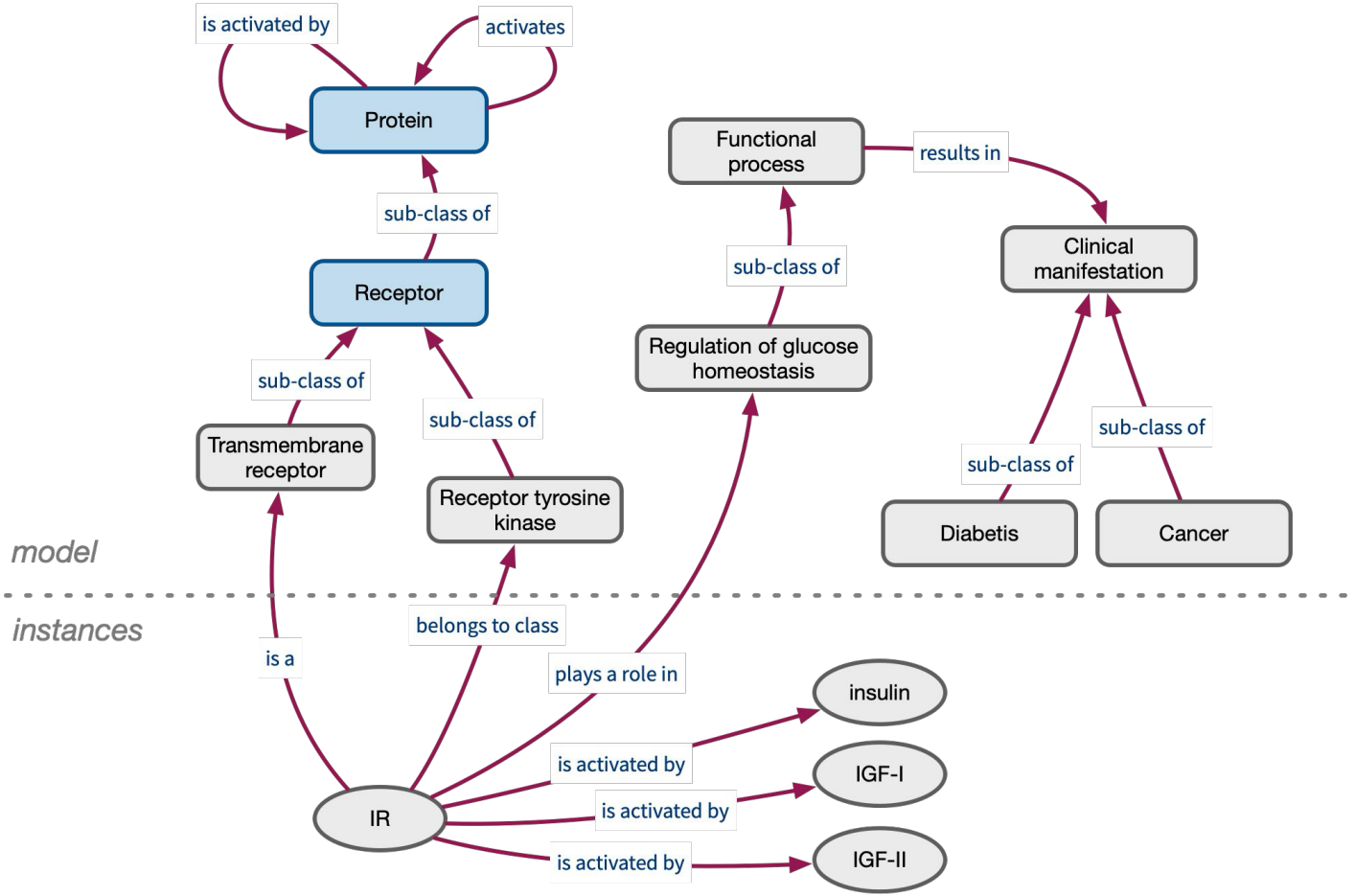



How these relations link concepts together ?

Do they allow deductions ?

https://en.wikipedia.org/wiki/Insulin_receptor

# Toy example


is activated by
activates
Protein
sub-class of
Receptor
sub-class of
sub-class of
Transmembrane receptor
Receptor tyrosine kinase
Functional process
results in
sub-class of
Regulation of glucose homeostasis
Clinical manifestation
sub-class of
sub-class of
Diabetis
Cancer
model
instances
is a
belongs to class
plays a role in
IR
insulin
IGF-I
IGF-II
is activated by
is activated by
is activated by

# RDF-S to represent simple ontologies

**RDF-Schema** aims at providing a simple vocabulary to **organize domain-specific knowledge** through classes (**concepts**) and properties (**relationships**).

## Class VS Instances

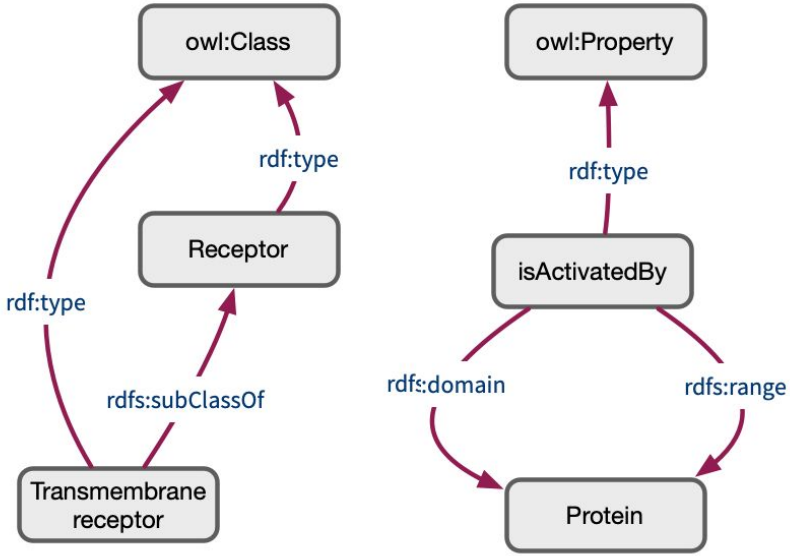
Resources may be classified into groups called **classes**. The members of a class are known as **instances** of the class. The rdf:**type** property is used to state that a resource is an instance of a class (« is a » relation ).

## Defining ontologies

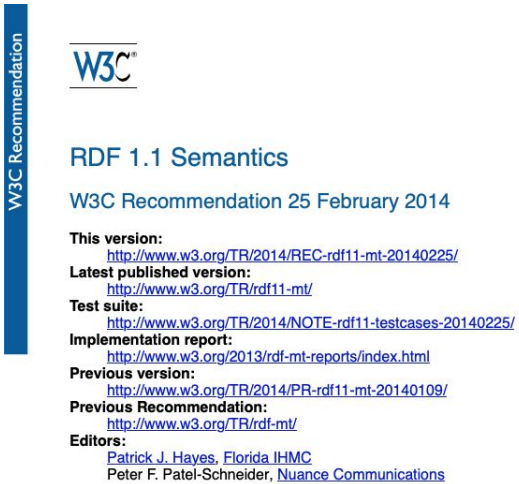
- rdf:**type**: to state that a resource is an instance of a class
- owl:**Class** & owl:**Property** to define specific classes or properties
- rdfs:**subClassOf**: to state that all the instances of one class are instances of another
- rdfs:**subPropertyOf**: to state that all resources related by one property are also related by another
- rdfs:**range**: a constraint on the class membership(s) for values of this property
- rdfs:**domain**: a constraint on the class membership(s) for resources having this property
- rdfs:**label**, rdfs:**comment**

# Sample RDF-S vocabulary

```
@prefix etbii: <http://our-namespace#> .
@prefix wikipedia: <https://en.wikipedia.org/wiki/>

etbii:TransmembraneReceptor rdf:type owl:Class ;
      rdfs:subClassOf etbii:Receptor ;
      rdfs:seeAlso wikipedia:Cell_surface_receptor .

etbii:Receptor rdf:type owl:Class ;
         rdfs:subClassOf etbii:Protein .

etbii:Protein rdf:type owl:Class .

etbii:isActivatedBy rdf:type owl:Property .
      rdfs:domain etbii:Protein ;
      rdfs:range etbii:Protein .
```



# Reasoning with RDFS-Entailments


W3C Recommendation

W3C

RDF 1.1 Semantics

W3C Recommendation 25 February 2014

**This version:**
http://www.w3.org/TR/2014/REC-rdf11-mt-20140225/
**Latest published version:**
http://www.w3.org/TR/rdf11-mt/
**Test suite:**
http://www.w3.org/TR/2014/NOTE-rdf11-testcases-20140225/
**Implementation report:**
http://www.w3.org/2013/rdf-mt-reports/index.html
**Previous version:**
http://www.w3.org/TR/2014/PR-rdf11-mt-20140109/
**Previous Recommendation:**
http://www.w3.org/TR/rdf-mt/
**Editors:**
Patrick J. Hayes, Florida IHMC
Peter F. Patel-Schneider, Nuance Communications

## Inference rules

**RDFS Core**

Derive new logical facts from existing triples, or verify logical consistency (satisfiability) of a knowledge graph.

## Type deduction via class hierarchies

**rdfs9/rdfs11**

If **TranscriptionFactor rdfs:subClassOf Protein**, any individual typed as **TranscriptionFactor** is also inferred as **Protein** — automatically.

## Type deduction via relations

**drfs2/rdfs3**

The **rdfs:domain** and **rdfs:range** of a property let the reasoner infer entity types from the relationships alone — no explicit annotations needed.

## Going further

OWL (Web Ontology Language) and Description Logics (DL) enable more expressive reasoning — cardinality constraints, class closure, inverse property inference, and more.

# Reasoning with RDFS-Entailments

## RDFS entailement rules (a selection)

| Rule | If S contains… | Then S entails… |
|---|---|---|
| rdfs2 | aaa rdfs:domain xxx<br>yyy aaa zzz | yyy rdf:type xxx |
| rdfs3 | aaa rdfs:range xxx<br>yyy aaa zzz | zzz rdf:type xxx |
| rdfs9 | xxx rdfs:subClassOf yyy<br>zzz rdf:type xxx | zzz rdf:type yyy |
| rdfs11 | xxx rdfs:subClassOf yyy<br>yyy rdfs:subClassOf zzz | xxx rdfs:subClassOf zzz |

*rdfs2 / rdfs3 highlighted — applied to the insulin/IR example below*

```
Schema:
isActivatedBy  rdfs:domain Protein.
isActivatedBy  rdfs:range Protein.

ASSERTED:
IR  isActivatedBy  insulin

INFERRED:
IR  rdf:type  Protein                    rdfs2
insulin  rdf:type  Protein               rdfs3
```

## Rule rdfs2 - step by step

**IF** — Property X declares a domain class Y

`X rdfs:domain Y`

*e.g. isActivatedBy rdfs:domain Protein*

**AND** — Individual a is linked to b via X

`a X b`

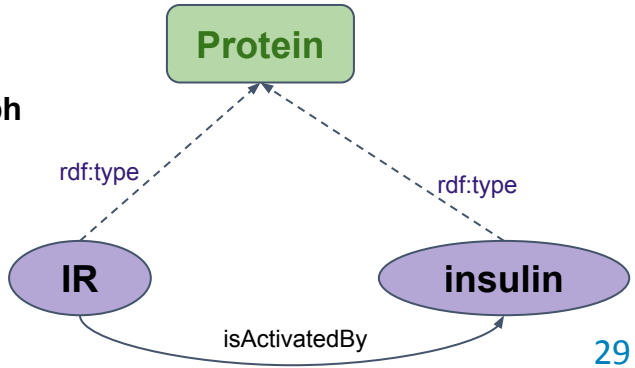
*e.g. IR isActivatedBy insulin*

**THEN** — a is inferred to be of type Y

`a rdf:type Y [new fact]`

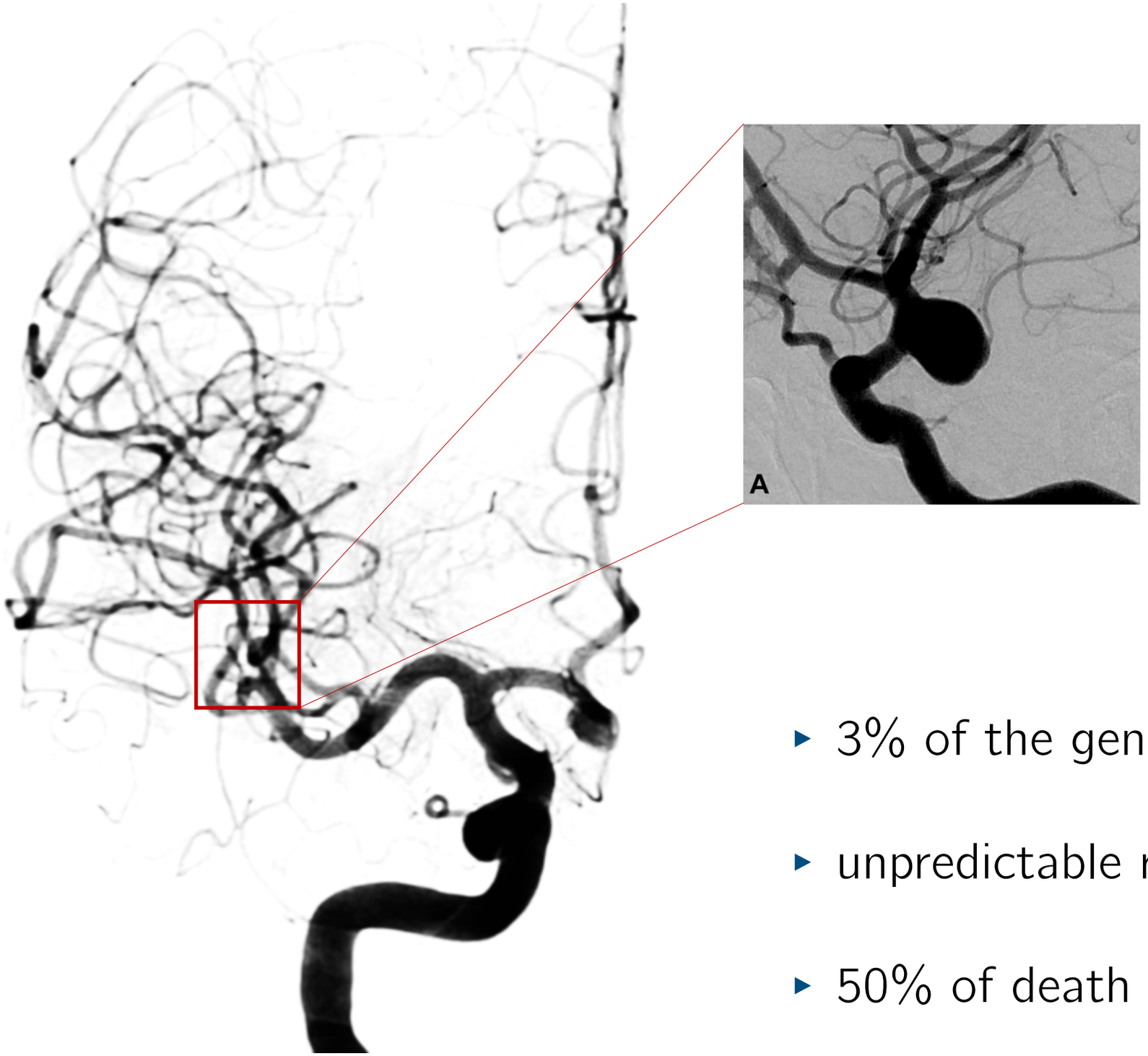
*e.g. IR rdf:type Protein*

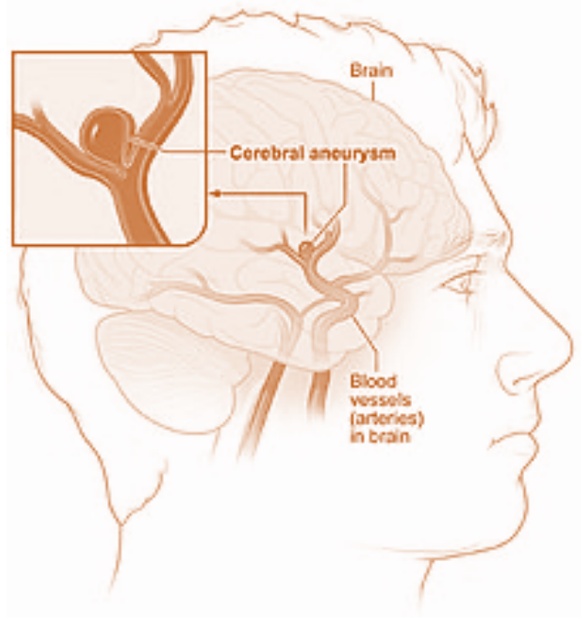
## Resulting RDF graph



# Biomedical application

# Intracranial aneurysms

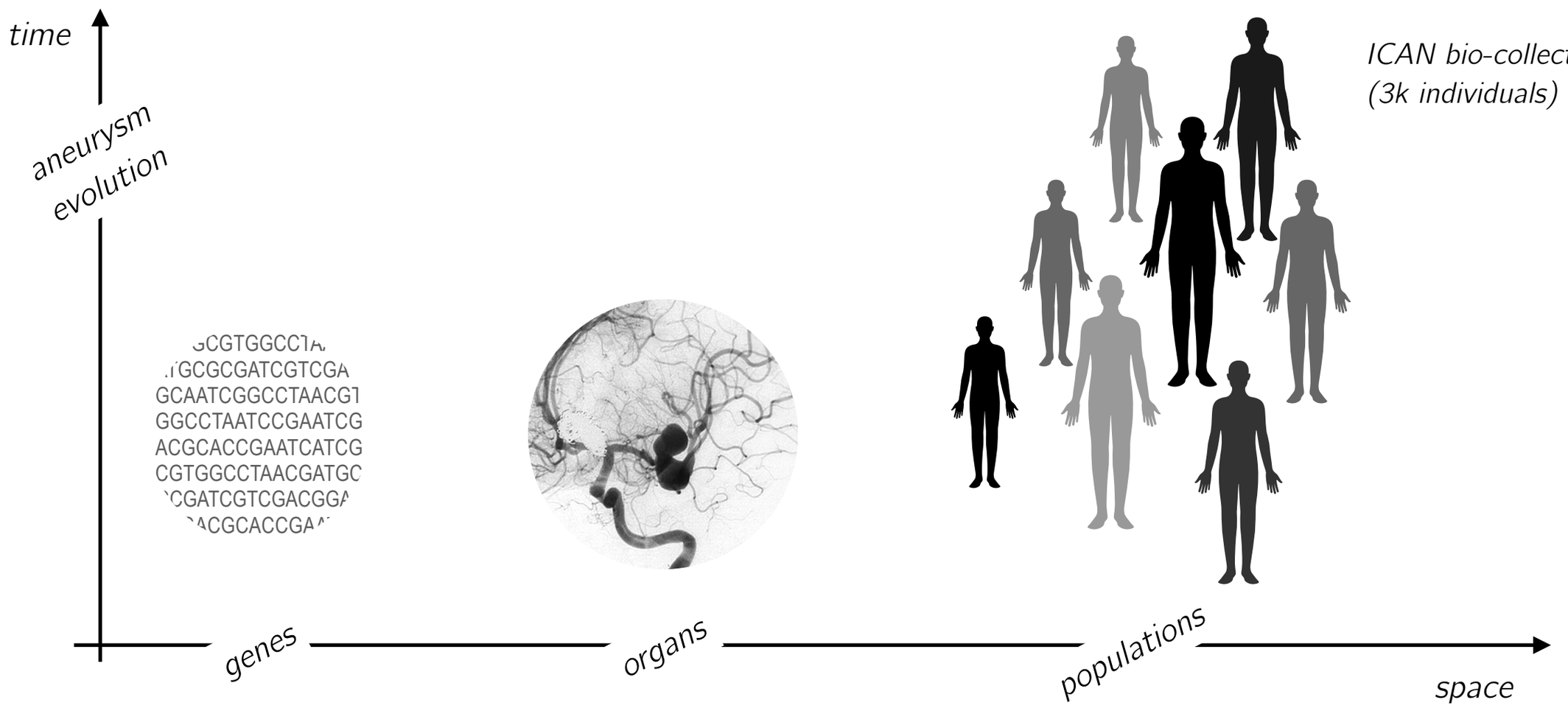



- 3% of the general population
- unpredictable rupture
- 50% of death in case of rupture

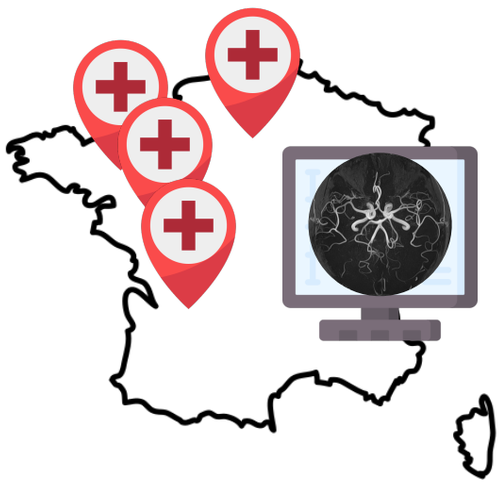
# Multi-factorial disease $\longrightarrow$ multi-scale data

Inter-disciplinary efforts needed for a better understanding of the pathology

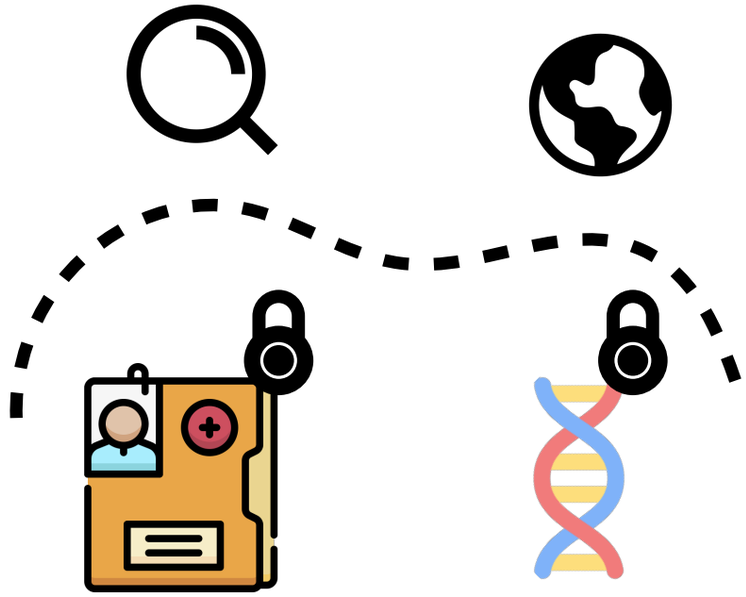


# Data integration & sharing challenges

❶ How to **collect high-quality medical images** from multiple hospitals/ MRIs ?



❷ How to **interlink and query multi-modal and multi-scale data** while preserving privacy constraints ?



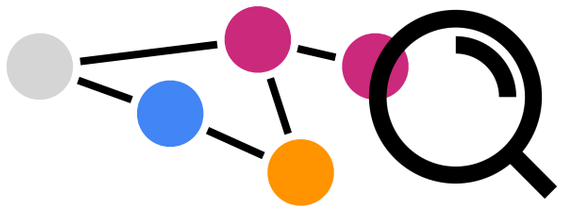
❸ How to mine and **model patient trajectories** from EHR data ? can we predict clinical outcomes ?

FRANCE 2030 PROGRAMME DE RECHERCHE SANTÉ NUMÉRIQUE

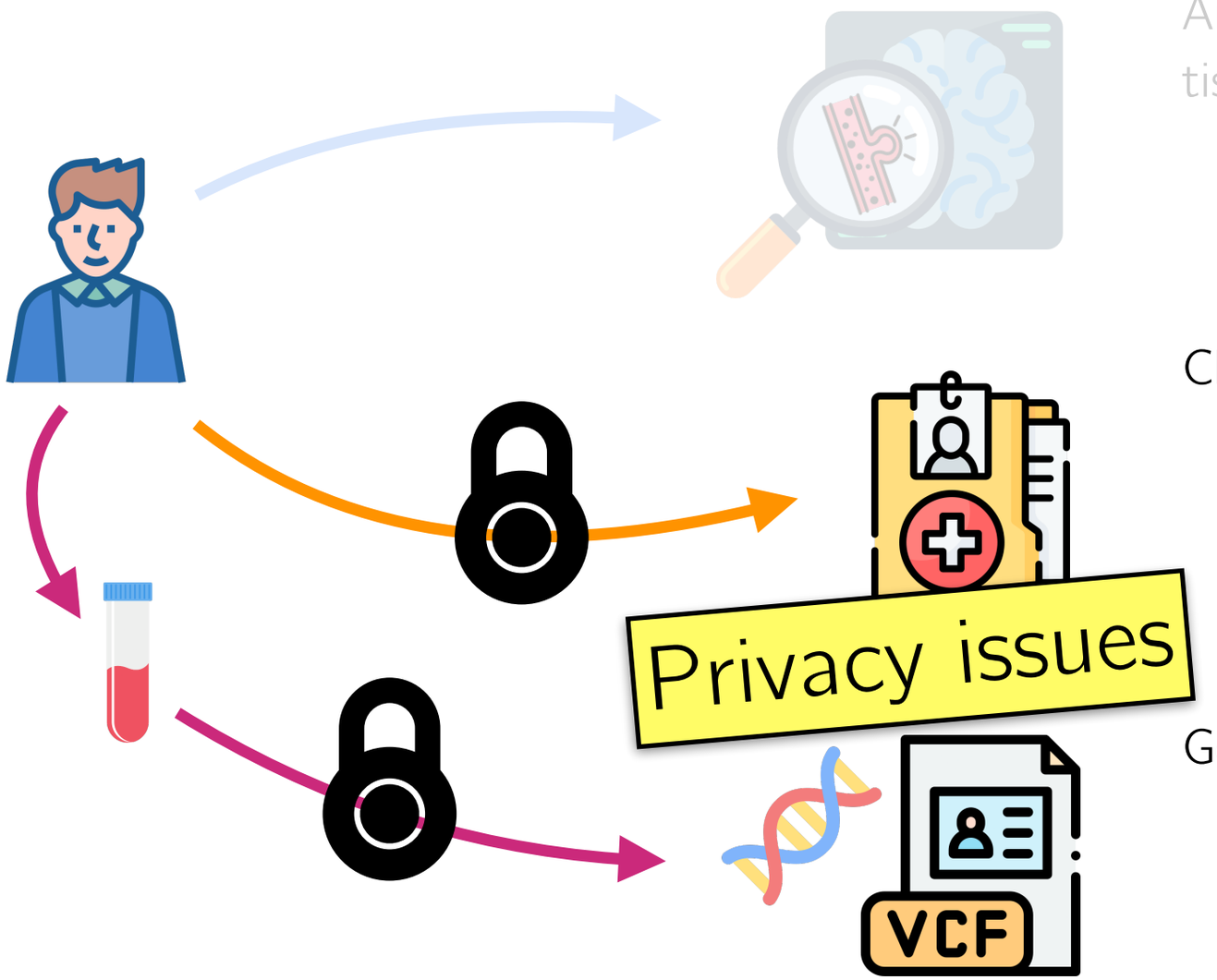
NEUROVASC PEPR SANTÉ NUMÉRIQUE

# Knowledge graphs for clinical & genomic data

A clinical and genomic intracranial aneurysm knowledge graph



to find & exchange phenotypes/variants with reference terminologies !




Anatomical structures ? Neuro-vascular tissues ?

- UBERON
- NCIT

Clinical data / phenotypes ?

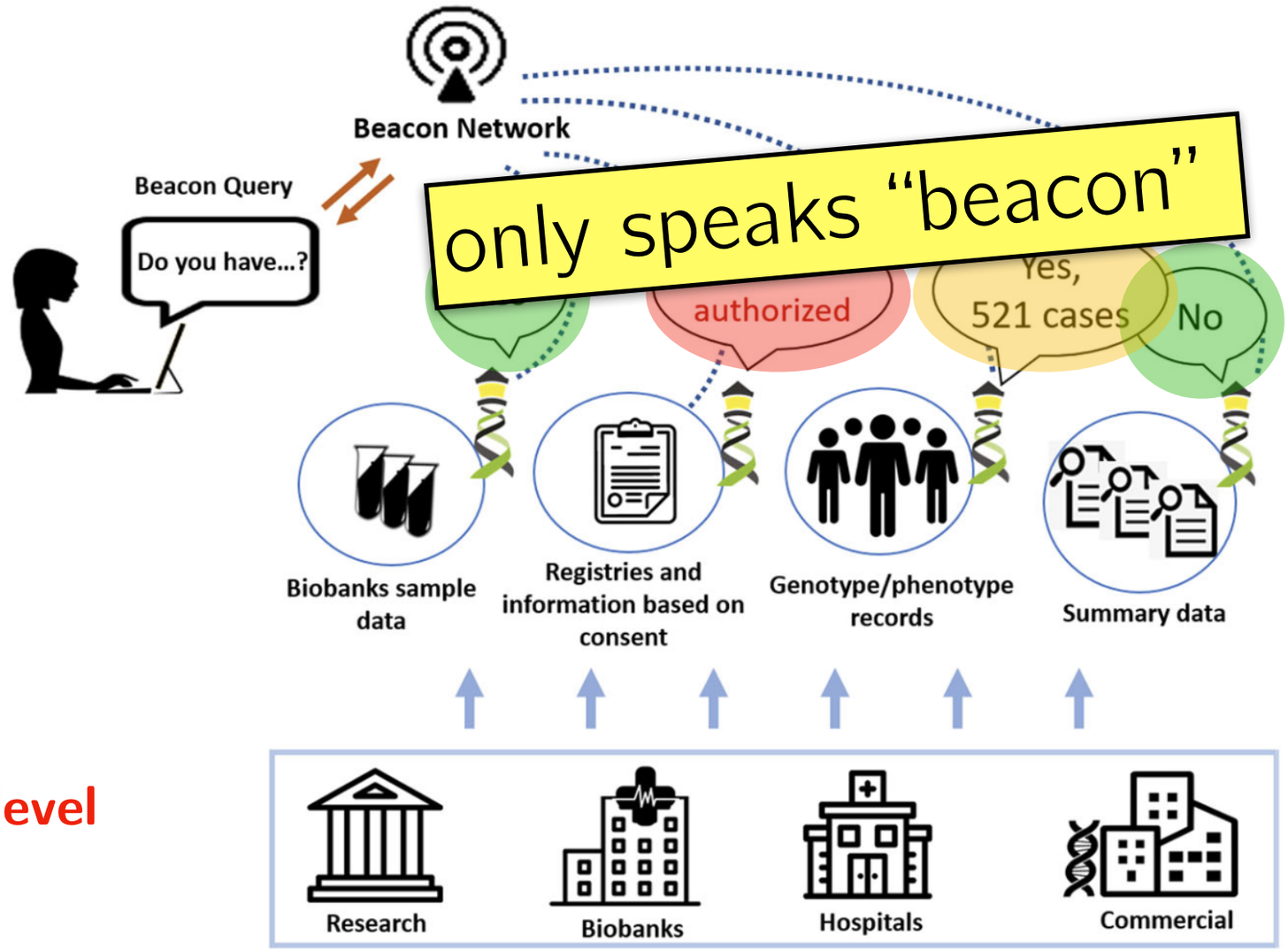
- **SPHN**
- **HPO**
- **DUO**

Genomic data ?

- **FALDO**
- **SO / GENO**
- **SIO**

# Beacon protocol

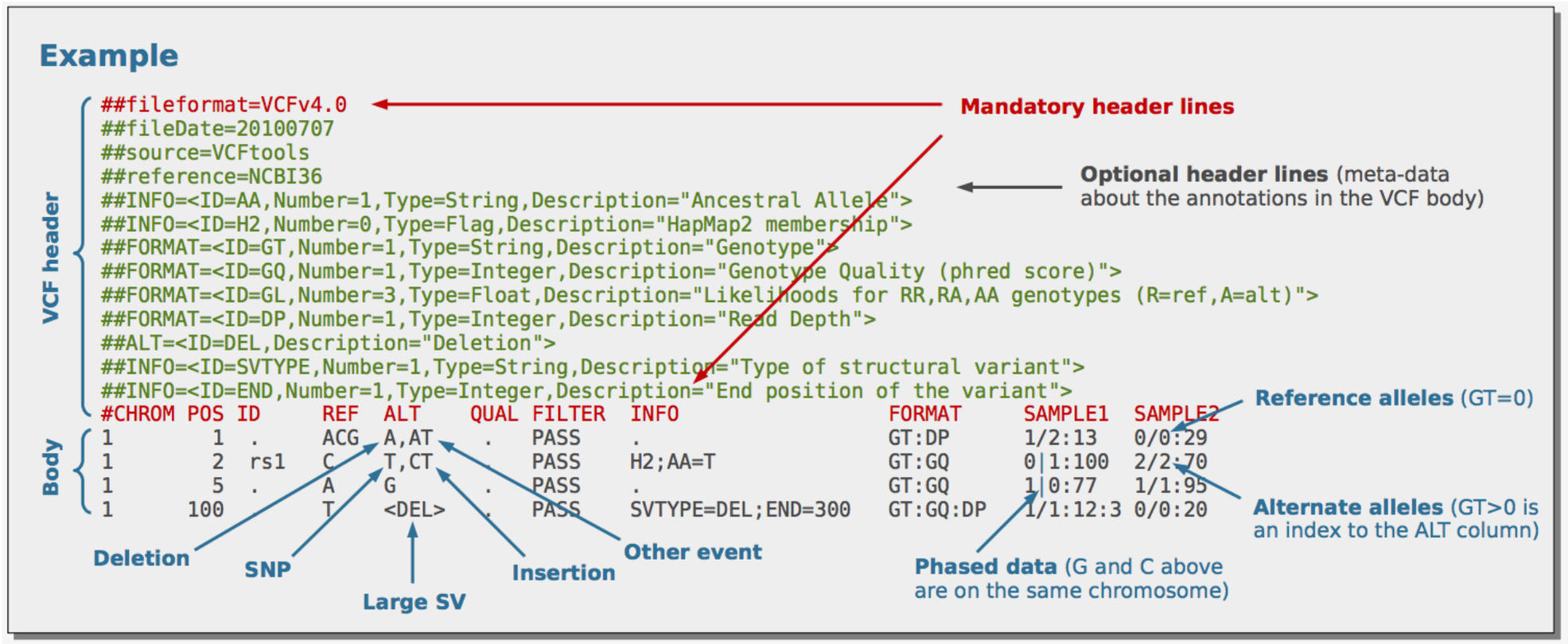
Beacon: a standard and exchange protocol for more decentralized biomedical research (promoted by Elixir and GA4GH)




- **Metadata model** for ‘Variation’, ‘Sample’, ‘Dataset’, ‘Individual’, etc
- Different **access models**: **boolean**, **aggregated data**, **record level**
- A framework with a **reference implementation**

Rambla, J., Baudis, M., Ariosa, R., Beck, T., Fromont, L. A., Navarro, A., Paloots, R., Rueda, M., Saunders, G., Singh, B., Spalding, J. D., Törnroos, J., Vasallo, C., Veal, C. D., & Brookes, A. J. (2022). **Beacon v2 and Beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond**. *Human mutation*, *43*(6), 791–799. https://doi.org/10.1002/humu.24369

# Genomic variation data & “annotation”

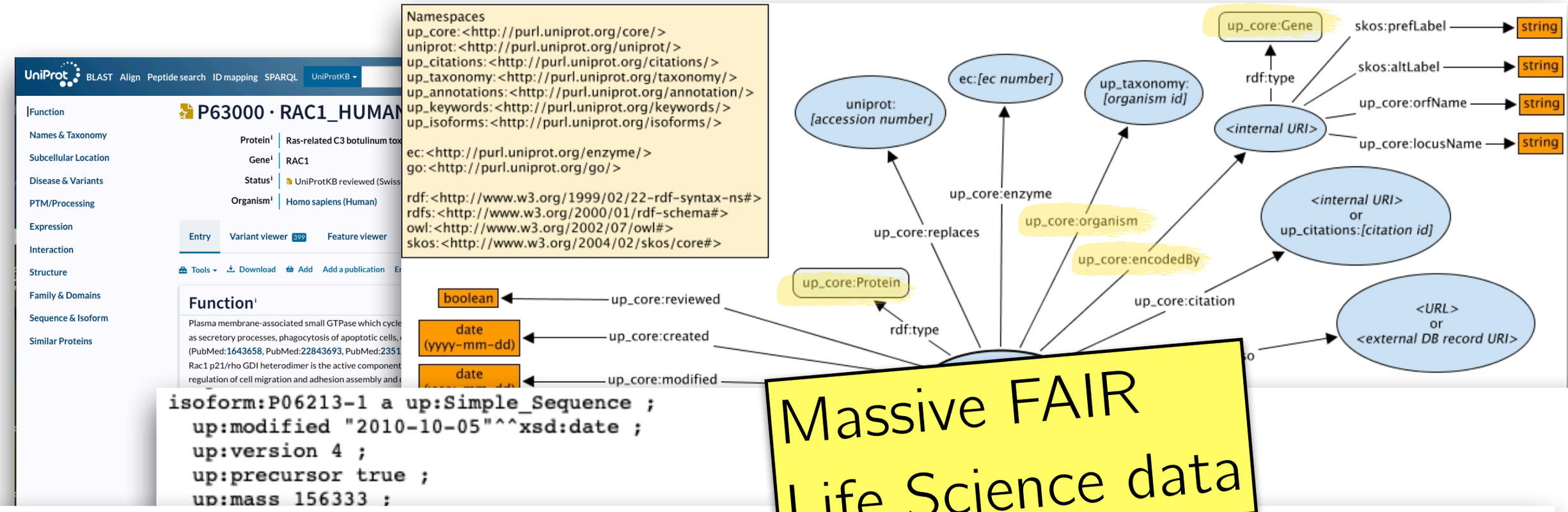



- Large tabular file: 1 line per genomic variation, 1 column per individual
- Specific columns for **locating** the variation in the **genome**
- INFO column for **annotations coming from external databases:** *e.g.* pathogenicity scores (CADD v1.7: whole genome annotation database, 625G)

Compute and storage intensive variant annotation

https://cadd.gs.washington.edu/download

# The UniProt public knowledge graph




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

| Graph | Documentation | Triples | Distinct subjects | Distinct predicates | Distinct classes | Distinct objects | License |
|---|---|---|---|---|---|---|---|
| uniparc | Documentation | 160,189,731,200 | 40,455,837,024 | 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/ |

# Issues & Objectives

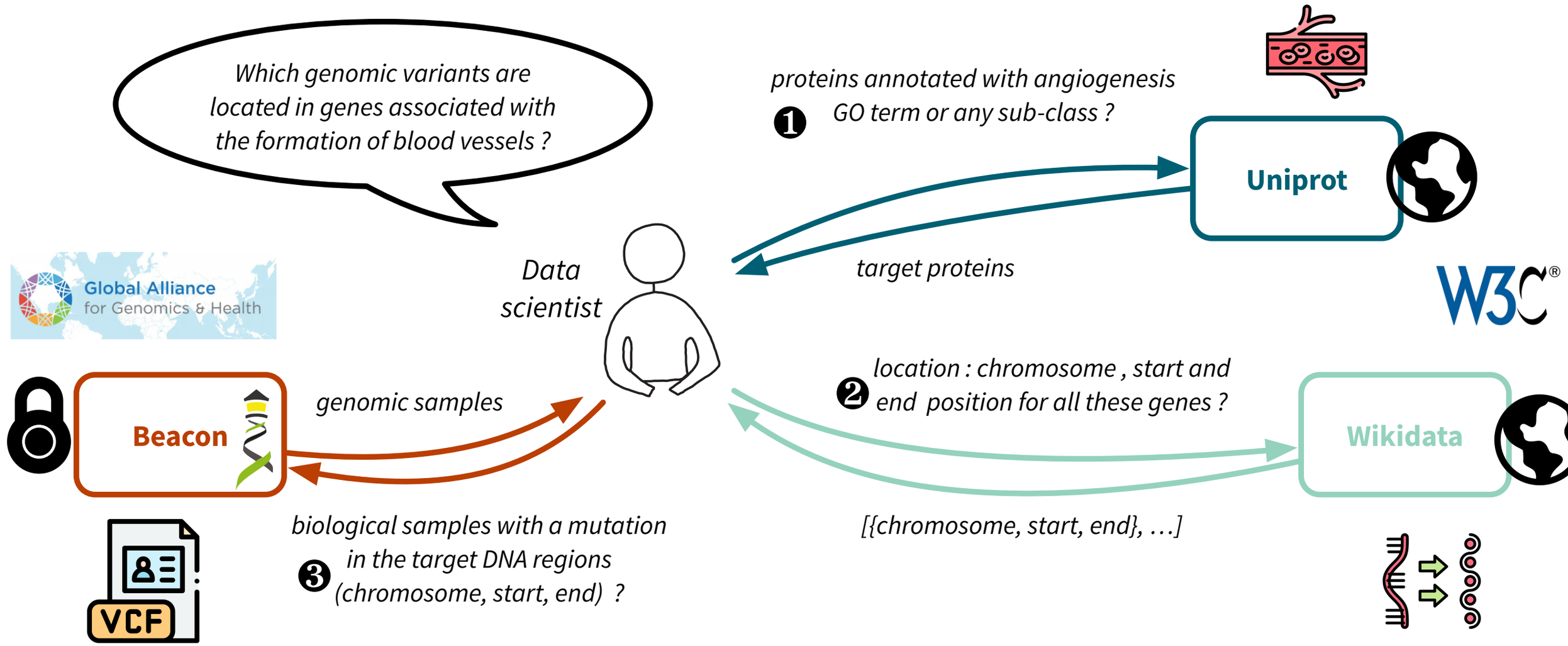
⚠️ **Genomic variants** must be safely kept **on-site**.

⚠️ **Annotating genomic variants** for biological interpretation is **costly** (data transfer + CPU).

🚀 Massive and diverse reference data already available in the form of **interoperable public knowledge graphs**.

How to enable **on-the-fly annotation** of genomic beacon data with public knowledge graphs ?

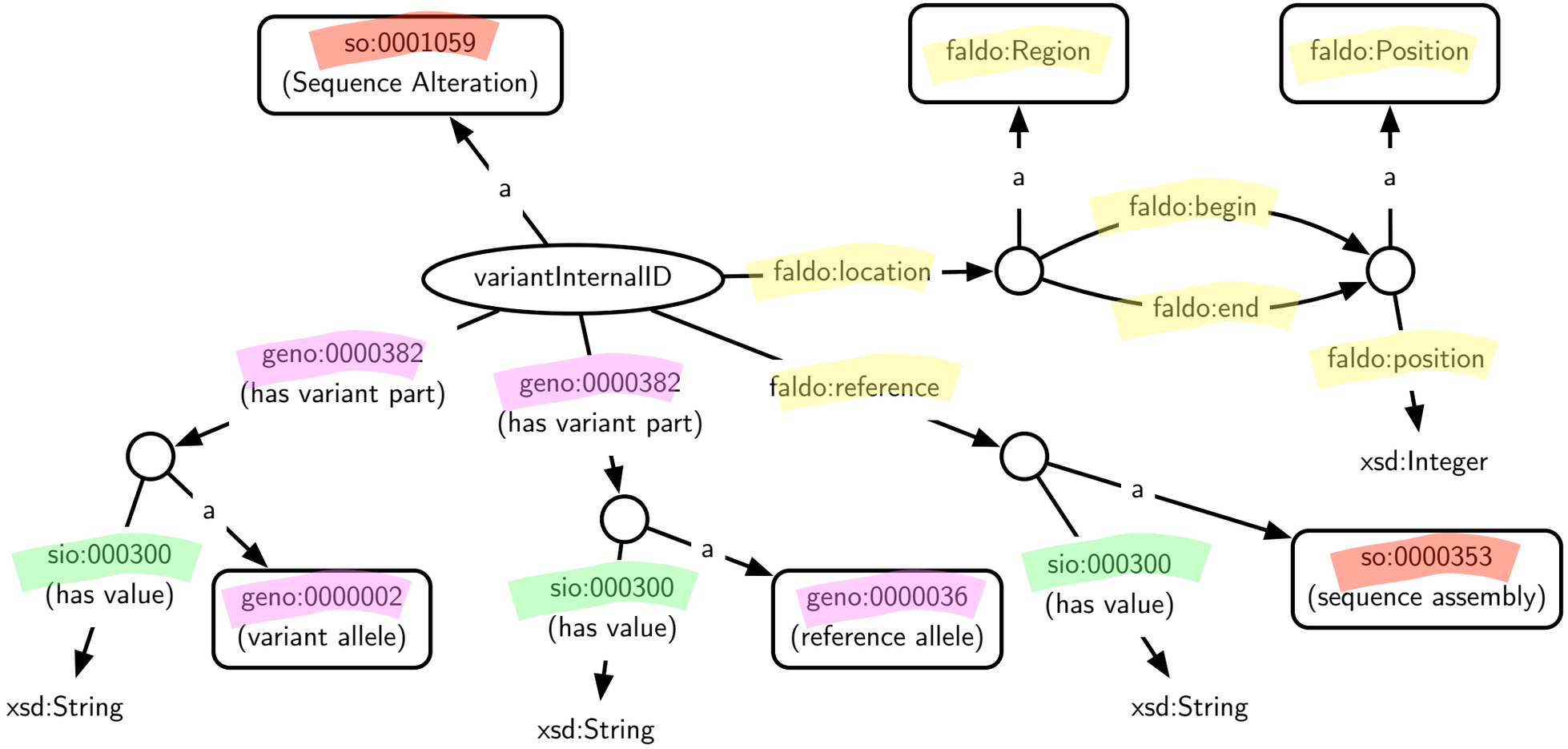
# Intracranial aneurysm motivating use case



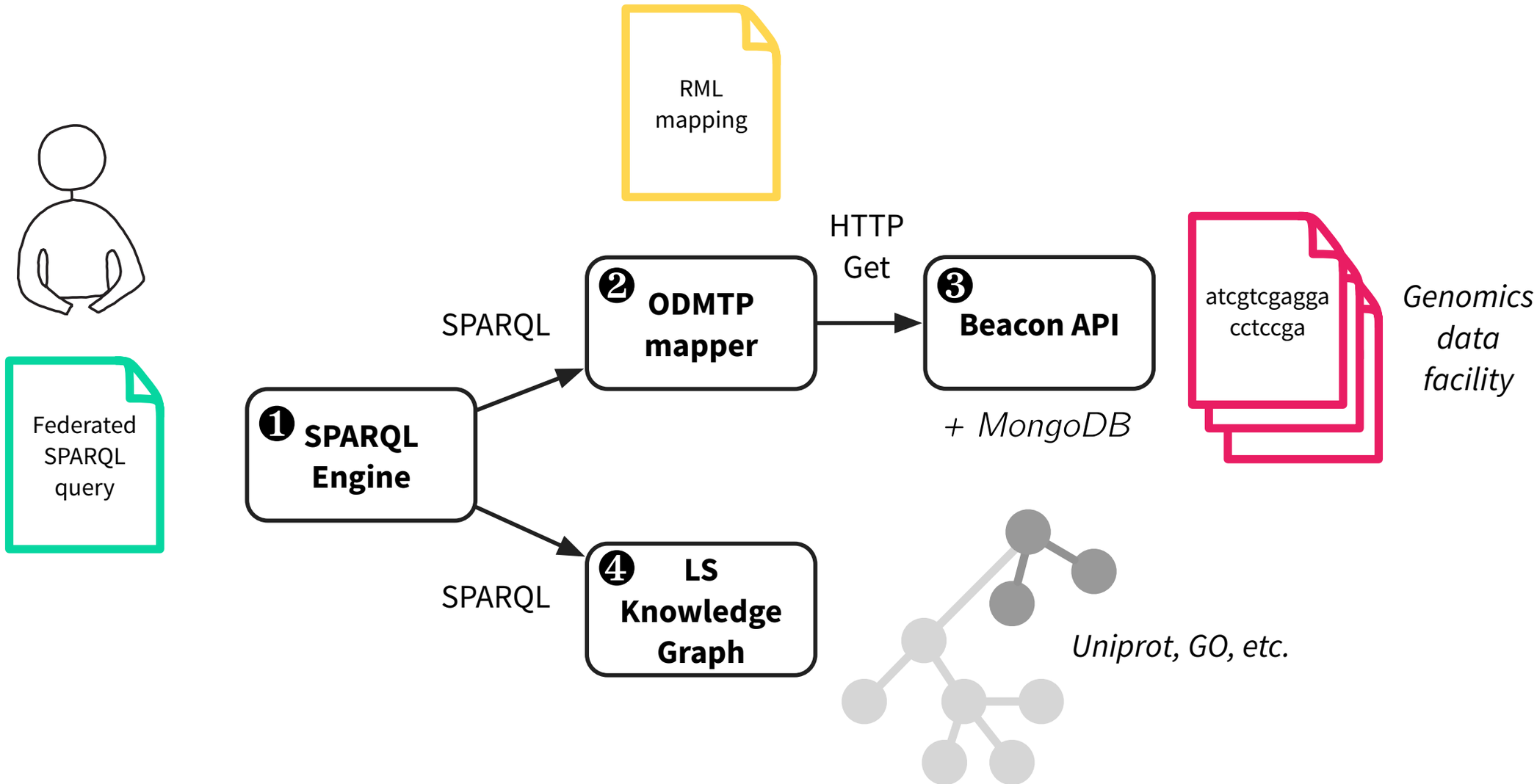
# Semantic representation of genomic variants

No need to create a new ontology ...

- SO: Sequence Ontology
- GENO: Genotype Ontology
- FALDO: Feature Annotation Location Description Ontology
- SIO: Semantic science Integrated Ontology



# Architecture for “Semantic” Beacons




Moreau, B., Serrano-Alvarado, P., Desmontils, E., & Thoumas, D. (2017, October). Querying non-RDF Datasets using Triple Patterns. In *16th International Semantic Web Conference (ISWC2017)*.

# Federated SPARQL query

```
SELECT * WHERE {
  SERVICE <https://sparql.uniprot.org/sparql> {
    ?protein a up:Protein ;
        up:organism taxon:9606 ;
        up:classifiedWith ?goTerm .
    ?goTerm rdfs:subClassOf* GO:0001525 .
  }
  BIND(SUBSTR(STR(?protein), STRLEN(STR(up:)) + 4) AS ?proteinID2)
.

  SERVICE <https://query.wikidata.org/sparql> {
    ?wp wdt:P352 ?proteinID2 ;
        wdt:P702 ?wg .
    ?wg wdp:P644 ?wgss ;
        wdp:P645 ?wgse .
    ?wgss wdps:P644 ?startcoordinate ;
        wdpq:P1057/wdt:P1813 ?chromosome ;
        wdpq:P659/rdfs:label ?assembly .
    ?wgse wdps:P645 ?endcoordinate ;
        wdpq:P1057/wdt:P1813 ?chromosome ;
        wdpq:P659/rdfs:label ?assembly .
    FILTER(lang(?assembly) = "en")
    FILTER(STR(?assembly) = "genome assembly GRCh38")
  }

  ?variant a so:0001059 ;
    faldo:reference/sio:SIO_000300 ?chromosome ;
    faldo:location/faldo:begin/faldo:position ?v_start ;
    faldo:location/faldo:end/faldo:position ?v_end .

  FILTER( (((?v_start >= xsd:integer(?startcoordinate)) &&
            (?v_start <= xsd:integer(?endcoordinate)) ))
        || ((?v_end >= xsd:integer(?startcoordinate)) &&
            (?v_end <= xsd:integer(?endcoordinate))) )
}
LIMIT 10
```

SERVICE clauses for each remote data source

All human proteins (`taxon:9606`) classified with all sub-classes of "angiogenesis" (`GO:0001525`)

For the matching proteins, get the location of the encoding genes for the GRCh38 reference human genome

All local genomic variants matching the localization constraints

# Take-home message

- Beacon is great for **privacy**-preserving **genomic data discovery**
- However, it has a **limited interoperability** with public knowledge graphs such as Uniprot
- Many ontologies are available to represent **genomic data as knowledge graphs**
- This approach preserves **decentralization and data source autonomy** through federated SPARQL queries.

- **Future works** include
  - addressing **scalability issues** (costly aggregate queries, non-selective queries on remote sources, distributed joins ...)
  - addressing **security issues in knowledge graph federations** → SAFE-KG ANR project
    - safe federated query formulation (LLM)
    - safe and efficient federated query execution
    - decentralized access and usage policies, traceability and explainability

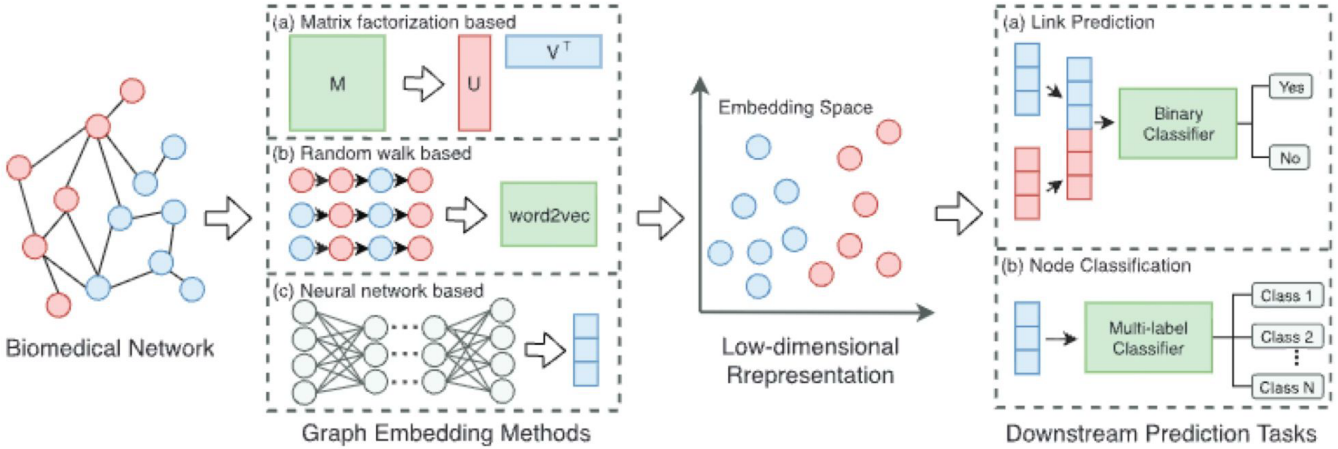
# KG Application in ML

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir FRANCE

cnrs INRAE Inserm cea

# Graph embeddings approach



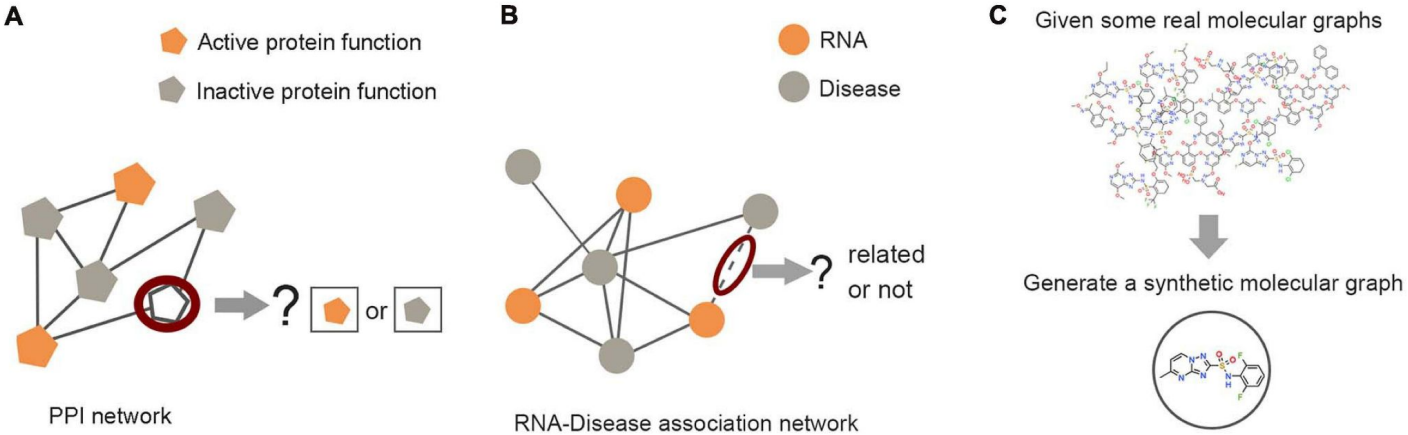

https://doi.org/10.1093/bioinformatics/btz718

**Some Tasks**

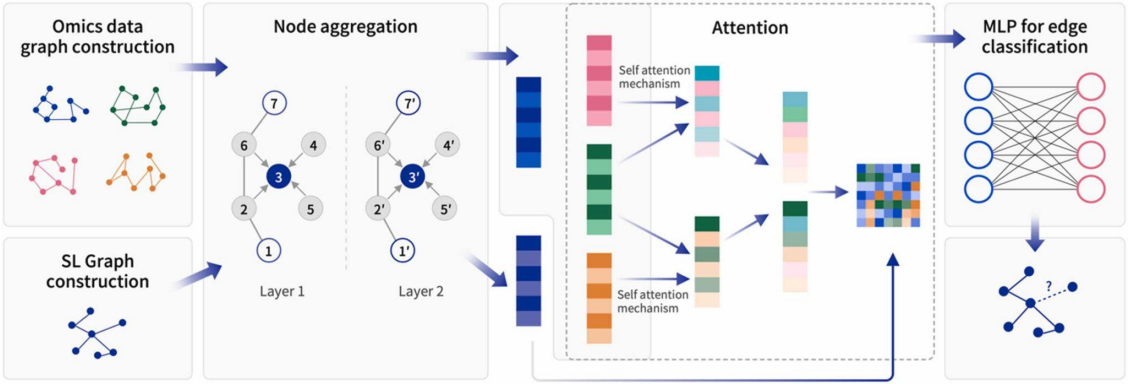
Disease-Gene association
Protein Function prediction
PPI prediction
Drug-Target (Drug) Interaction
Drug response …

Zhang X-M, Liang L, Liu L and Tang M-J (2021) Graph Neural Networks and Their Current Applications in Bioinformatics. Front. Genet. 12:690049. doi: 10.3389/fgene.2021.690049

# Node classification / Link prediction / Synth. data generation

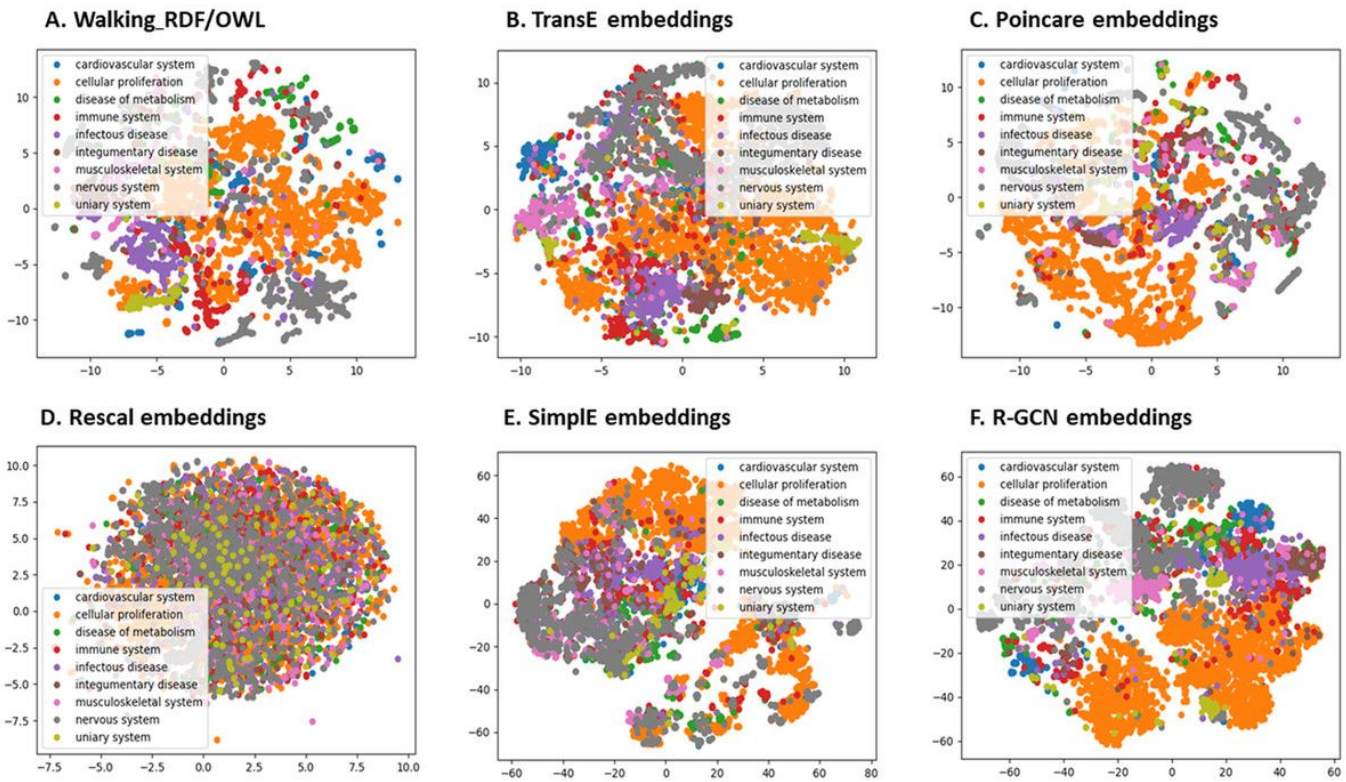



Zhang X-M, Liang L, Liu L and Tang M-J (2021) Graph Neural Networks and Their Current Applications in Bioinformatics. *Front. Genet.* 12:690049. https://doi.org/10.3389/fgene.2021.690049




Using graph-based model to identify cell specific synthetic lethal effects
https://doi.org/10.1016/j.csbj.2023.10.011

# Embeddings methods on knowledge graphs give different results




**KG embeddings models**

Uniprot Embeddings

RDF2VEC

DistMult

MultiKE

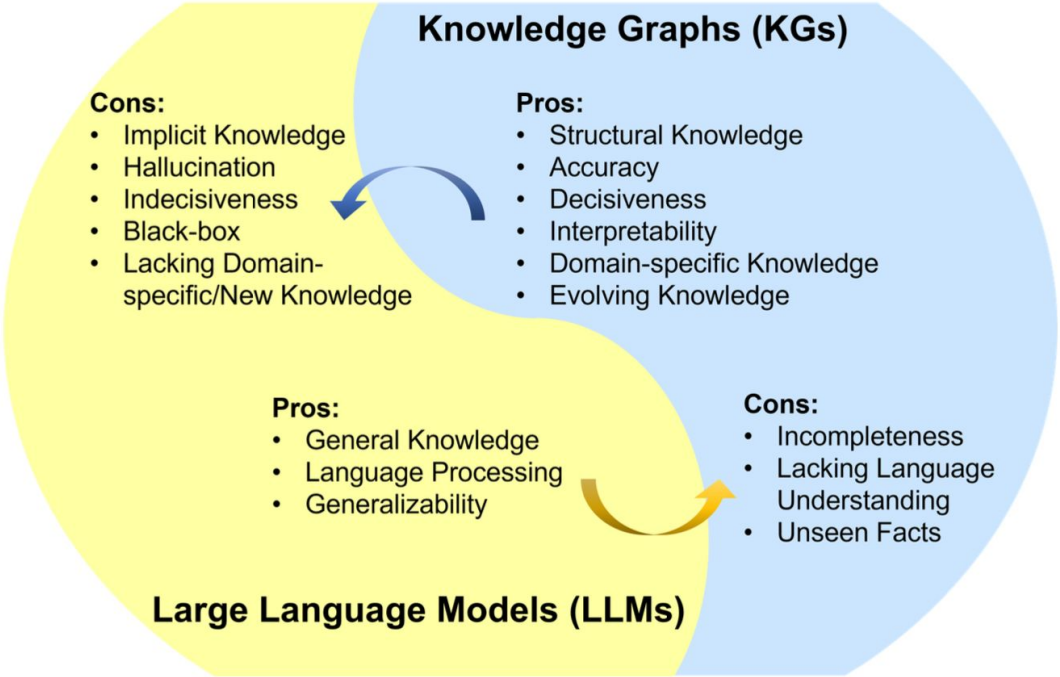
ComplEx

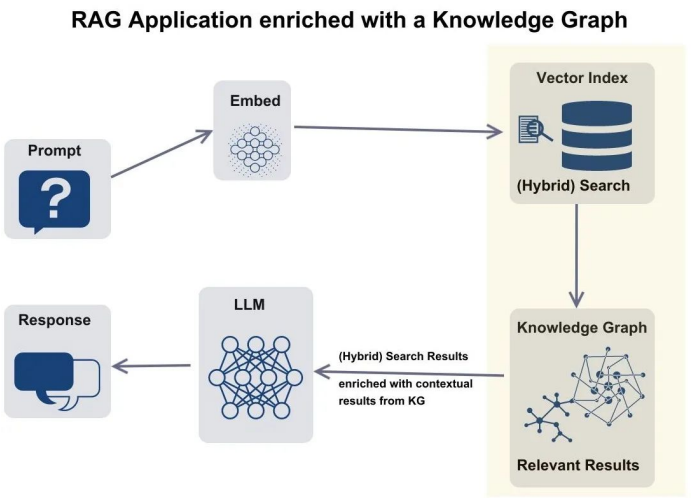
RDGCN

i-Align

Alshahrani M, Thafar MA, Essack M. 2021. Application and evaluation of knowledge graph embeddings in biomedical data. *PeerJ Computer Science* 7:e341 https://doi.org/10.7717/peerj-cs.341

# LLMs vs KGs pros and cons



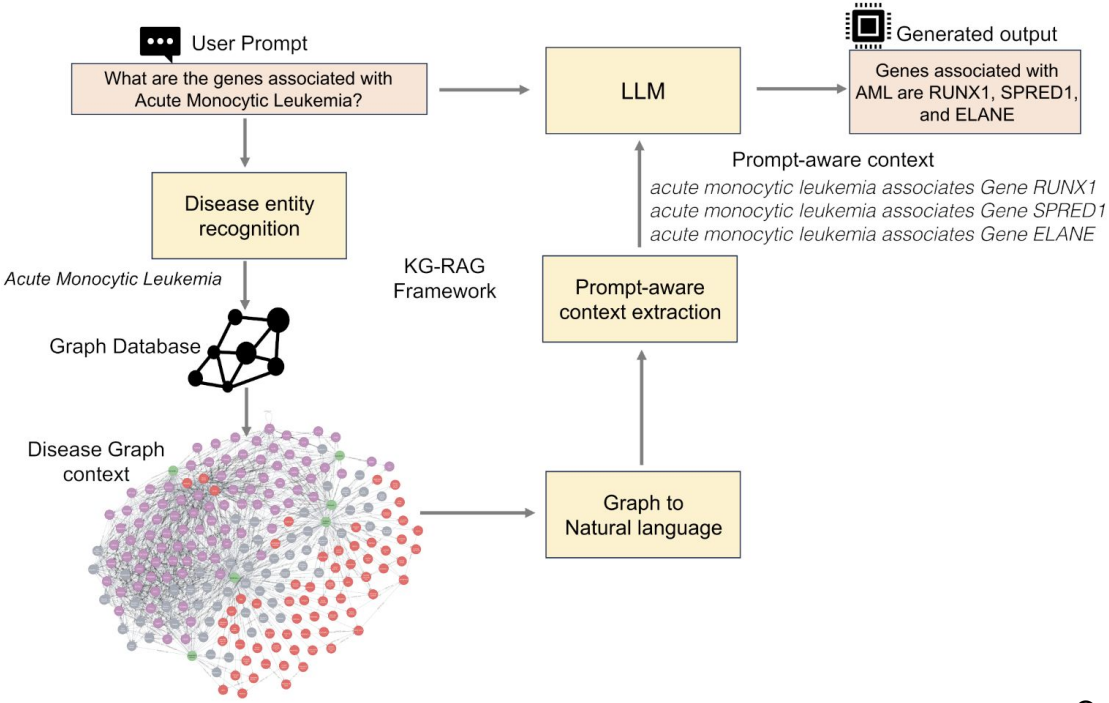

Retrieval Augmented Generation and Knowledge Graphs

https://gradientflow.com/boosting-llms-with-external-knowledge-the-case-for-knowledge-graphs/

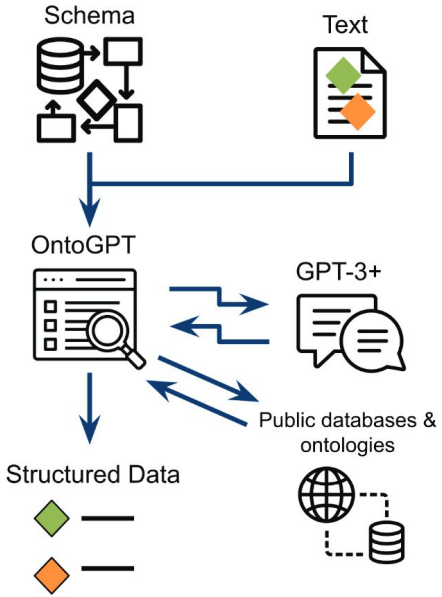
**=> Boosting LLMs with External Knowledge: The Case for Knowledge Graphs**

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2023). Unifying Large Language Models and Knowledge Graphs: A Roadmap. *ArXiv, abs/2306.08302*.

# Some Life Science application of LLMs + KG




Soman et al. https://doi.org/10.1093/bioinformatics/btae560




Overview of the SPIRES approach.
SPIRES is available as part of the open source OntoGPT package: https://github.com/monarch-initiative/ontogpt

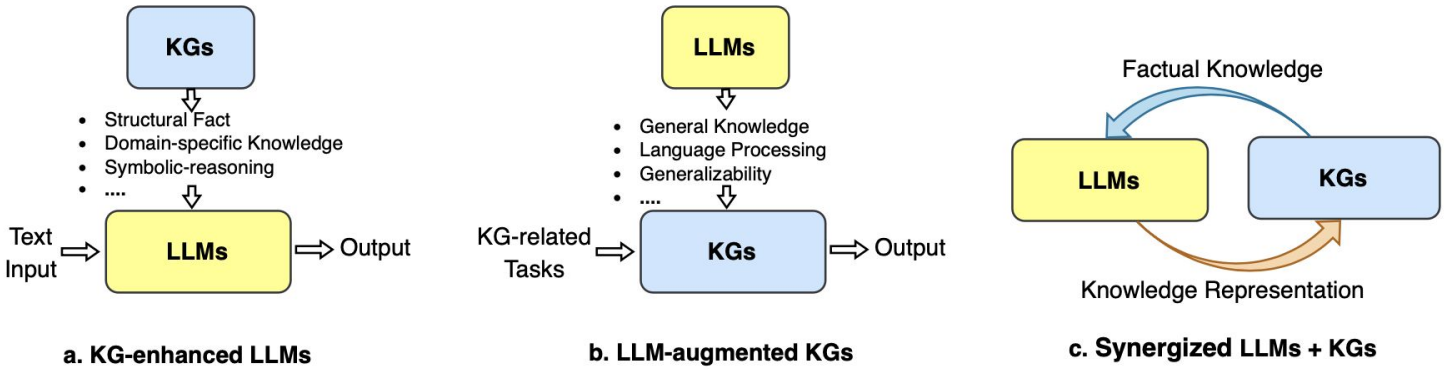
# Toward combining KGs and LLMs

KG to LLM

- Document / explain generated text with “facts” coming from knowledge graphs : e.g. Explicability https://arxiv.org/pdf/2309.01029.pdf https://arxiv.org/abs/2311.09188
- Explain data with complex structure : e.g. map data with ontologies, ontology alignment
- Answer domain-specific questions : e.g. guided fine-tuning

LLM to KG

- Natural language interface to generate SPARQL queries : e.g. SPARQL Generation: an analysis on fine-tuning OpenLLaMA for Question Answering over a Life Science Knowledge Graph. Ana Claudia Sima et al. SWAT4HCLS 2024 *https://arxiv.org/abs/2402.04627*
- Augment knowledge with synthetic data for better prediction in graph embedding approaches e.g. https://arxiv.org/abs/2203.13965




Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2023). Unifying Large Language Models and Knowledge Graphs: A Roadmap. *ArXiv, abs/2306.08302*.

**Phenomics Assistant: An Interface for LLM-based Biomedical Knowledge Graph Exploration**

Shawn T O'Neil, Kevin Schaper, Glass Elsarboukh, Justin T Reese, Sierra A T Moxon, Nomi L Harris, Monica C Munoz-Torres, Peter N Robinson, Melissa A Haendel, Christopher J Mungall

**doi:** https://doi.org/10.1101/2024.01.31.578275

**BioMedGPT: Open Multimodal Generative Pre-trained Transformer for BioMedicine**

Yizhen Luo et al. https://arxiv.org/abs/2308.09442

*Vol. 06, No. 02, pp. 342 –357 (2025)*
**ISSN: 2708-0757**

JASTT
JOURNAL OF APPLIED SCIENCE AND TECHNOLOGY TRENDS
www.jastt.org

A Hybrid LLM–Knowledge Graph Framework for Accurate Biomedical Question Answering

Havraz Y. Omar[1,2]*, Abdulhakeem O. Mohammed[3]

https://jastt.org/index.php/jasttpath/article/download/404/107/2294

**Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES): a method for populating knowledge bases using zero-shot learning**

J Harry Caufield, Harshad Hegde, Vincent Emonet, Nomi L Harris, Marcin P Joachimiak, Nicolas Matentzoglu, HyeongSik Kim, Sierra Moxon, Justin T Reese, Melissa A Haendel ... Show more

*Bioinformatics*, Volume 40, Issue 3, March 2024, btae104,
https://doi.org/10.1093/bioinformatics/btae104

**Dynamic Retrieval Augmented Generation of Ontologies using Artificial Intelligence (DRAGON-AI)**

Sabrina Toro *et al*

**https://arxiv.org/abs/2312.10904**

**Biomedical knowledge graph-enhanced prompt generation for large language models**

Soman et al. https://arxiv.org/pdf/2311.17330.pdf

# Multi-Agent Systems for scientific discovery

## KG4Diagnosis: A Hierarchical Multi-Agent LLM Framework with Knowledge Graph Enhancement for Medical Diagnosis

Kaiwen Zuo, Yirui Jiang, Fan Mo, Pietro Lio

https://doi.org/10.48550/arXiv.2412.16833

## KGARevion: An AI Agent for Knowledge-Intensive Biomedical QA

Xiaorui Su, Yibo Wang, Shanghua Gao, Xiaolong Liu, Valentina Giunchiglia, Djork-Arné Clevert, Marinka Zitnik

## CellAgent: An LLM-driven Multi-Agent Framework for Automated Single-cell Data Analysis

Yihang Xiao, Jinyi Liu, Yan Zheng, Xiaohan Xie, Jianye Hao, Mingzhi Li, Ruitao Wang, Fei Ni, Yuxiao Li, Jintian Luo, Shaoqing Jiao, Jiajie Peng

## VarChat: the generative AI assistant for the interpretation of human genomic variations

Federica De Paoli, Silvia Berardelli, Ivan Limongelli, Ettore Rizzo, Susanna Zucca ✉

Author Notes

*Bioinformatics*, Volume 40, Issue 4, April 2024, btae183,
https://doi.org/10.1093/bioinformatics/btae183

## MEGA-GPT: Artificial Intelligence Guidance and Building Analytical Protocols Using MEGA Software

John B Allard, Sudhir Kumar ✉ Author Notes

*Molecular Biology and Evolution*, Volume 42, Issue 6, June 2025, msaf101,
https://doi.org/10.1093/molbev/msaf101

## Autonomous chemical research with large language models

Daniil A. Boiko, Robert MacKnight, Ben Kline & Gabe Gomes ✉

*Nature* **624**, 570–578 (2023) | Cite this article

## Artificial Intelligence agents for biological research: a survey

Cong Qi, Wenbo Wang, Siqi Jiang, Qin Liu, Xun Song, Hanzhang Fang, Zhi Wei ✉

Author Notes

*Briefings in Bioinformatics*, Volume 27, Issue 1, January 2026, bbag075,
https://doi.org/10.1093/bib/bbag075

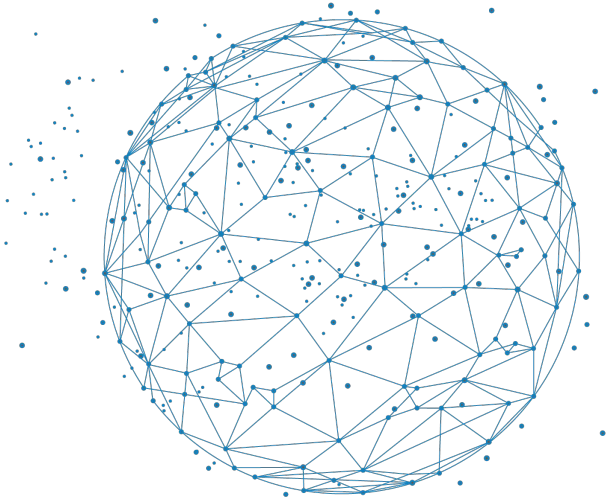
## MRAgent: an LLM-based automated agent for causal knowledge discovery in disease via Mendelian randomization

Wei Xu, Gang Luo, Weiyu Meng, Xiaobing Zhai, Keli Zheng, Ji Wu, Yanrong Li, Abao Xing, Junrong Li, Zhifan Li ... Show more

Author Notes

*Briefings in Bioinformatics*, Volume 26, Issue 2, March 2025, bbaf140,
https://doi.org/10.1093/bib/bbaf140

# Keep in mind



*Complex data analyses require fine-grained, explicit descriptions*

- Annotate your data with **RDF** to assemble **knowledge graphs** (KGs)
- Support future **integration** by referring to other Knowledge Graphs: **URIs**
- Formalize domain knowledge with **ontologies**: **RDFS**, **OWL**
- Mine (multiple) KGs with **graph patterns**: (federated) **SPARQL** queries

# Some available public Knowledge Graphs

RDFportal.org: catalog of 50+ life science KGs

Uniprot: proteins

Rhea: chemical reactions

Bgee: gene expressions

SemOpenAlex: academic papers

IDSM: small molecules

Wikidata: general knowledge

- Covid related queries

Wikipathways

# Going further

- Bob DuCharme
  - What is RDF?
    http://www.bobdc.com/blog/whatisrdf/
  - What is RDFS?
    http://www.bobdc.com/blog/whatisrdfs/
  - SPARQL in 11 minutes
    https://www.youtube.com/watch?v=FvGndkpa4K0
  - Learning SPARQL, 2nd ed. O'Reilly
- https://www.w3.org/TR/rdf11-primer/
- https://www.w3.org/TR/sparql11-query/
- https://www.slideshare.net/LeeFeigenbaum/sparql-cheat-sheet
- http://www.wikipathways.org/index.php/Help:WikiPathways_Sparql_queries
- https://www.fun-mooc.fr/fr/cours/web-semantique-et-web-de-donnees/

# Questions ?

olivier.dameron@univ-rennes1.fr
alban.gaignard@univ-nantes.fr
pierre.larmande@ird.fr

IFB
INSTITUT FRANÇAIS DE BIOINFORMATIQUE

elixir
FRANCE

cnrs INRAE Inserm cea INVESTIR L'AVENIR