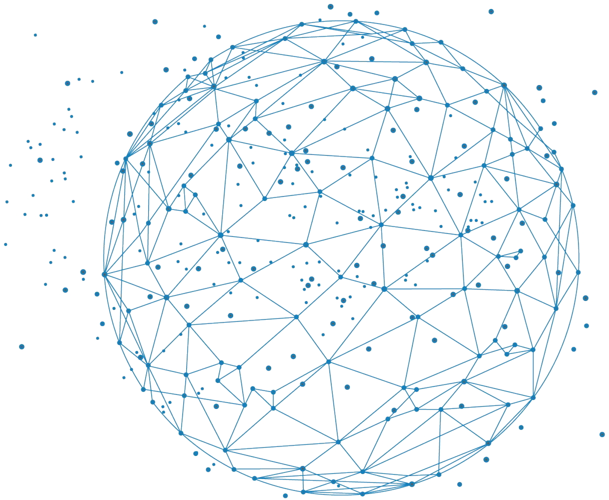
ÉCOLE THÉMATIQUE
ETBII
BIOINFORMATIQUE INTÉGRATIVE IFB

Edition 2024 in Fréjus



Theoretical part
Semantic Web

Olivier DAMERON -
Alban GAIGNARD -
Pierre LARMANDE

CC BY SA

DOI version final

IFB INSTITUT FRANÇAIS DE BIOINFORMATIQUE
elixir FRANCE
cnrs
INRAE
Inserm
cea
INVESTIR L'AVENIR

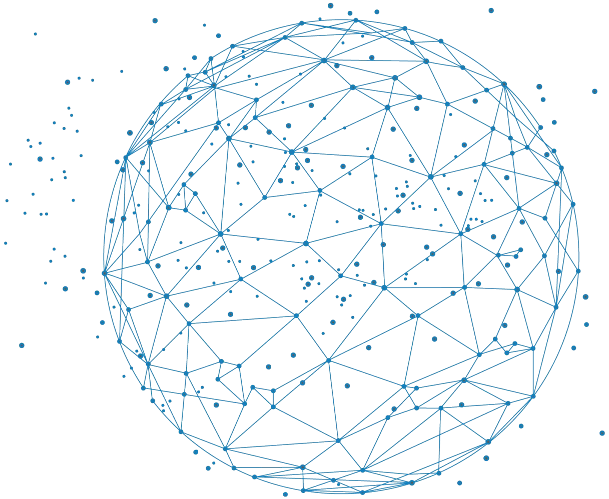
Outline

- Life science data require
  - Integration
  - Knowledge-based reasoning
- The Semantic Web provide a relevant framework
- Use RDF to represent knowledge graphs
- Use SPARQL to query knowledge graphs
- Use RDFS and OWL to formalize knowledge as ontologies

What you will learn (hopefully):

- A general understanding of metadata and (symbolic) knowledge…
- … that relies on surprisingly simple principles

Life science data from an information science perspectives



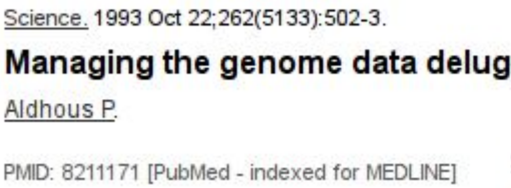
Big data and the deluge of life science data

Big data [Laney, 2001]

Datasets so **large** and **complex** that traditional data processing is inadequate

Life science: data deluge since the 90s [Aldhous, 1993]

- Computerized biomedical data (evidence-based medicine, translational medicine, precision medicine)
- Genomics and bioinformatics

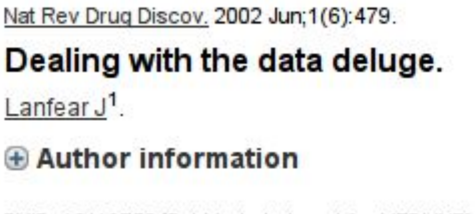


Science. 1993 Oct 22;262(5133):502-3.

Managing the genome data deluge.

Aldhous P.

PMID: 8211171 [PubMed - indexed for MEDLINE]



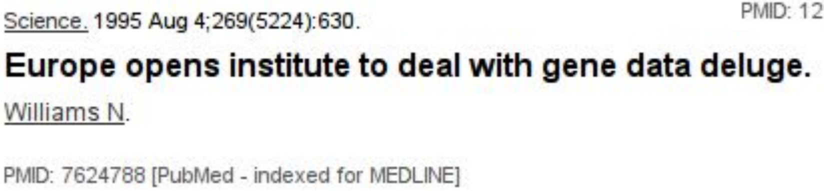
Nat Rev Drug Discov. 2002 Jun;1(6):479.

Dealing with the data deluge.

Lanfear J[1].

⊕ **Author information**

PMID: 12119750 [PubMed - indexed for MEDLINE]



Science. 1995 Aug 4;269(5224):630.

Europe opens institute to deal with gene data deluge.

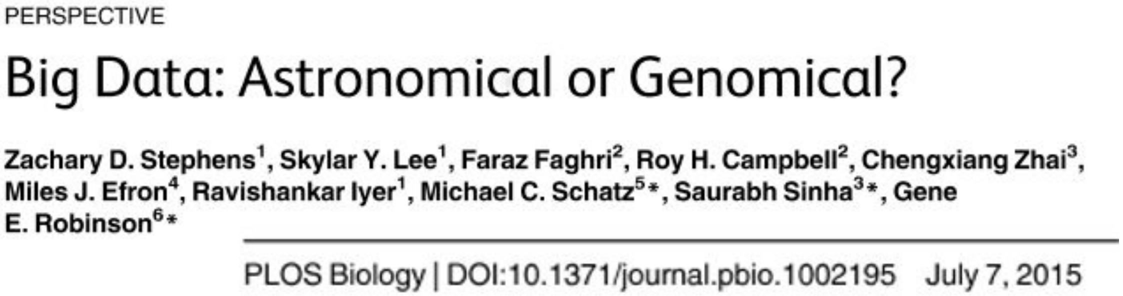
Williams N.

PMID: 7624788 [PubMed - indexed for MEDLINE]

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



PERSPECTIVE

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

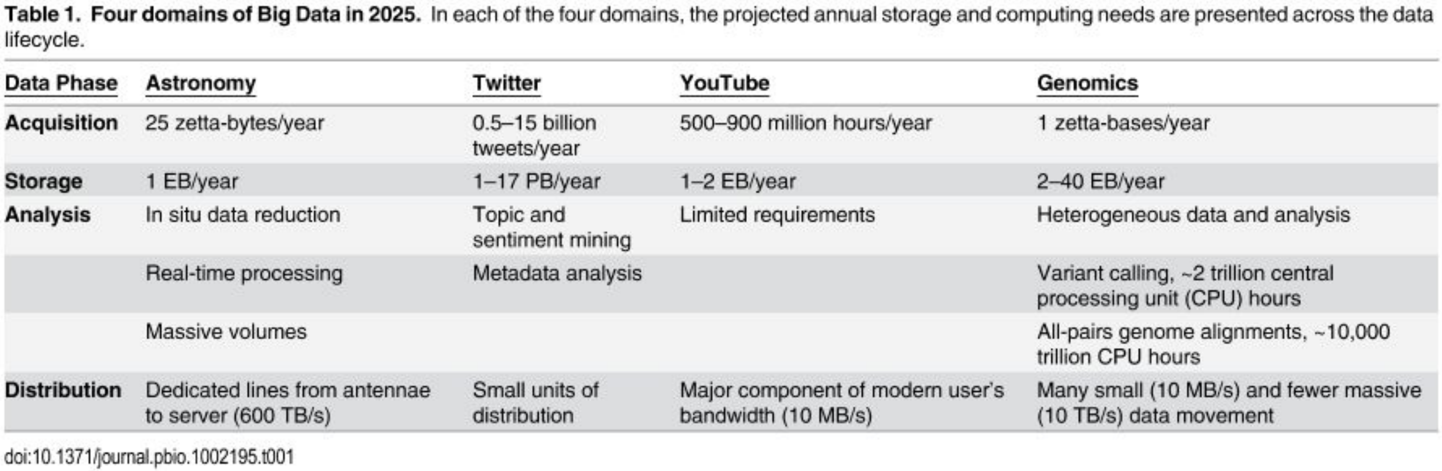


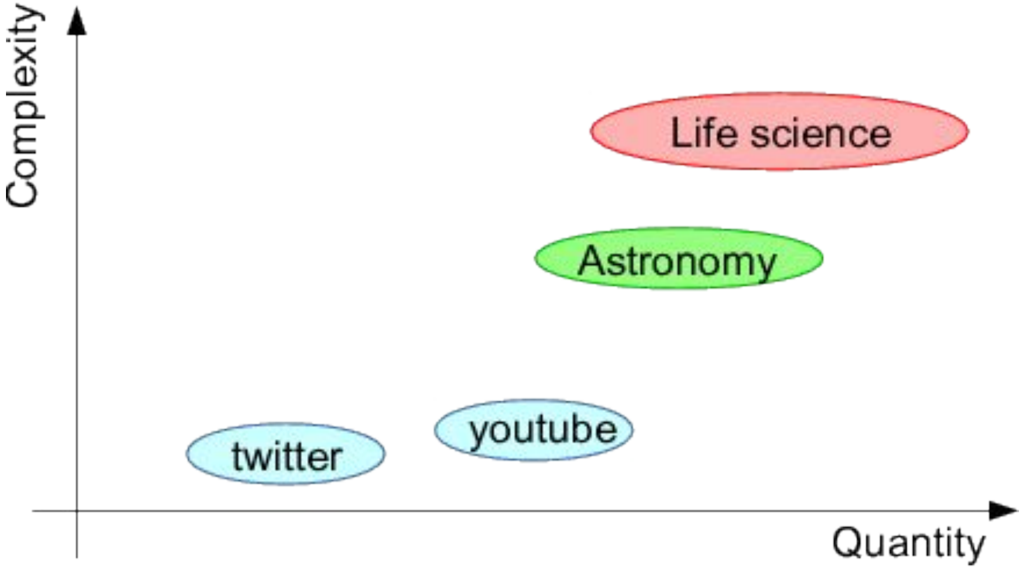
Table 1. Four domains of Big Data in 2025. In each of the four domains, the projected annual storage and computing needs are presented across the data lifecycle.

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

Nature of data complexity

- multiple scale (heterogeneity)
- (highly) interdependent at each scale
- interdependent between scales
- variability
- incompleteness
- evolutive
- distributed
  (and lack of interoperability)



Challenge (computational): How to handle this complexity?

- Experts are very good at analyzing data on their domain (hint…)
  - on their domain
  - on their data
  - with their limited human capacity
  - … help is (badly) needed!
- The difficulty is to analyze data systematically
- Expertise = ability to use knowledge for interpreting data
  - aggregating data
  - inferring connections
- How to use their expertise, instead of trying to re-discover it?
  - assumption: we may not always need to re-discover naively the whole biology at each experiment

Requirements for handling the complexity of life science data

- *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 resources
    (e.g. the transcript associated to a gene, the transcription factors that regulate it…)
  - the categories they belong to
- *Requirement 3:* combine descriptions from different origins
- *Requirement 4:* query these descriptions
- *Requirement 5:* support semantically-rich querying and reasoning (because of the inner complexity) using domain knowledge

Metadata is a love note to the people and machines after you

``Metadata, you see, is really a love note - it might be to yourself, but in fact it’s a love note to the person after you, or the machine after you, where you’ve saved someone that amount of time to find something by telling them what this thing is.”

Jason Scott - http://ascii.textfiles.com/archives/3181

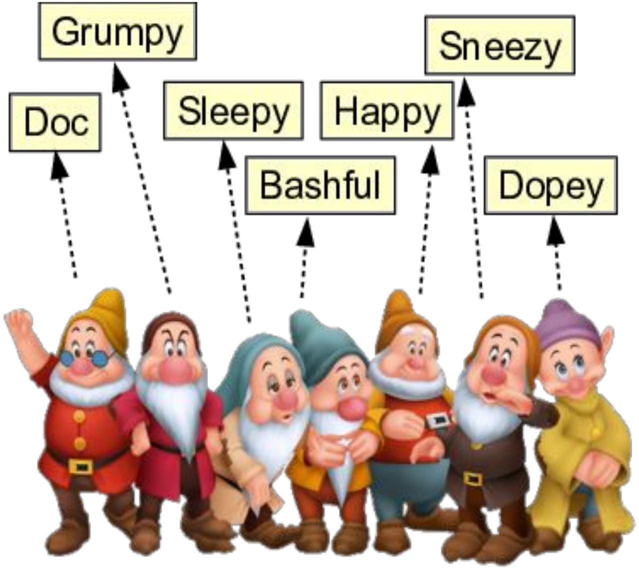
Capturing expertise with annotations

Annotation = explicit representation of the result of some interpretation process

Capturing expertise with annotations

Annotation = explicit representation of the result of some interpretation process

- ideally by an expert (from big data to smart data)
- requires some background knowledge
- formalization spanning the whole semantic spectrum, ranging
  - from free text…
  - …to controlled vocabularies…
  - …to (shared) semantic frameworks



The data annotation process

Annotate data =

- Describe **explicitly**…
  - … the **relevant elements** in your data…
  - … their **characteristics**…
  - … and the **relations between them**
- So that users (you + the non-experts) or programs do not have to go once again through the (tedious, complicated) process of interpreting them
  - It is important
  - It seems easy…
    - *Exercice: describe a set of images*
  - …but turns out more difficult than expected
- 2 aspects:
  - Interpreting and describing data is for domain experts
  - Formalizing and representing the annotations and their dependencies is for data engineers

Fundamental principles

Data / Observations

#1 Annotate data = describe their interpretation -> metadata

#2 There are some dependencies btw some annotations -> knowledge graph

Abstractions / Models

#3 The dependencies between categories support multiple layers of generalization -> ontologies

Graph of data U knowledge graph U ontologies

#4 Reasoning = rules for traversing the graphs

Using annotations for overcoming data complexity

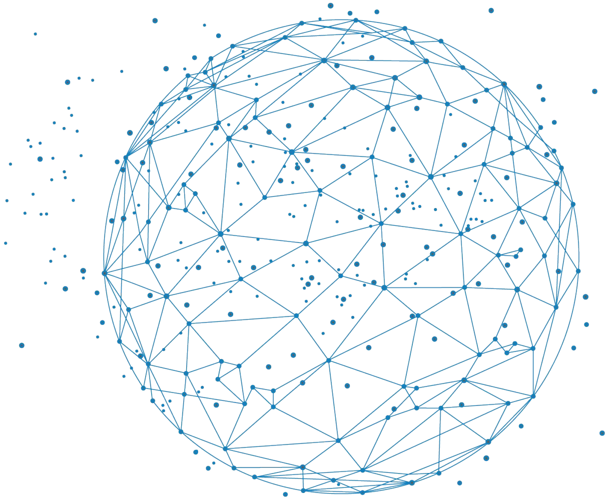
Add annotations? But we have too much data already!

Benefits

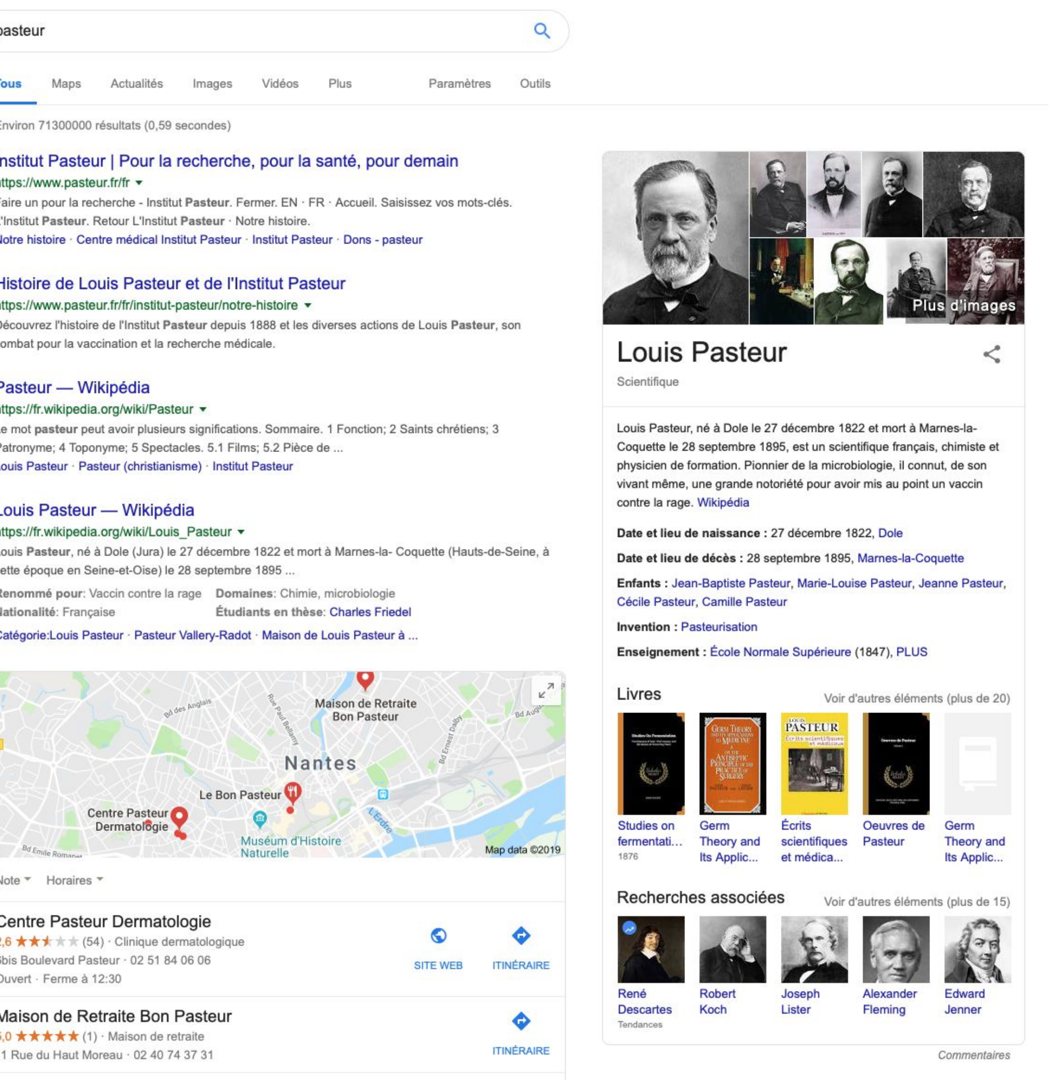
- Can be used as proxy to complex data
- Simplifies by providing a compact abstraction
- Overcomes variability
- Enriches by making explicit the underlying meaning

Storing, sharing and reusing these annotations is the key to life science systematic data analysis

General introduction to semantic web



Knowledge Graphs



nviron 71300000 résultats (0,59 secondes)

nstitut Pasteur | Pour la recherche, pour la santé, pour demain
ttps://www.pasteur.fr/fr
aire un pour la recherche - Institut **Pasteur**. Fermer. EN · FR · Accueil. Saisissez vos mots-clés. 'Institut **Pasteur**. Retour L'Institut **Pasteur** · Notre histoire.
otre histoire · Centre médical Institut Pasteur · Institut Pasteur · Dons - pasteur

Histoire de Louis Pasteur et de l'Institut Pasteur
ttps://www.pasteur.fr/fr/institut-pasteur/notre-histoire
écouvrez l'histoire de l'Institut **Pasteur** depuis 1888 et les diverses actions de Louis **Pasteur**, son ombat pour la vaccination et la recherche médicale.

Pasteur — Wikipédia
ttps://fr.wikipedia.org/wiki/Pasteur
e mot **pasteur** peut avoir plusieurs significations. Sommaire. 1 Fonction; 2 Saints chrétiens; 3 atronyme; 4 Toponyme; 5 Spectacles. 5.1 Films; 5.2 Pièce de ...
ouis Pasteur · Pasteur (christianisme) · Institut Pasteur

ouis Pasteur — Wikipédia
ttps://fr.wikipedia.org/wiki/Louis_Pasteur
ouis **Pasteur**, né à Dole (Jura) le 27 décembre 1822 et mort à Marnes-la- Coquette (Hauts-de-Seine, à ette époque en Seine-et-Oise) le 28 septembre 1895 ...
enommé pour: Vaccin contre la rage **Domaines**: Chimie, microbiologie
ationalité: Française **Étudiants en thèse**: Charles Friedel
atégorie:Louis Pasteur · Pasteur Vallery-Radot · Maison de Louis Pasteur à ...

entre Pasteur Dermatologie
,6 ★★★☆☆ (54) · Clinique dermatologique
bis Boulevard Pasteur · 02 51 84 06 06
uvert · Ferme à 12:30

aison de Retraite Bon Pasteur
,0 ★★★★★ (1) · Maison de retraite
1 Rue du Haut Moreau · 02 40 74 37 31

Louis Pasteur
Scientifique

Louis Pasteur, né à Dole le 27 décembre 1822 et mort à Marnes-la-Coquette le 28 septembre 1895, est un scientifique français, chimiste et physicien de formation. Pionnier de la microbiologie, il connut, de son vivant même, une grande notoriété pour avoir mis au point un vaccin contre la rage. Wikipédia

Date et lieu de naissance : 27 décembre 1822, Dole
Date et lieu de décès : 28 septembre 1895, Marnes-la-Coquette
Enfants : Jean-Baptiste Pasteur, Marie-Louise Pasteur, Jeanne Pasteur, Cécile Pasteur, Camille Pasteur
Invention : Pasteurisation
Enseignement : École Normale Supérieure (1847), PLUS

Livres
Voir d'autres éléments (plus de 20)

Studies on fermentati... 1876
Germ Theory and Its Applic...
Écrits scientifiques et médica...
Oeuvres de Pasteur
Germ Theory and Its Applic...

Recherches associées
Voir d'autres éléments (plus de 15)

René Descartes Tendances
Robert Koch
Joseph Lister
Alexander Fleming
Edward Jenner

Several facets of Artificial Intelligence

Named entity recognition (NLP)

➞ « pasteur » refers to Louis Pasteur

Knowledge Representation

➞ Louis Pasteur is a scientist

➞ a scientist is a kind of person

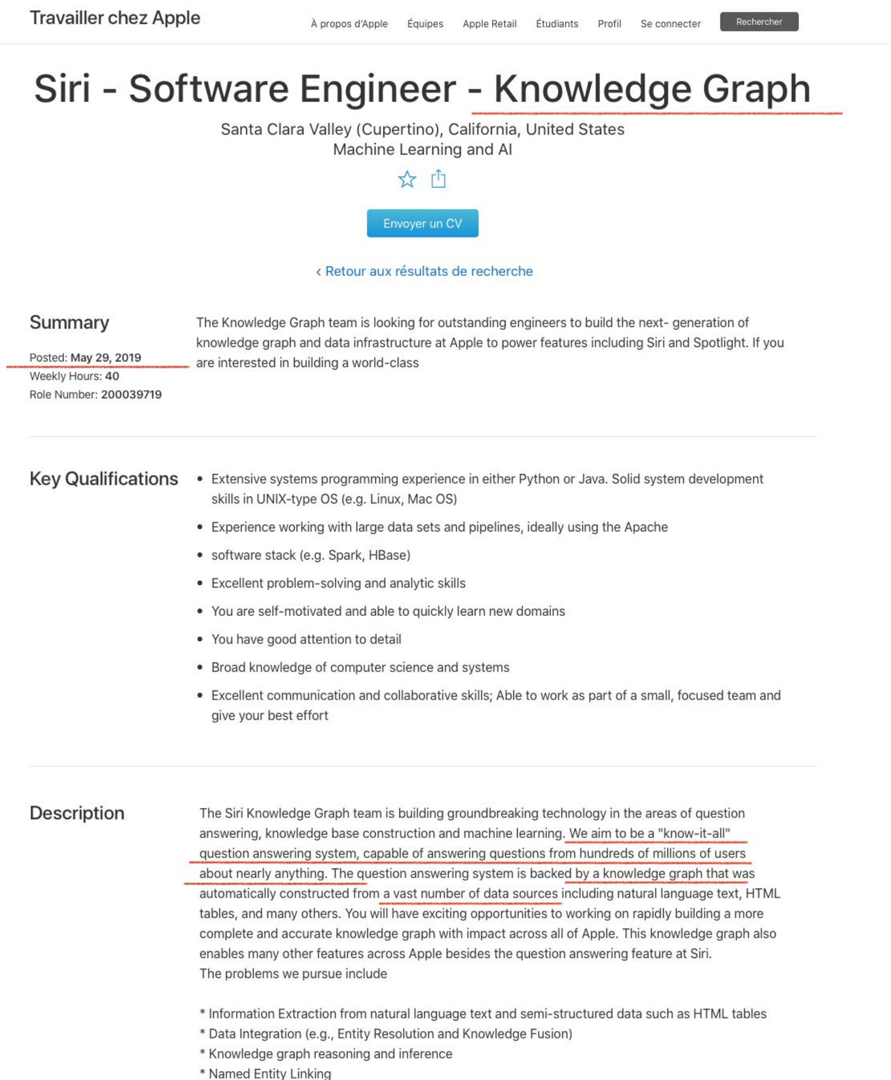
Reasoning

➞ « if an entity is a scientists then its also a person »

➞ « If a web resource is a person then display his/her portrait »

… and many other such as Computer Vision, Machine Learning

Knowledge representation and reasoning : still needed for question answering systems !



Siri - Software Engineer - Knowledge Graph

Santa Clara Valley (Cupertino), California, United States
Machine Learning and AI

Summary

Posted: **May 29, 2019**
Weekly Hours: **40**
Role Number: **200039719**

The Knowledge Graph team is looking for outstanding engineers to build the next- generation of knowledge graph and data infrastructure at Apple to power features including Siri and Spotlight. If you are interested in building a world-class

Key Qualifications

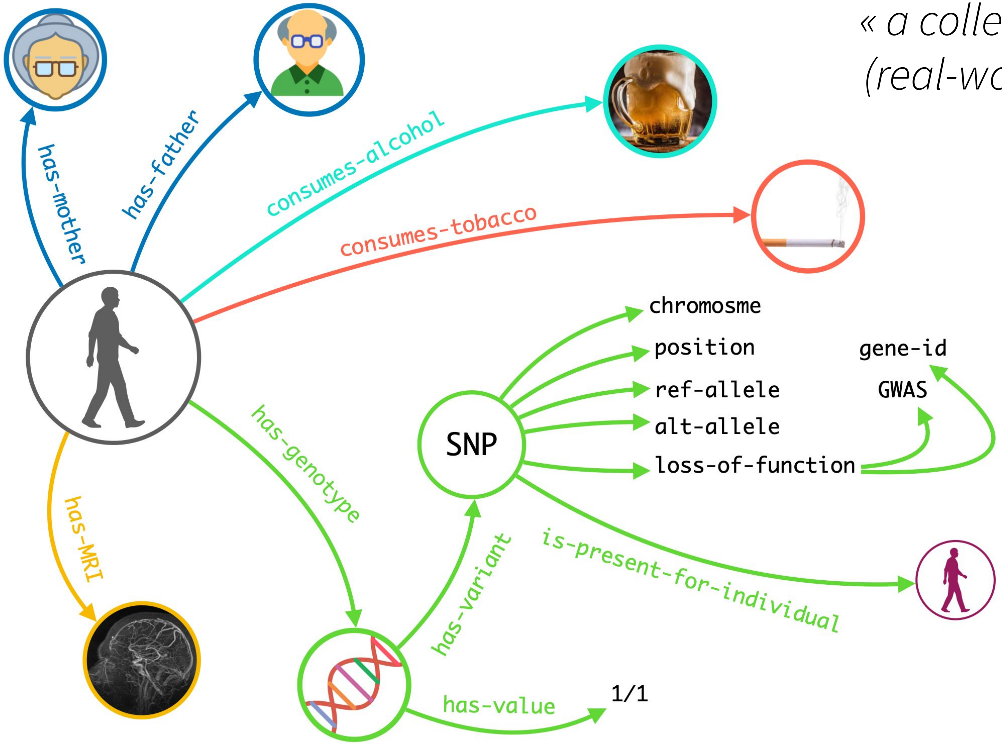
- Extensive systems programming experience in either Python or Java. Solid system development skills in UNIX-type OS (e.g. Linux, Mac OS)
- Experience working with large data sets and pipelines, ideally using the Apache
- software stack (e.g. Spark, HBase)
- Excellent problem-solving and analytic skills
- You are self-motivated and able to quickly learn new domains
- You have good attention to detail
- Broad knowledge of computer science and systems
- Excellent communication and collaborative skills; Able to work as part of a small, focused team and give your best effort

Description

The Siri Knowledge Graph team is building groundbreaking technology in the areas of question answering, knowledge base construction and machine learning. We aim to be a "know-it-all" question answering system, capable of answering questions from hundreds of millions of users about nearly anything. The question answering system is backed by a knowledge graph that was automatically constructed from a vast number of data sources including natural language text, HTML tables, and many others. You will have exciting opportunities to working on rapidly building a more complete and accurate knowledge graph with impact across all of Apple. This knowledge graph also enables many other features across Apple besides the question answering feature at Siri.
The problems we pursue include

* Information Extraction from natural language text and semi-structured data such as HTML tables
* Data Integration (e.g., Entity Resolution and Knowledge Fusion)
* Knowledge graph reasoning and inference
* Named Entity Linking

What is a Knowledge Graph ?



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

a **database**
→ information storage / extraction

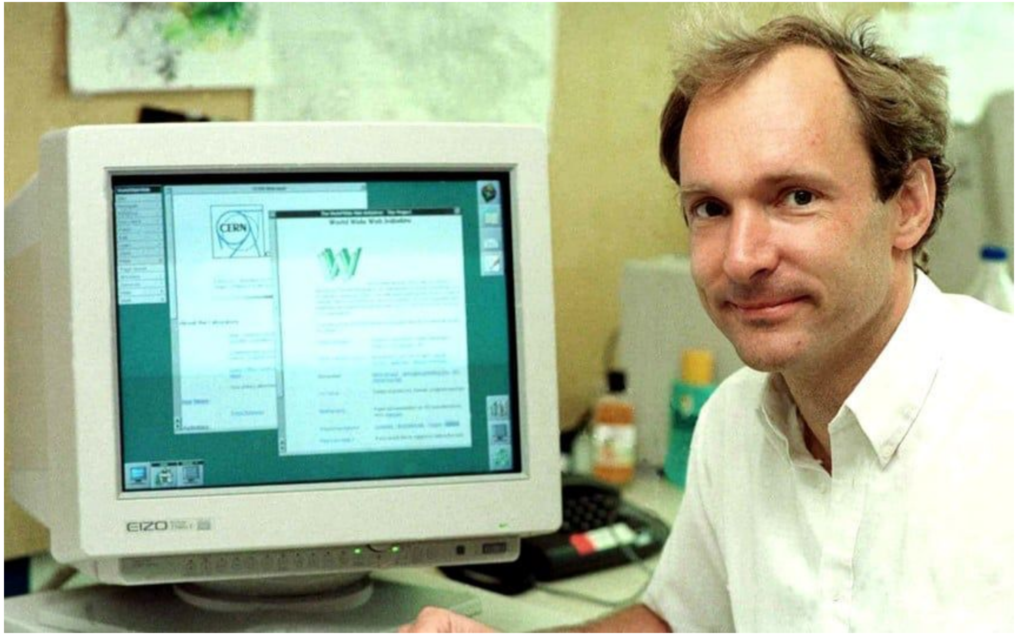
a **graph**
→ network analysis

a **knowledge base**
→ formal semantics
(logical facts, logical inferences)

From **linked documents** (Web) …

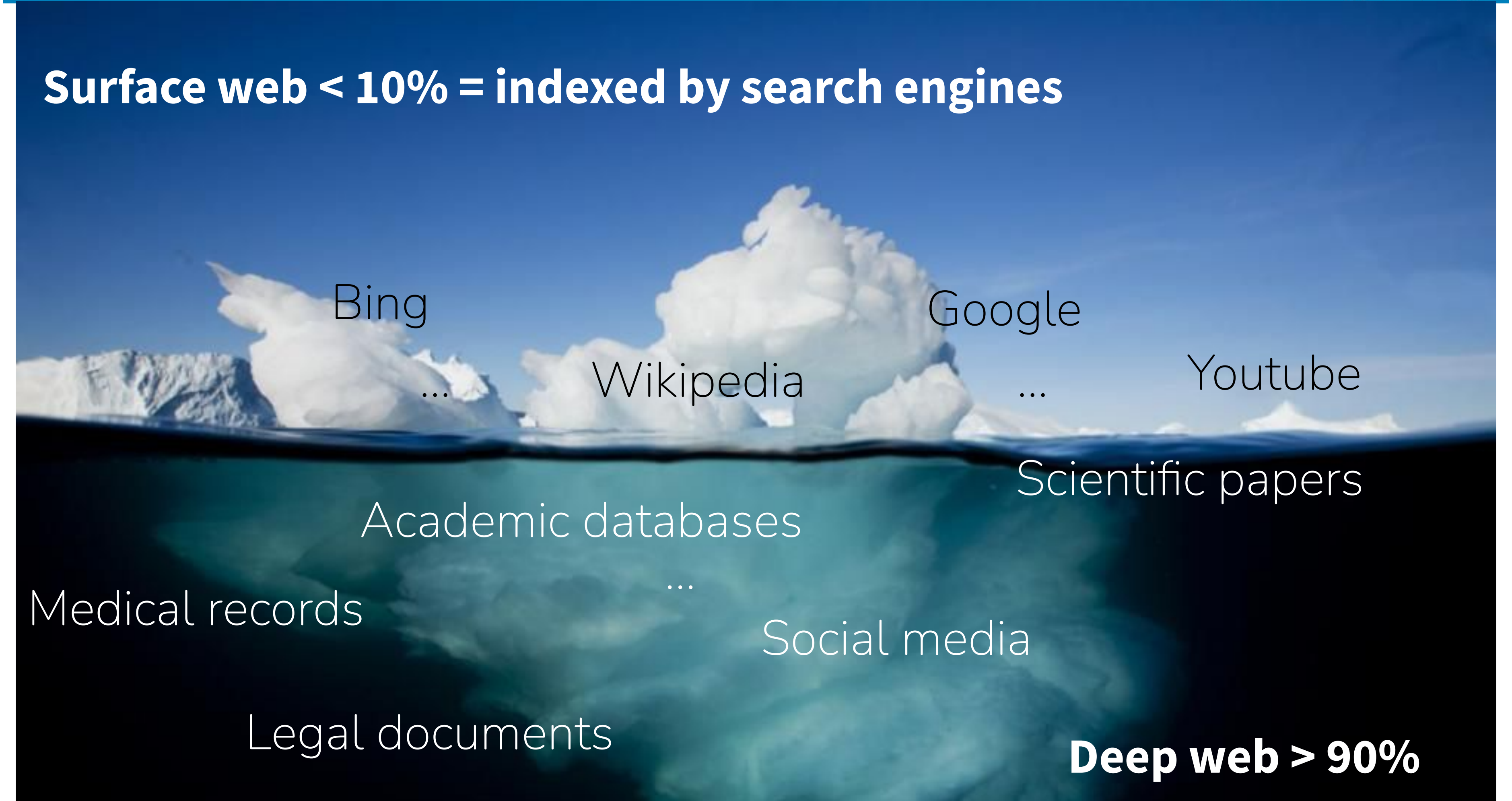
… to **linked data** (Semantic Web)

90’s: web of documents



A **de-centralized** system of **hypertext documents** based on URL, HTTP, and HTML **standards** (World Wide Web consortium, W3C).

Nowadays: Web of data



Surface web < 10% = indexed by search engines

Bing

Google

Wikipedia

Youtube

...

...

Scientific papers

Academic databases

...

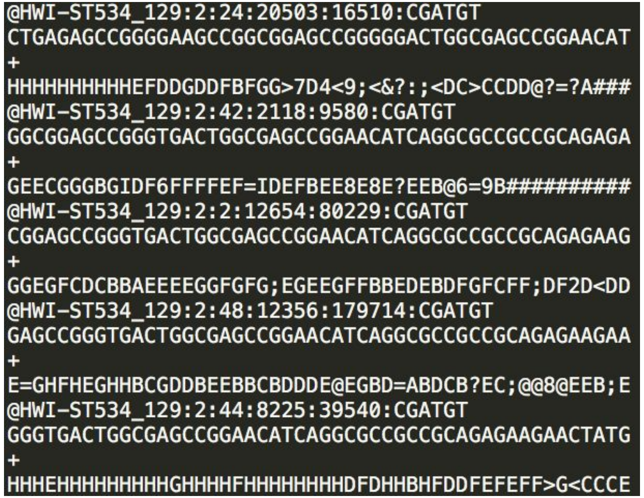
Medical records

Social media

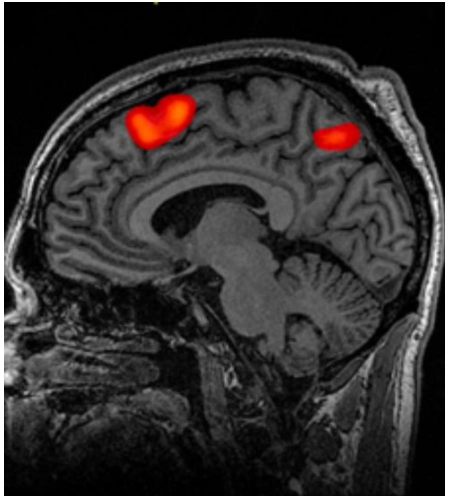
Legal documents

Deep web > 90%

Why so much data is not findable ?



```
@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
+
HHHEHHHHHHHHGHHHHHFHHHHHHHHDFDHHBHFDDFEFEFF>G<CCCE
```



Genetic sequences

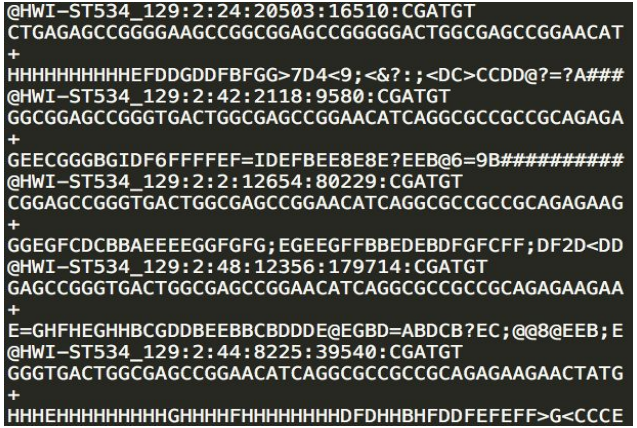
- 1st line = label
- 2nd line = raw sequence (A,T,C,G)
- 4th line = quality score for each base

Medical image

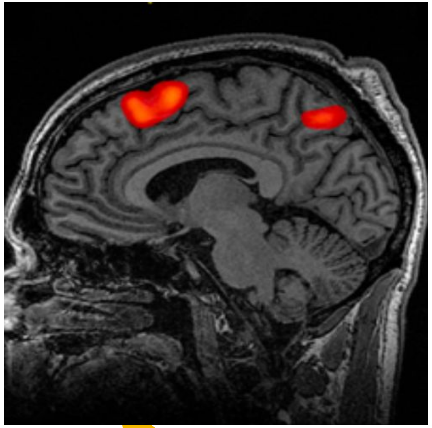
- MRI as imaging modality
- brain as imaging target
- labeled data (intensity/volume ?)

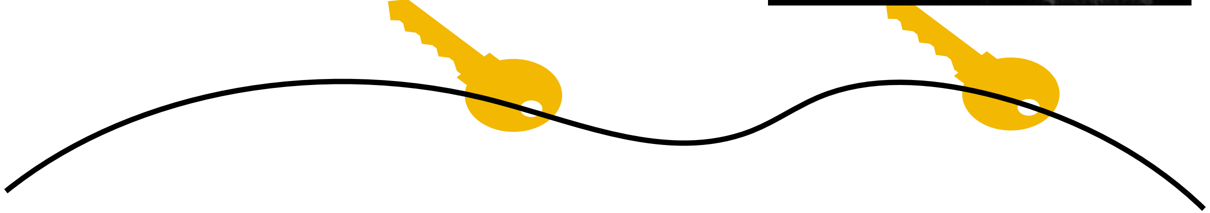
Semantic web

Make 🤖 and 🧍 better exchange, interpret & reason on diverse data !!

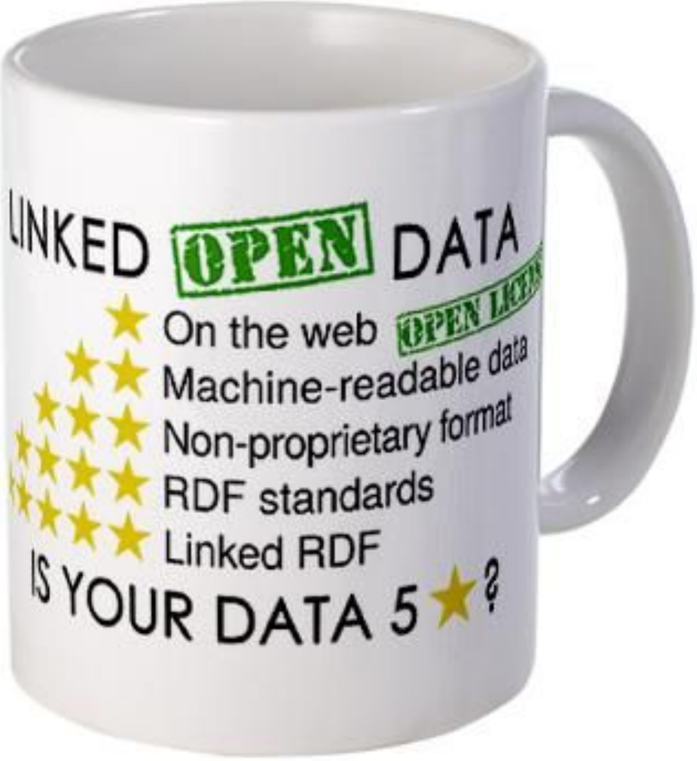


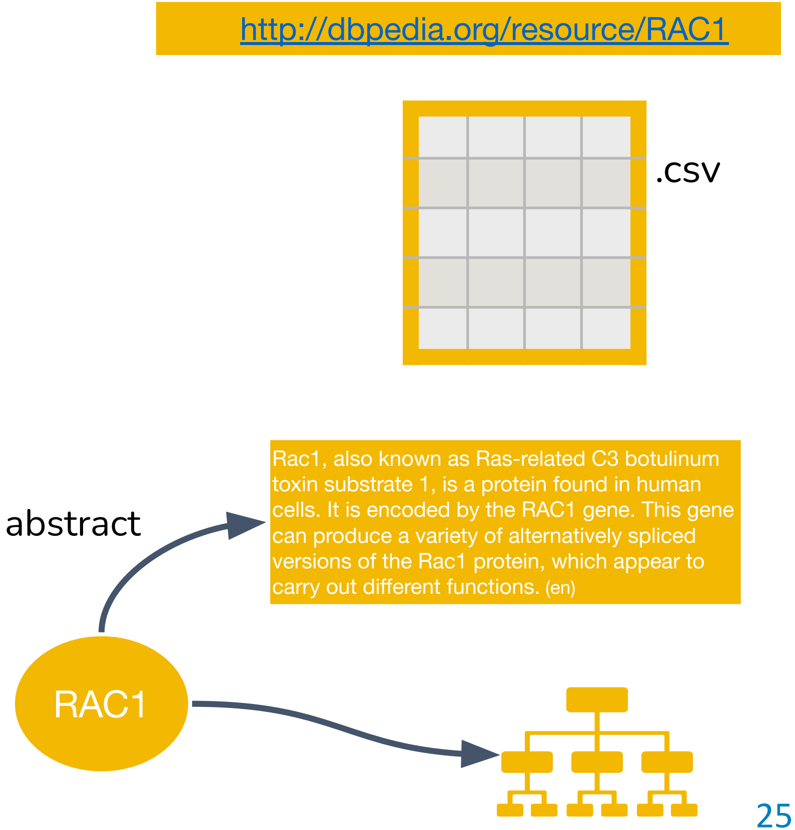
```
@HWI-ST534_129:2:24:20503:16510:CGATGT
CTGAGAGCCGGGGAAGCCGGCGGAGCCGGGGGACTGGCGAGCCGGAACAT
+
HHHHHHHHHHEFDDGDDFBFGG>7D4<9;<&?:;<DC>CCDD@?=?A###
@HWI-ST534_129:2:42:2118:9580:CGATGT
GGCGGAGCCGGGTGACTGGCGAGCCGGAACATCAGGCGCCGCCGCAGAGA
+
GEECGGGBGIDF6FFFFEF=IDEFBEE8E8E?EEB@6=9B##########
@HWI-ST534_129:2:2:12654:80229:CGATGT
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
```





W3C recommendations





http://dbpedia.org/resource/RAC1

.csv

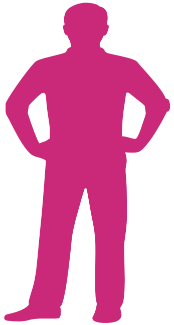
abstract

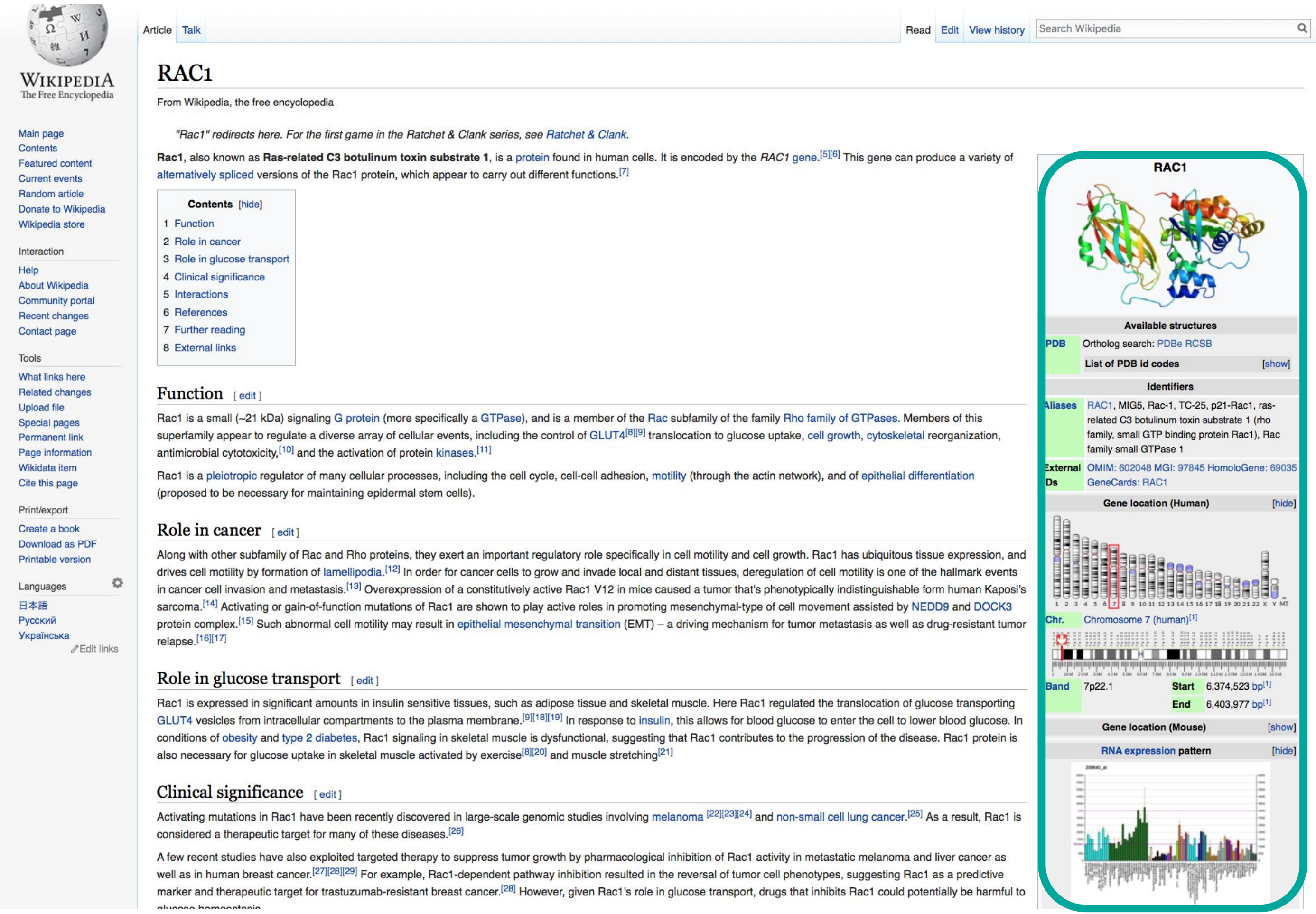
Rac1, also known as Ras-related C3 botulinum toxin substrate 1, is a protein found in human cells. It is encoded by the RAC1 gene. This gene can produce a variety of alternatively spliced versions of the Rac1 protein, which appear to carry out different functions. (en)

RAC1

→ **technological stack to implement FAIR principles** 😀

Wikipedia … for humans





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 of this superfamily appear to regulate a diverse array of cellular events, including the control of GLUT4[8][9] translocation to glucose uptake, cell growth, cytoskeletal reorganization, 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 expression, 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 hallmark events in cancer cell invasion and metastasis.[13] Overexpression of a constitutively active Rac1 V12 in mice caused a tumor that's phenotypically indistinguishable form human Kaposi's 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 and DOCK3 protein complex.[15] Such abnormal cell motility may result in epithelial mesenchymal transition (EMT) – a driving mechanism for tumor metastasis as well as drug-resistant tumor 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 transporting 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 glucose. In conditions of obesity and type 2 diabetes, Rac1 signaling in skeletal muscle is dysfunctional, suggesting that Rac1 contributes to the progression of the disease. Rac1 protein is also necessary for glucose uptake in skeletal muscle activated by exercise[8][20] and muscle stretching[21]

Clinical significance [edit]

Activating mutations in Rac1 have been recently discovered in large-scale genomic studies involving melanoma [22][23][24] and non-small cell lung cancer.[25] As a result, Rac1 is considered a therapeutic target for many of these diseases.[26]

A few recent studies have also exploited targeted therapy to suppress tumor growth by pharmacological inhibition of Rac1 activity in metastatic melanoma and liver cancer as well as in human breast cancer.[27][28][29] For example, Rac1-dependent pathway inhibition resulted in the reversal of tumor cell phenotypes, suggesting Rac1 as a predictive marker and therapeutic target for trastuzumab-resistant breast cancer.[28] However, given Rac1's role in glucose transport, drugs that inhibits Rac1 could potentially be harmful to

RAC1

Available structures

PDB	Ortholog search: PDBe RCSB **List of PDB id codes** [show]

Identifiers

Aliases	RAC1, MIG5, Rac-1, TC-25, p21-Rac1, ras-related C3 botulinum toxin substrate 1 (rho family, small GTP binding protein Rac1), Rac family small GTPase 1
External IDs	OMIM: 602048 MGI: 97845 HomoloGene: 69035 GeneCards: RAC1

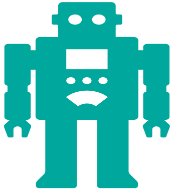
Gene location (Human) [hide]

Chr.	Chromosome 7 (human)[1]		
Band	7p22.1	**Start**	6,374,523 bp[1]
		End	6,403,977 bp[1]

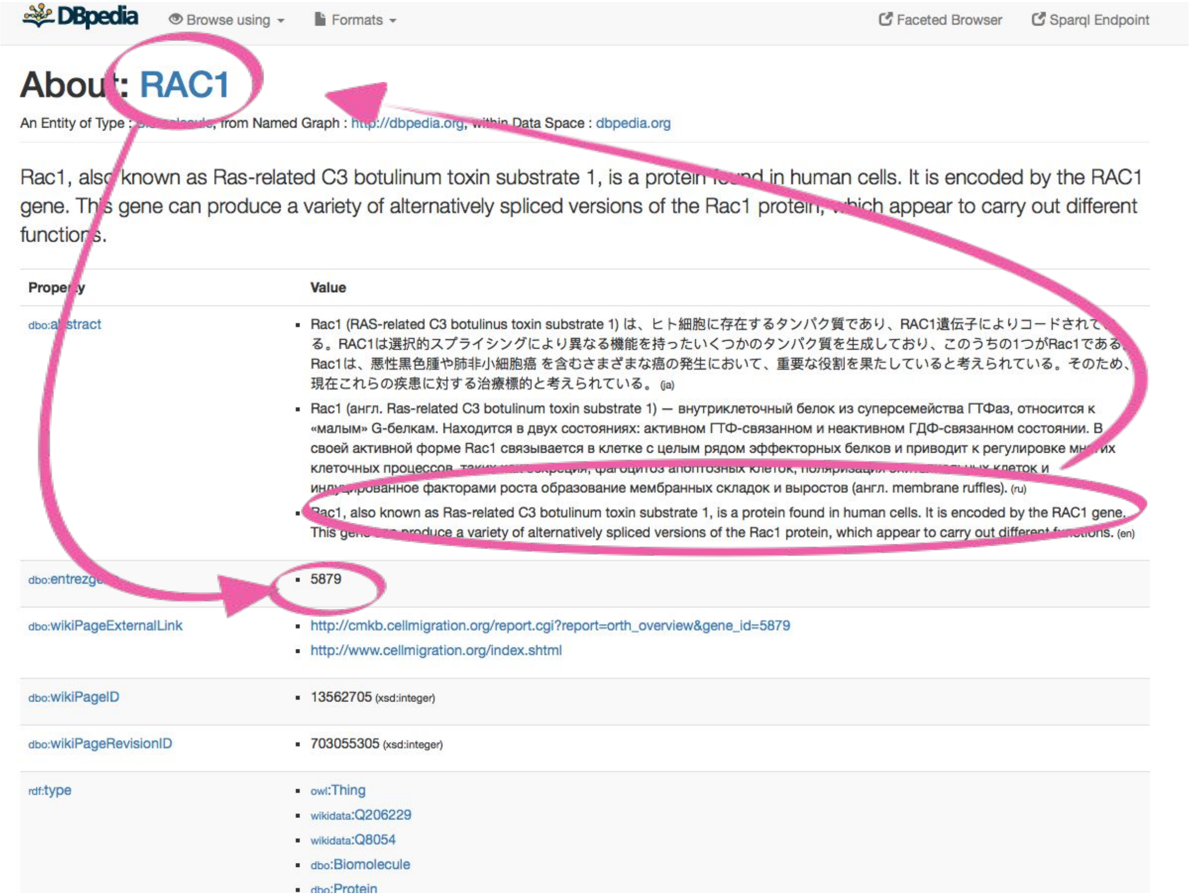
Gene location (Mouse) [show]

RNA expression pattern [hide]

Wikipedia … for machines (DBpedia)



Give me all gene IDs described with « toxin »



About: RAC1

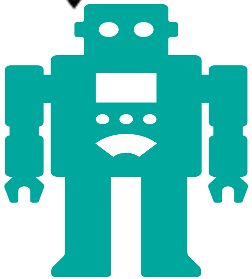
An Entity of Type : Biomolecule, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

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.

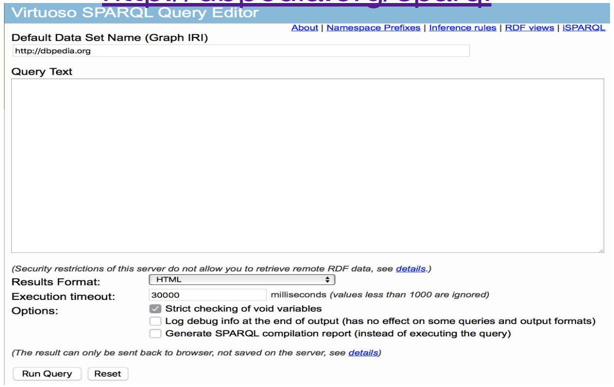
Property	Value
dbo:abstract	▪ Rac1 (RAS-related C3 botulinus toxin substrate 1) は、ヒト細胞に存在するタンパク質であり、RAC1遺伝子によりコードされている。RAC1は選択的スプライシングにより異なる機能を持ったいくつかのタンパク質を生成しており、このうちの1つがRac1である。Rac1は、悪性黒色腫や肺非小細胞癌 を含むさまざまな癌の発生において、重要な役割を果たしていると考えられている。そのため、現在これらの疾患に対する治療標的と考えられている。 (ja) ▪ Rac1 (англ. Ras-related C3 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 клеточных процессов, таких как секреция, фагоцитоз апоптозных клеток, поляризация эпителиальных клеток и индуцированное 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)
dbo:entrezgene	▪ 5879
dbo:wikiPageExternalLink	▪ http://cmkb.cellmigration.org/report.cgi?report=orth_overview&gene_id=5879 ▪ http://www.cellmigration.org/index.shtml
dbo:wikiPageID	▪ 13562705 (xsd:integer)
dbo:wikiPageRevisionID	▪ 703055305 (xsd:integer)
rdf:type	▪ owl:Thing ▪ wikidata:Q206229 ▪ wikidata:Q8054 ▪ dbo:Biomolecule ▪ dbo:Protein

Querying DBpedia with SPARQL

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



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

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

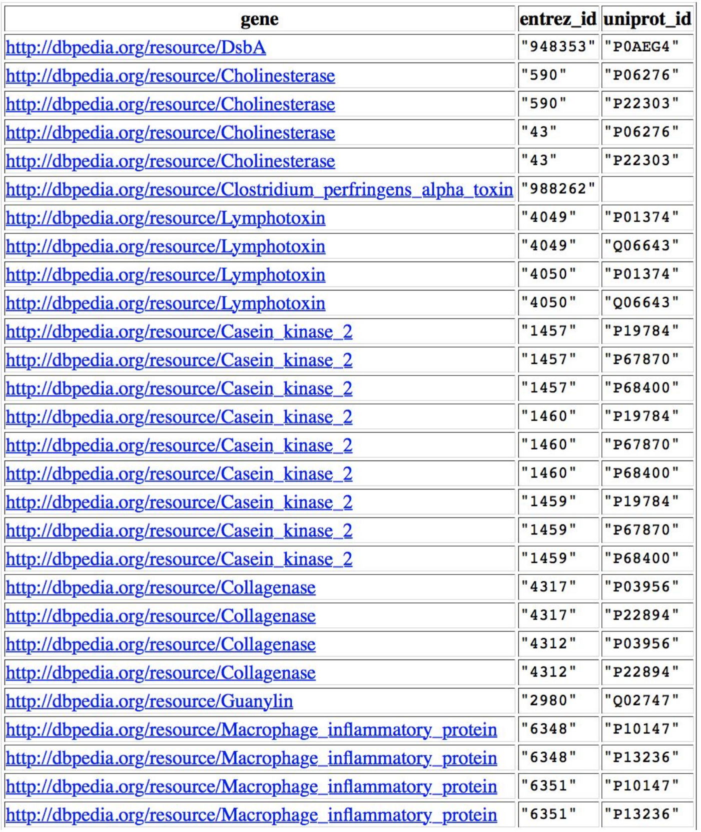
Results Format: HTML

Execution timeout: 30000 milliseconds (values less than 1000 are ignored)

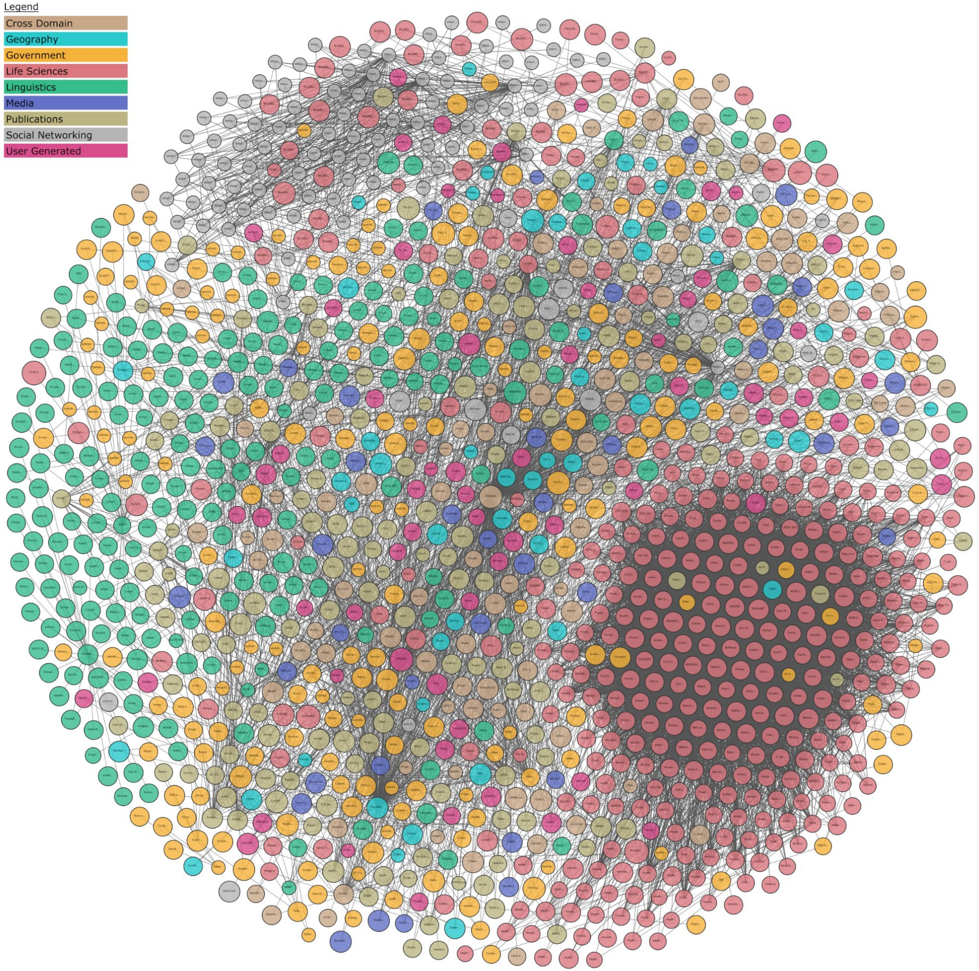
Options:
- ☑ Strict checking of void variables
- ☐ Log debug info at the end of output (has no effect on some queries and output formats)
- ☐ Generate SPARQL compilation report (instead of executing the query)

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

Run Query
Reset



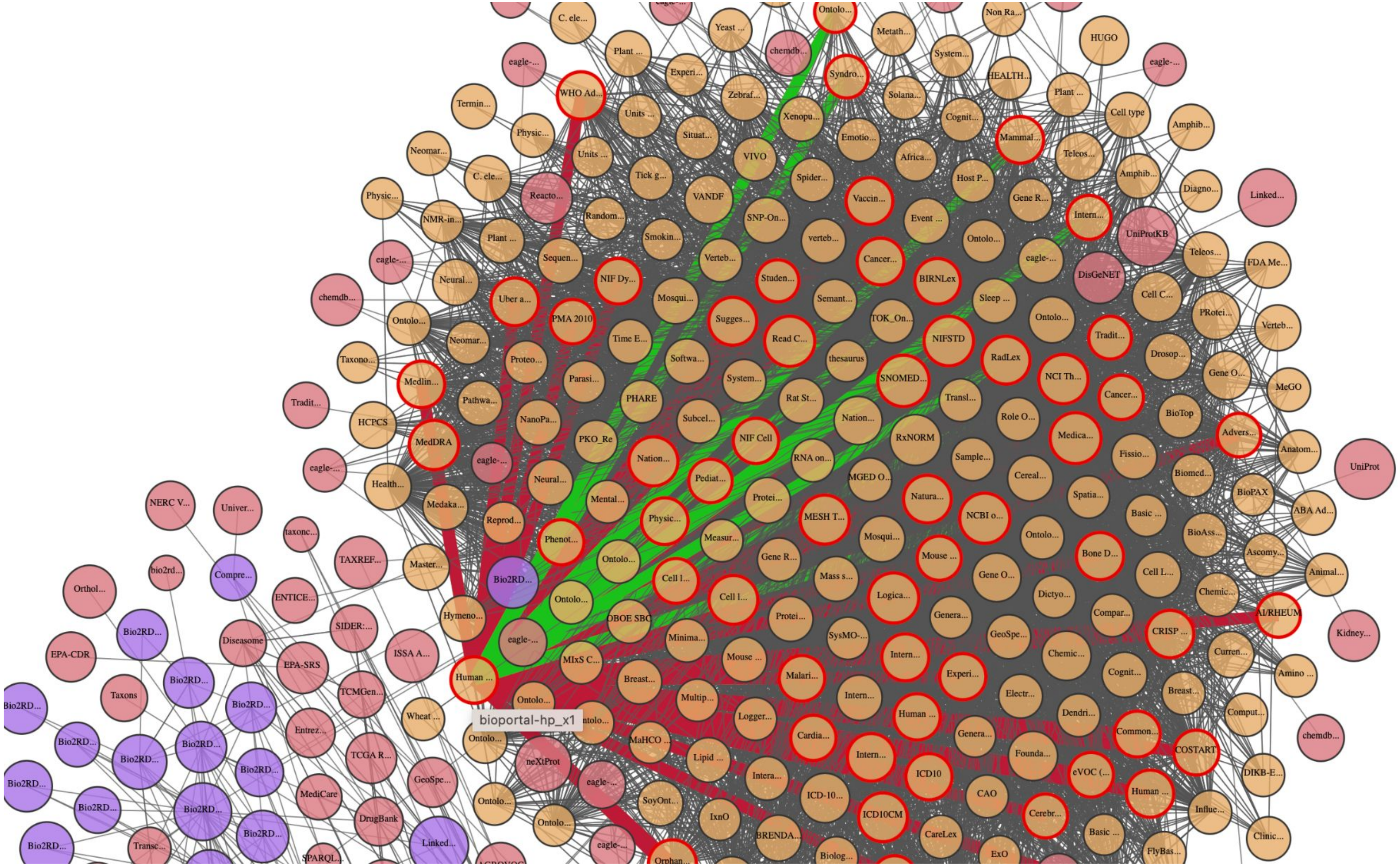
gene	entrez_id	uniprot_id
http://dbpedia.org/resource/DsbA	"948353"	"P0AEG4"
http://dbpedia.org/resource/Cholinesterase	"590"	"P06276"
http://dbpedia.org/resource/Cholinesterase	"590"	"P22303"
http://dbpedia.org/resource/Cholinesterase	"43"	"P06276"
http://dbpedia.org/resource/Cholinesterase	"43"	"P22303"
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"

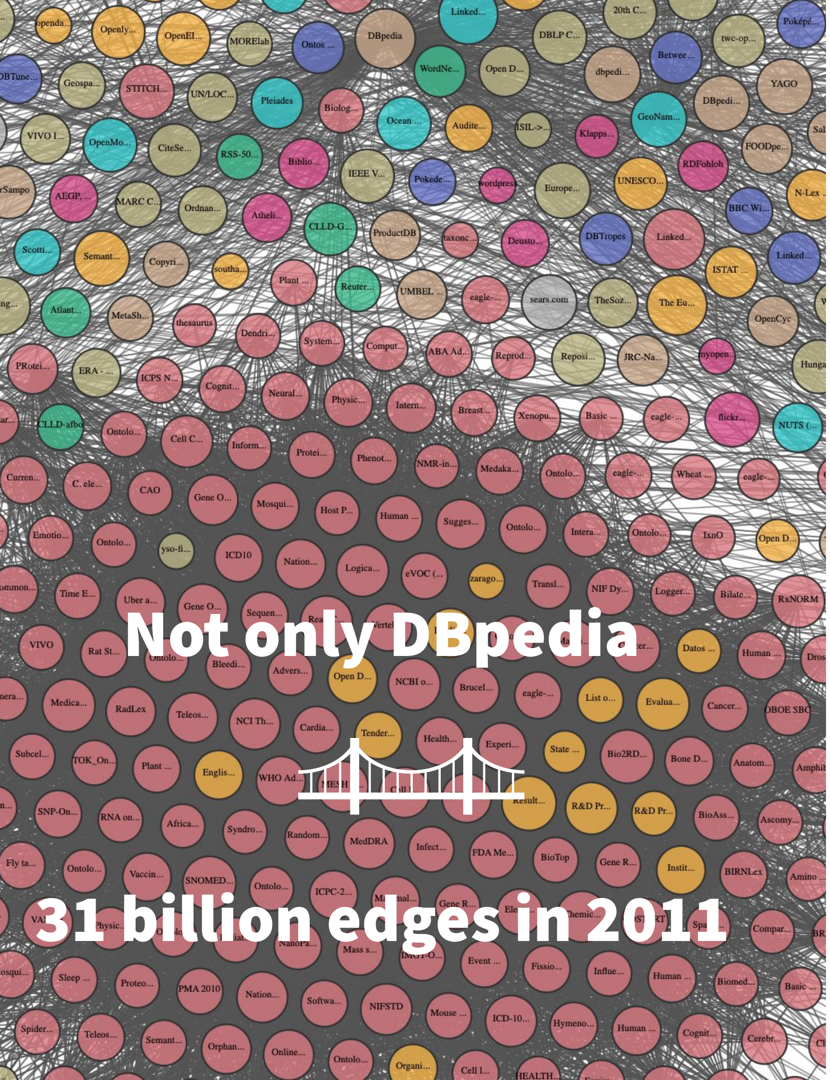


The Linked Open Data Cloud from lod-cloud.net

CC BY

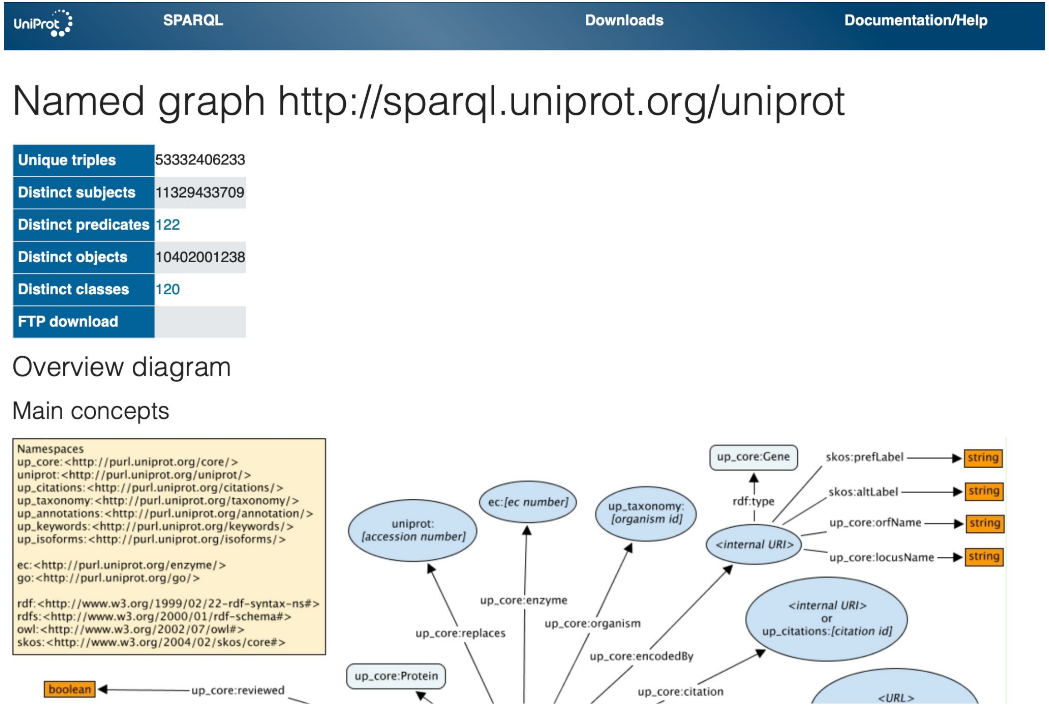
Human Phenotype Ontology linked to Life Science ontologies





Not only DBpedia

31 billion edges in 2011



Named graph http://sparql.uniprot.org/uniprot

Unique triples	53332406233
Distinct subjects	11329433709
Distinct predicates	122
Distinct objects	10402001238
Distinct classes	120
FTP download	

Overview diagram

Main concepts

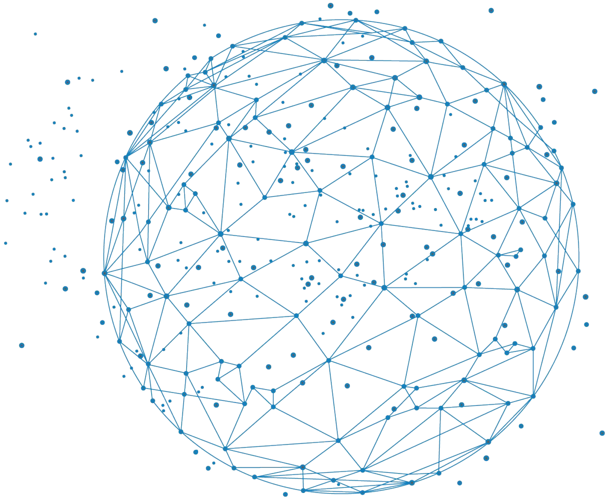
Namespaces
up_core:<http://purl.uniprot.org/core/>
uniprot:<http://purl.uniprot.org/uniprot/>
up_citations:<http://purl.uniprot.org/citations/>
up_taxonomy:<http://purl.uniprot.org/taxonomy/>
up_annotations:<http://purl.uniprot.org/annotation/>
up_keywords:<http://purl.uniprot.org/keywords/>
up_isoforms:<http://purl.uniprot.org/isoforms/>

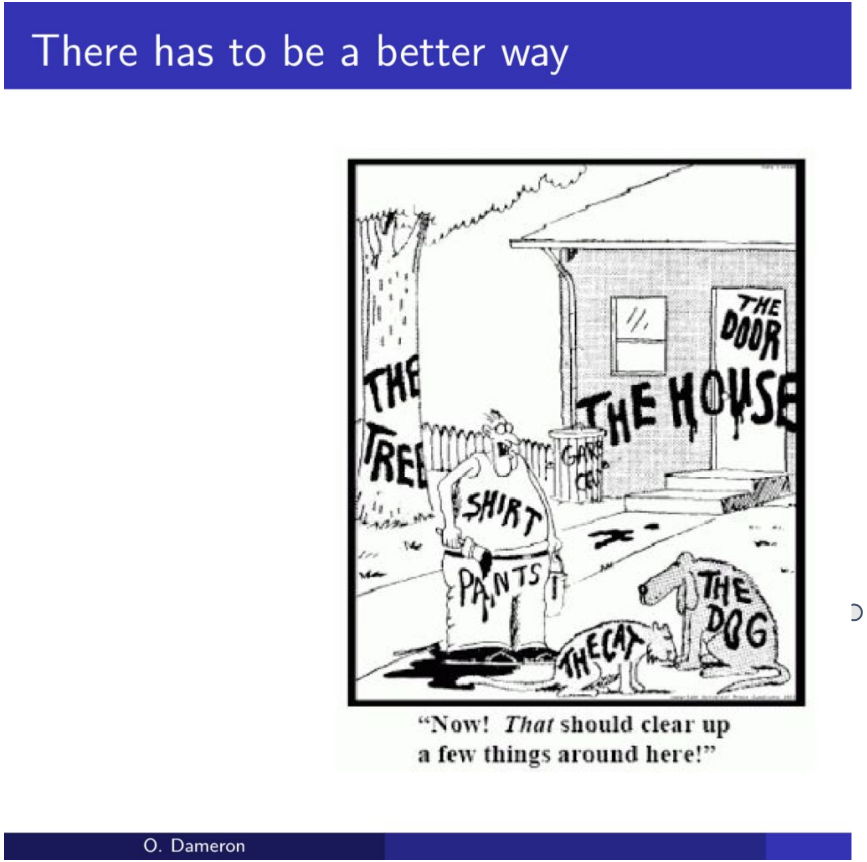
ec:<http://purl.uniprot.org/enzyme/>
go:<http://purl.uniprot.org/go/>

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#>

53 billion triples in Uniprot KG (Jan. 2024)

Representing Knowledge Graphs

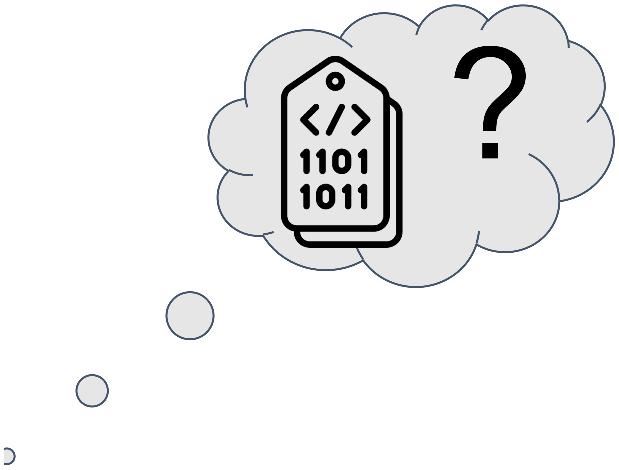




There has to be a better way

“Now! *That* should clear up a few things around here!”

O. Dameron

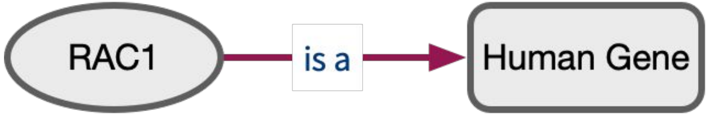


RDF: formalize statements

Simple sentence “subject - verb - object”

RAC1 is a human gene.

Graphical representation



Machine readable syntax (RDF)

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

https://www.w3.org/TR/rdf11-primer/

RDF: syntax to describe (and link) resources

Definition

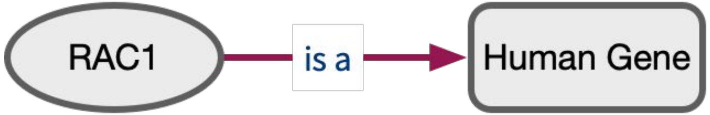
(1) an RDF statement represents a **relationship** between two entities: the **subject** and the **object**

(2) the **predicate** represents the nature of their relationship

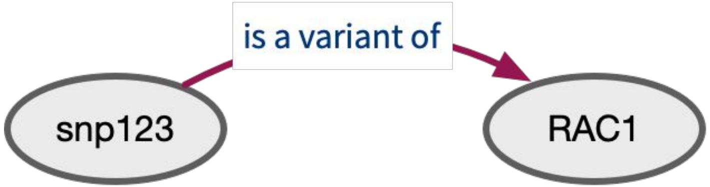
(3) the relationship is phrased in a **directional** way (from subject to object) and is called in RDF a **property**

(4) RDF statements are called **triples**: they consist of three elements they

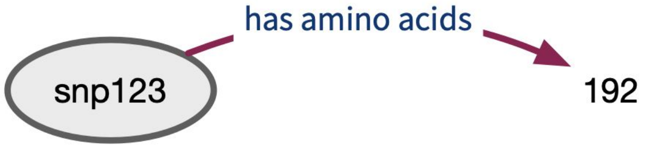
(5) Nodes are **URIs** to identify **named entities** on the web or **Literals** to represents text, numbers



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



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



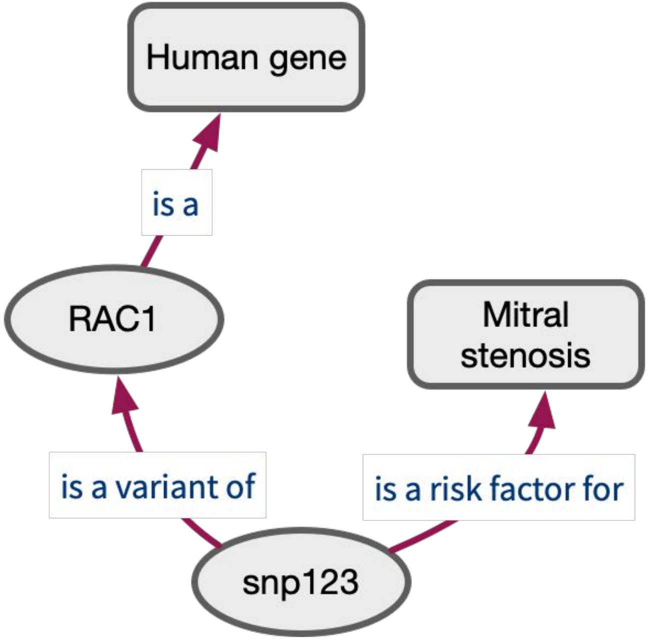
```
<http://RAC1>  <http://has_amino_acids>  192 .
```

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 in the Web.



Writing RDF graphs with Turtle

Definitions

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

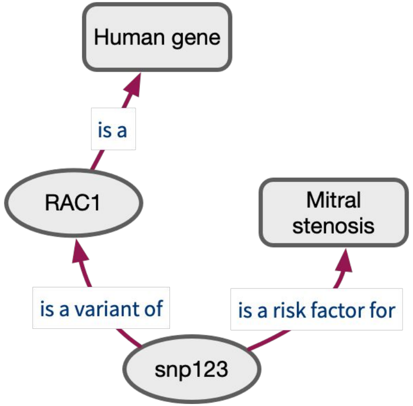
```
S  P  O .
```

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

$$\begin{array}{llll} S & P_1 & O_1 & ; \\ & P_2 & O_2 & . \end{array}$$

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

$$S \quad P \quad O_1 \; , \; O_2 \; .$$



```
@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_of ns:Mitral_stenosis
.
```

In practice … from text to KG

Exercise

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.”*

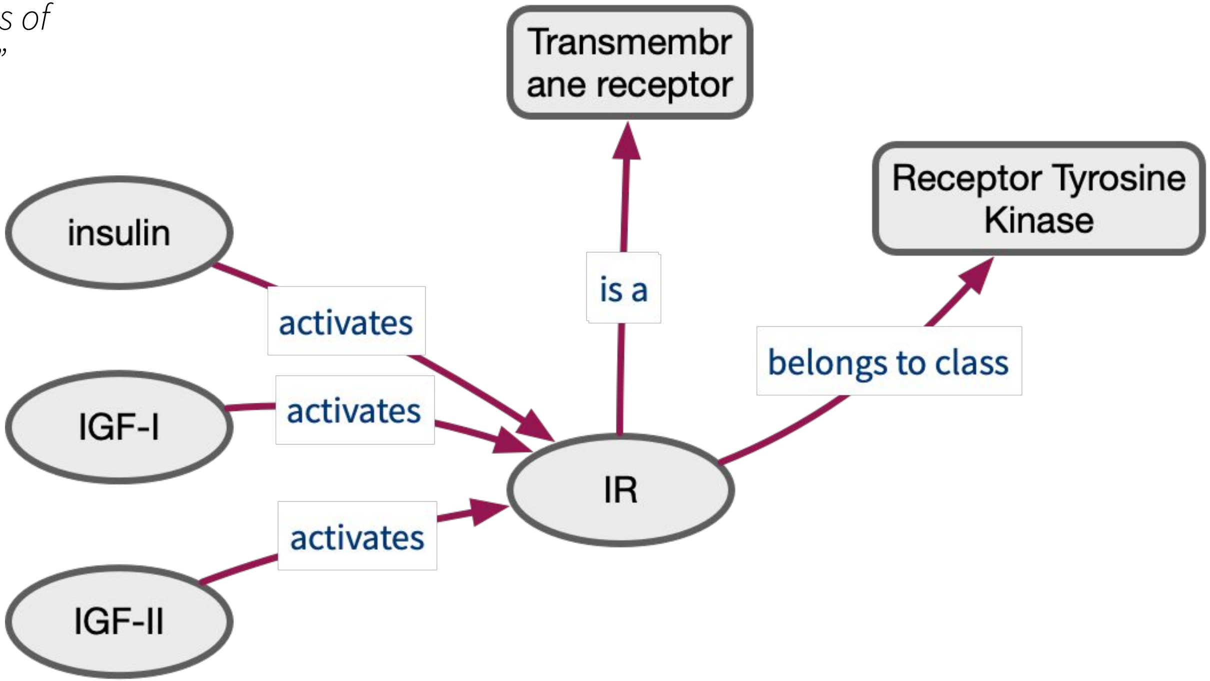
1. Draft a **graphical representation** of the associated knowledge graph.
   - ✓ Identify verbs → predicates
   - ✓ Identify linked entities,
     who is a subject of a relation ?
     who is the object of a relation ?

2. Give the **RDF syntax** for this KG.

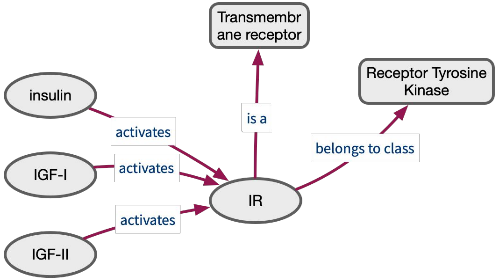
Practice …

In practice …

"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."

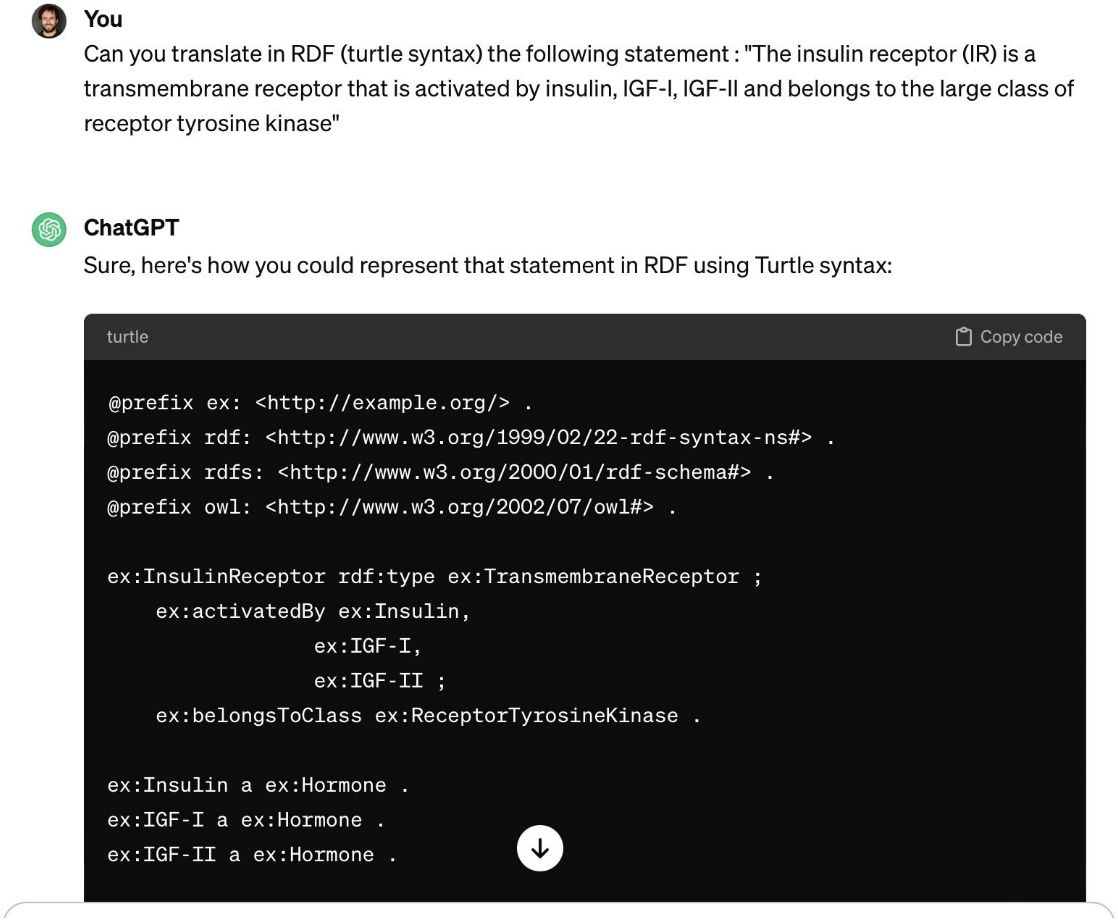


In practice …



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

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 .
```

Practice …



You

Can you translate in RDF (turtle syntax) the following statement : "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"

ChatGPT

Sure, here's how you could represent that statement in RDF using Turtle syntax:

```turtle
@prefix ex: <http://example.org/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .

ex:InsulinReceptor rdf:type ex:TransmembraneReceptor ;
    ex:activatedBy ex:Insulin,
                 ex:IGF-I,
                 ex:IGF-II ;
    ex:belongsToClass ex:ReceptorTyrosineKinase .

ex:Insulin a ex:Hormone .
ex:IGF-I a ex:Hormone .
ex:IGF-II a ex:Hormone .
```

In practice … from KG to text

https://legacy.uniprot.org/uniprot/P06213.ttl

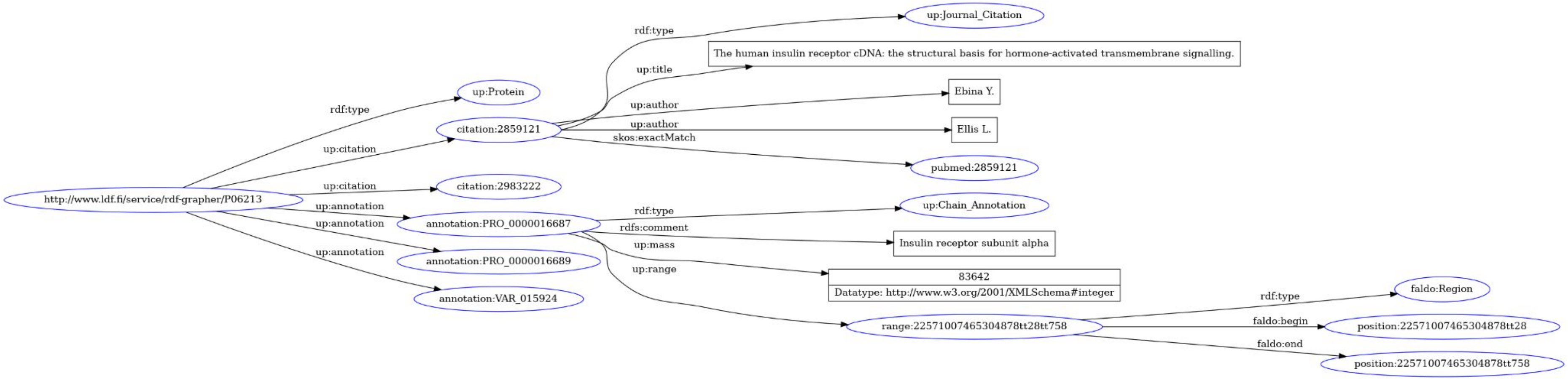
```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix up: <http://purl.uniprot.org/core/> .
@prefix annotation: <http://purl.uniprot.org/annotation/> .

<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 .
```

Exercice

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

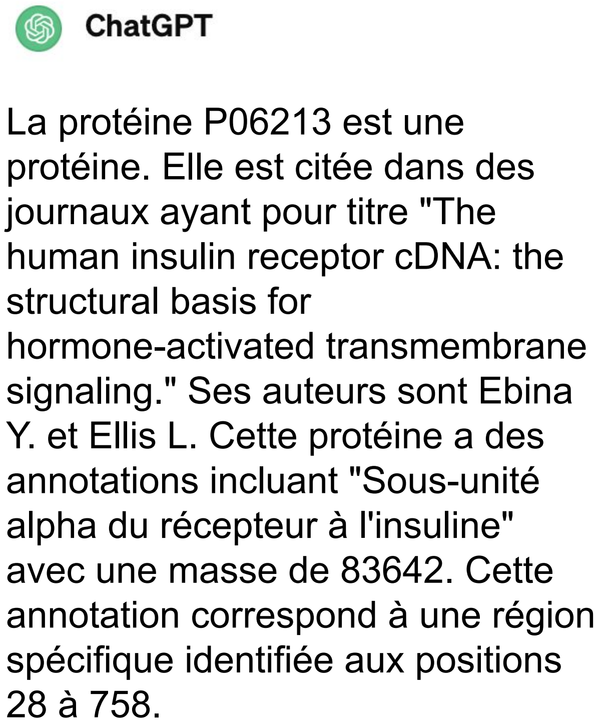
Practice … from KG to text

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


http://www.ldf.fi/service/rdf-grapher/P06213
rdf:type
up:Protein
up:citation
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.
up:author
Ellis L.
skos:exactMatch
pubmed:2859121
up:citation
citation:2983222
up:annotation
annotation:PRO_0000016687
rdf:type
up:Chain_Annotation
rdfs:comment
Insulin receptor subunit alpha
up:mass
83642
Datatype: http://www.w3.org/2001/XMLSchema#integer
up:range
range:22571007465304878tt28tt758
rdf:type
faldo:Region
faldo:begin
position:22571007465304878tt28
faldo:end
position:22571007465304878tt758
up:annotation
annotation:PRO_0000016689
up:annotation
annotation:VAR_015924

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

Practice … from KG to text



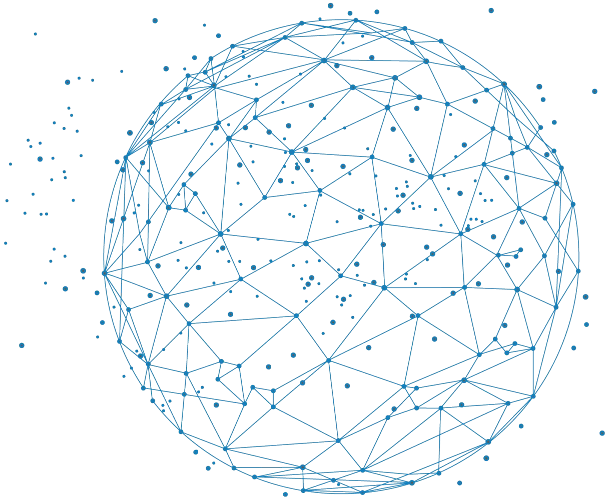
ChatGPT

La protéine P06213 est une protéine. Elle est citée dans des journaux ayant pour titre "The human insulin receptor cDNA: the structural basis for hormone-activated transmembrane signaling." Ses auteurs sont Ebina Y. et Ellis L. Cette protéine a des annotations incluant "Sous-unité alpha du récepteur à l'insuline" avec une masse de 83642. Cette annotation correspond à une région spécifique identifiée aux positions 28 à 758.

!! inaccurate positions

!! no link to the pubmed paper

Querying with graph patterns



Triple patterns (TP)

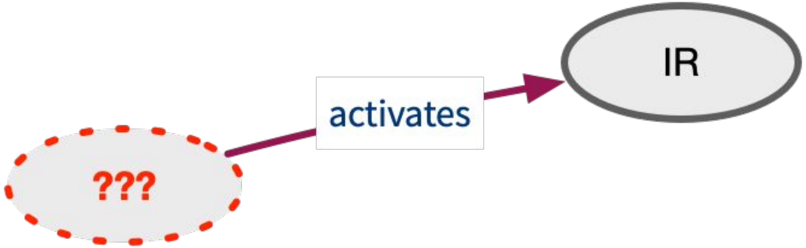
SPARQL is the W3C language to query multiple data sources expressed in RDF.

The principle consists in defining a graph pattern to be matched against an RDF graph.

Definition

Triple Patterns (TPs) are like RDF triples except that each of the subject, predicate and object may be a **variable**. Variables are prefixed with a **?** .

Give me all known activators of IR ?

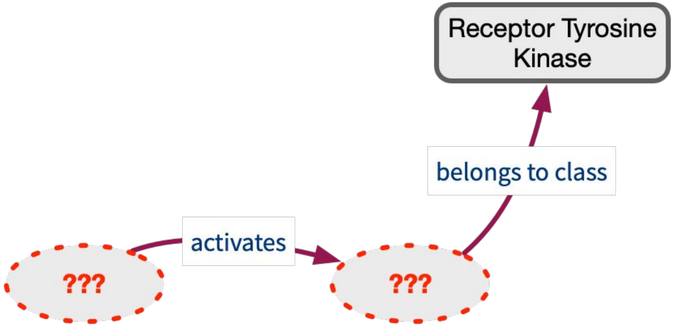


```
?x    ns:activates    ns:IR .
```

Basic graph patterns (BGP)

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

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



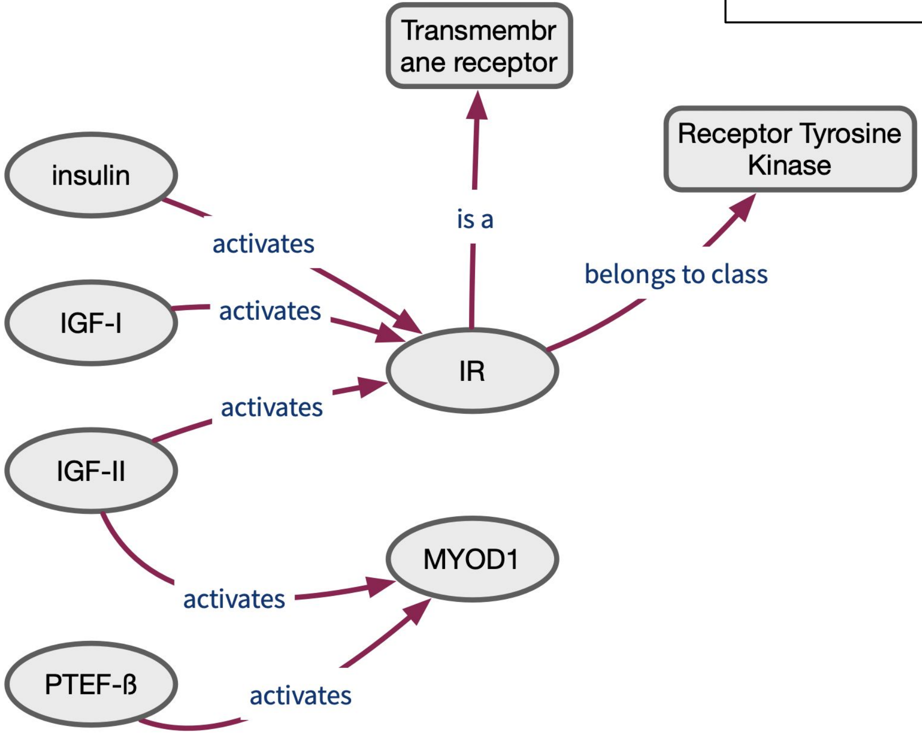
Definition

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

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

SPARQL query evaluation

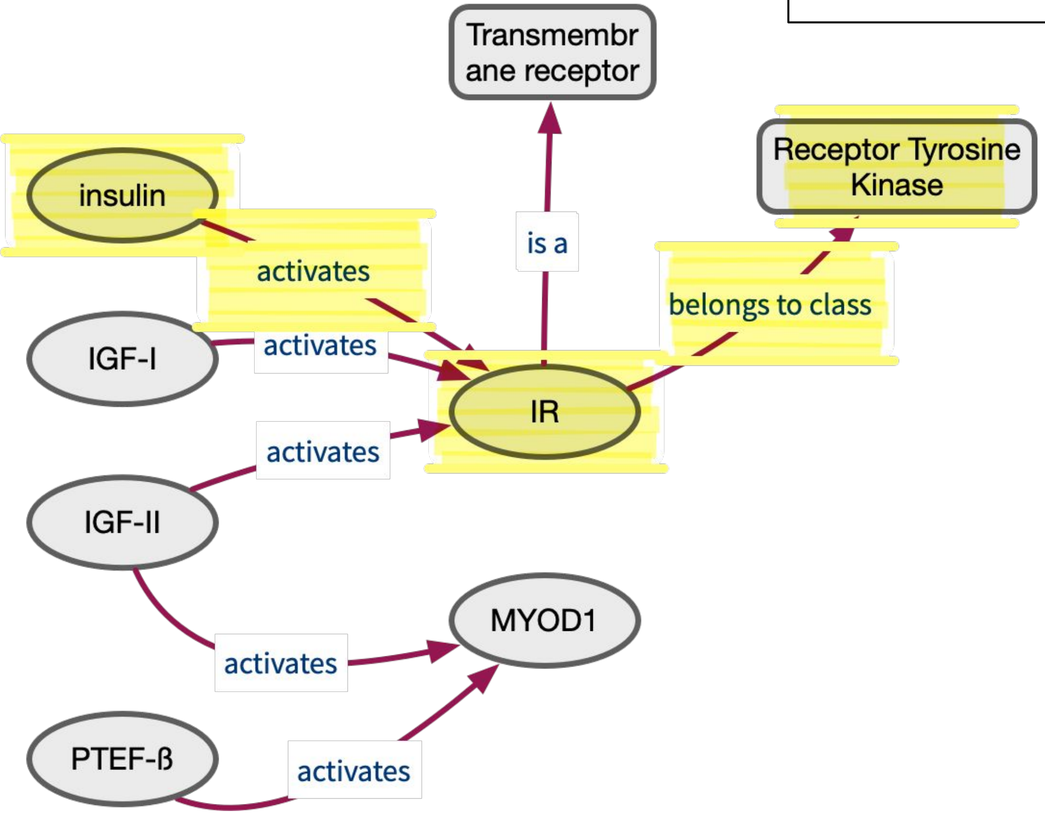
```
?x  ns:activates     ?y .
?y  ns:belongs_to_class ns:ReceptorTyrosineKinase .
```



?x	?y

SPARQL query evaluation

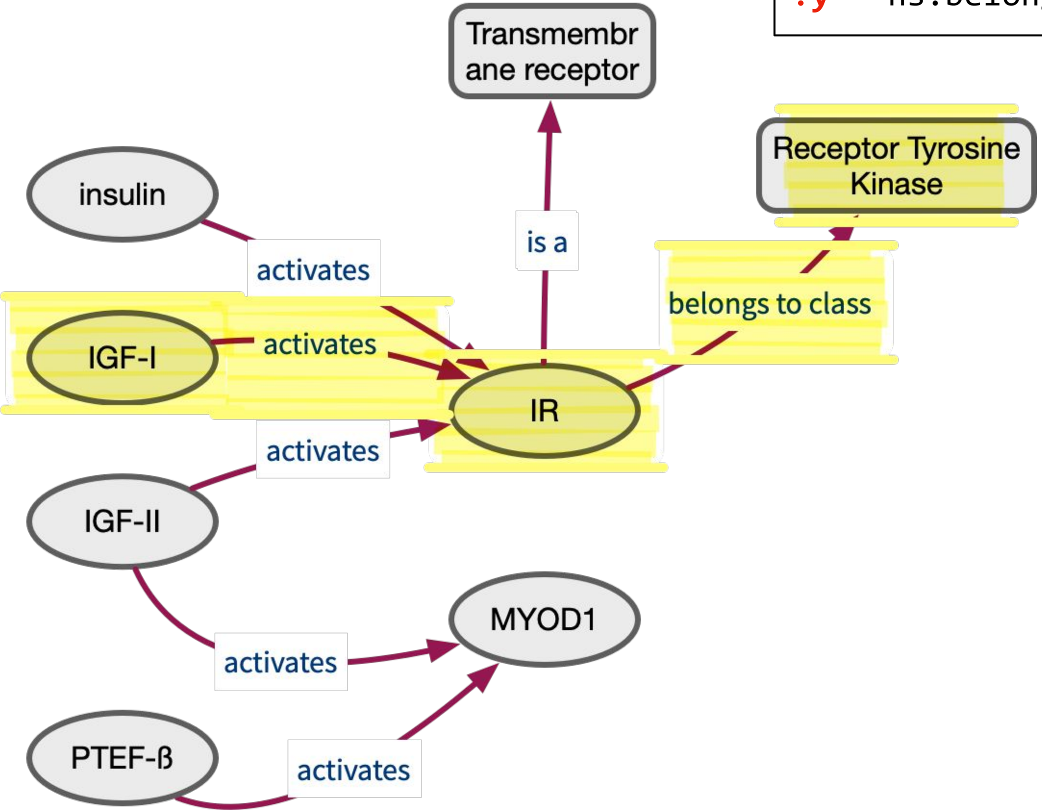
```
?x   ns:activates     ?y .
?y   ns:belongs_to_class ns:ReceptorTyrosineKinase .
```



?x	?y
insulin	IR

SPARQL query evaluation

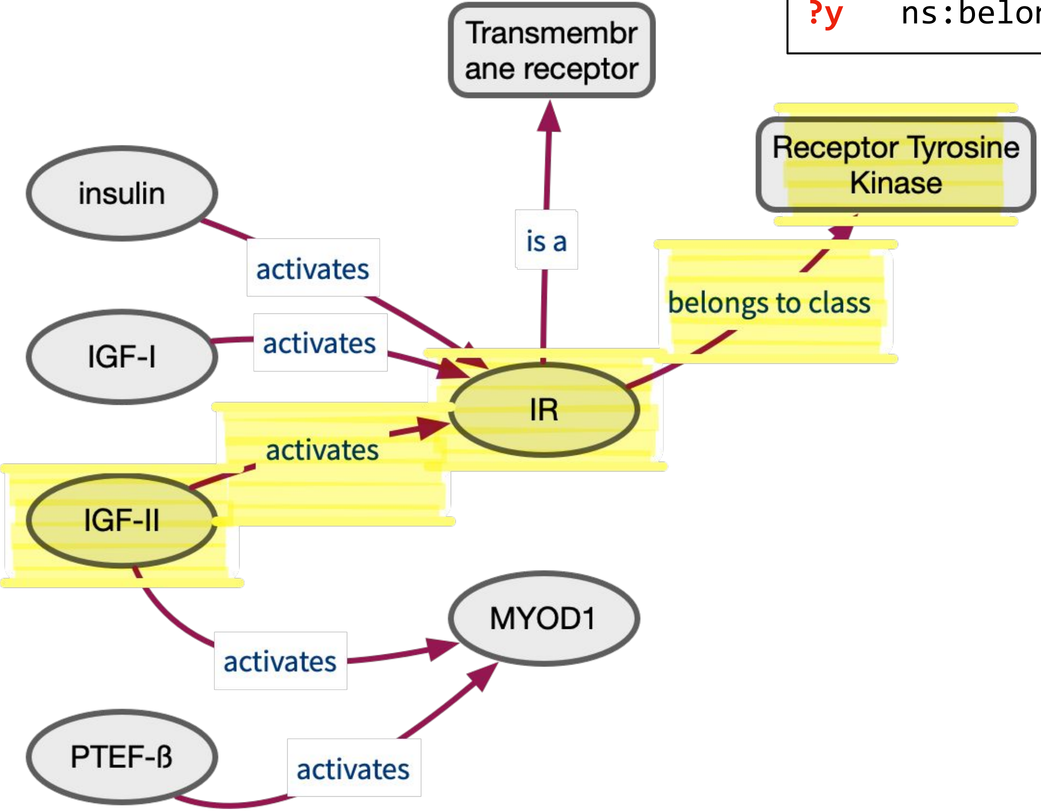
```
?x   ns:activates     ?y .
?y   ns:belongs_to_class ns:ReceptorTyrosineKinase .
```



?x	?y
IGF-I	IR
insulin	IR

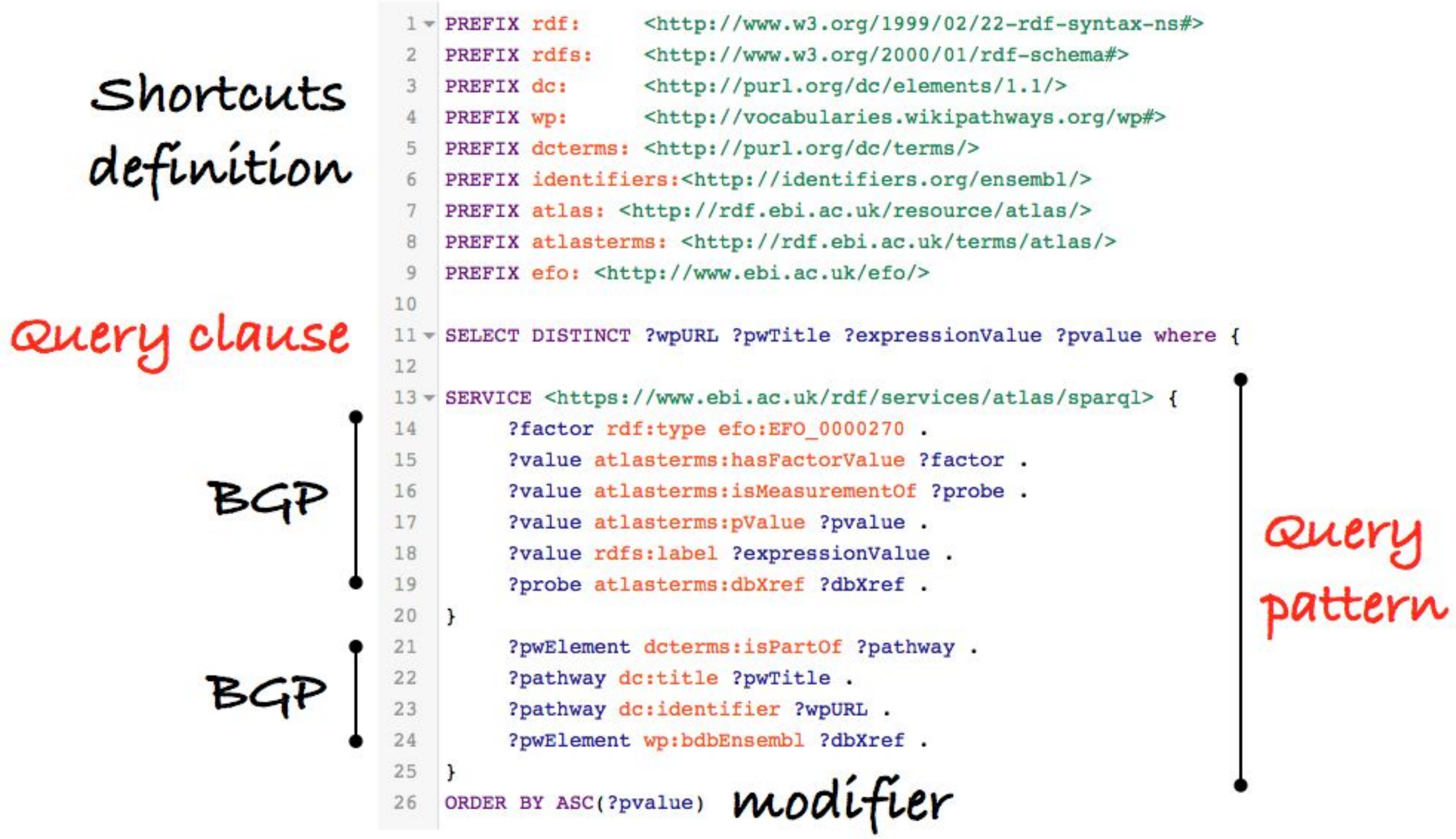
SPARQL query evaluation

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



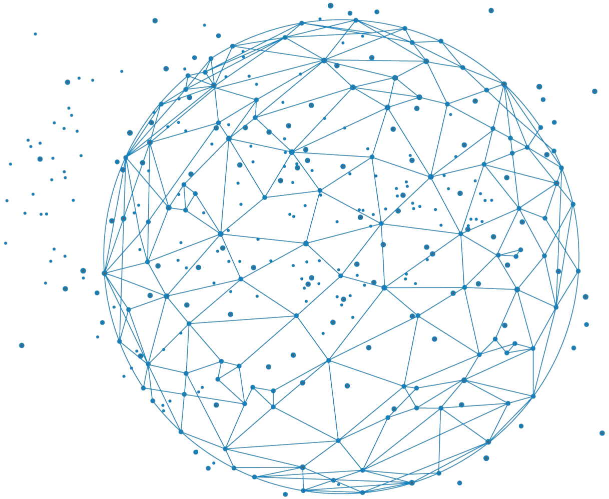
?x	?y
IGF-II	IR
IGF-I	IR
insulin	IR

Sample SPARQL query



```
PREFIX rdf:     <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:    <http://www.w3.org/2000/01/rdf-schema#>
PREFIX dc:      <http://purl.org/dc/elements/1.1/>
PREFIX wp:      <http://vocabularies.wikipathways.org/wp#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX identifiers:<http://identifiers.org/ensembl/>
PREFIX atlas: <http://rdf.ebi.ac.uk/resource/atlas/>
PREFIX atlasterms: <http://rdf.ebi.ac.uk/terms/atlas/>
PREFIX efo: <http://www.ebi.ac.uk/efo/>

SELECT DISTINCT ?wpURL ?pwTitle ?expressionValue ?pvalue where {

SERVICE <https://www.ebi.ac.uk/rdf/services/atlas/sparql> {
     ?factor rdf:type efo:EFO_0000270 .
     ?value atlasterms:hasFactorValue ?factor .
     ?value atlasterms:isMeasurementOf ?probe .
     ?value atlasterms:pValue ?pvalue .
     ?value rdfs:label ?expressionValue .
     ?probe atlasterms:dbXref ?dbXref .
}
     ?pwElement dcterms:isPartOf ?pathway .
     ?pathway dc:title ?pwTitle .
     ?pathway dc:identifier ?wpURL .
     ?pwElement wp:bdbEnsembl ?dbXref .
}
ORDER BY ASC(?pvalue)
```

Reasoning with Knowledge graphs



Ontologies in use: use the MeSH to query PubMed

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

Ontologies in use: the Gene Ontology

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)?

What is an ontology ?

Definitions

"[...] an explicit specification of a conceptualization" (Gruber, 1993)

"[...] a formal specification of a shared conceptualization" (Borst, 1997)

Explicit → a machine does not spontaneously "understand", "infer" or "reason"

Conceptualization → a knowledge model aimed at reducing the complexity (generalizing) real facts

Formal → reasoning mechanisms must be correct for reliable deductions

Shared → domain knowledge result from the consensus of expert communities

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.

Transmembrane receptor

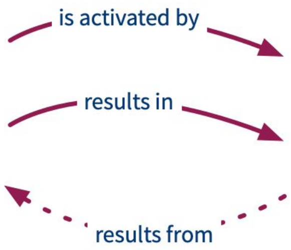
Receptor Tyrosine Kinase

Regulation of glucose homeostasis

Functional process

Clinical manifestation

How these concepts are related together ?

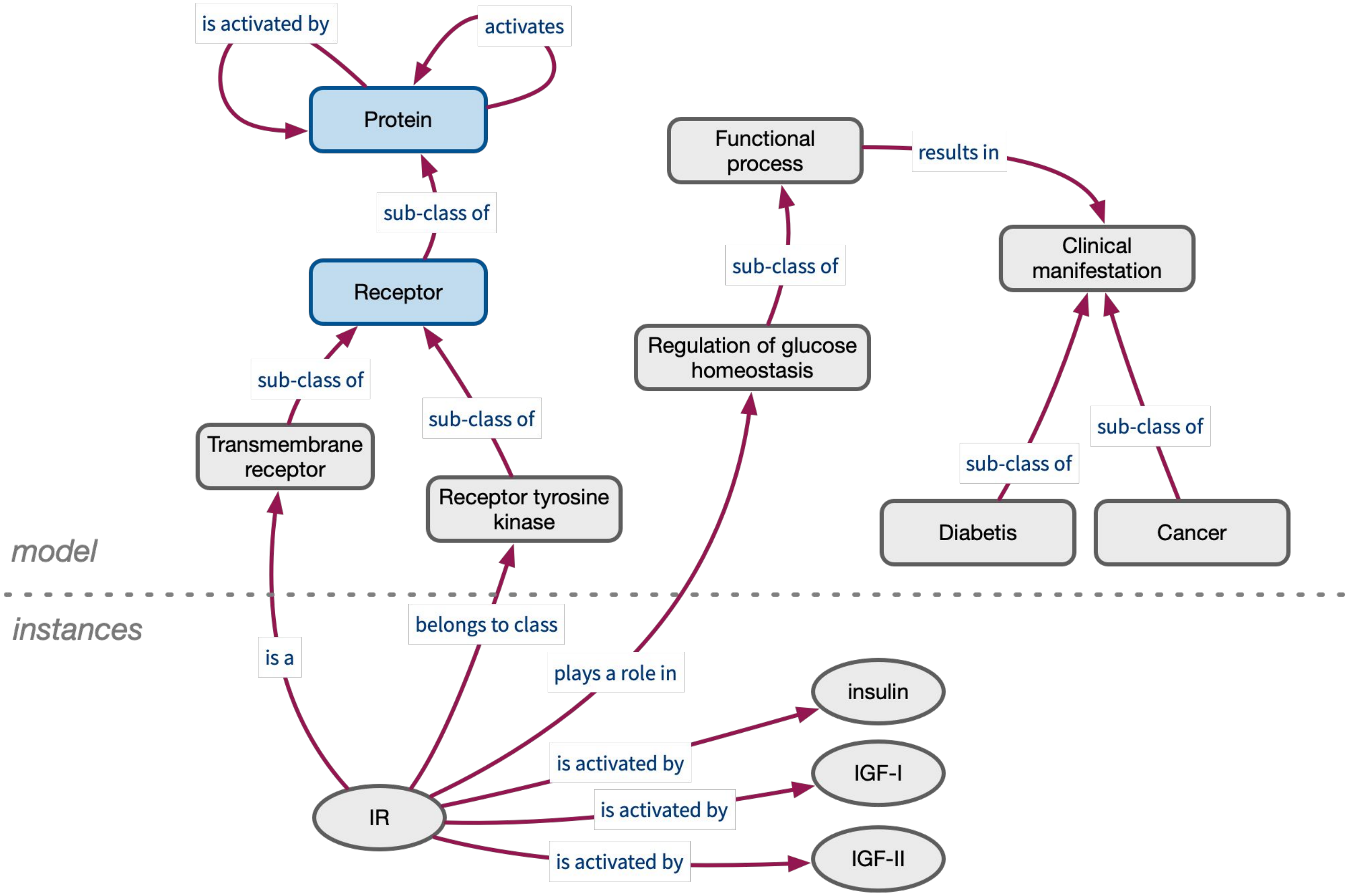

is activated by
results in
results from

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
Clinical manifestation
sub-class of
Regulation of glucose homeostasis
sub-class of
sub-class of
Diabetis
Cancer
model
instances
is a
belongs to class
plays a role in
insulin
is activated by
IGF-I
is activated by
IGF-II
is activated by
IR

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

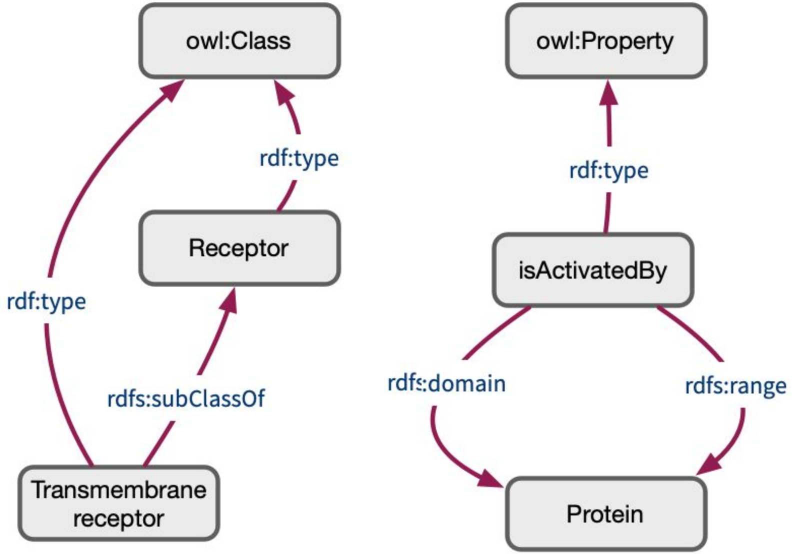
Resources may be classified into groups called **classes**. The members of a class are known as **instances** of the class. The rdf:**type** property is used to state that a resource is an instance of a class (« is a » relation).

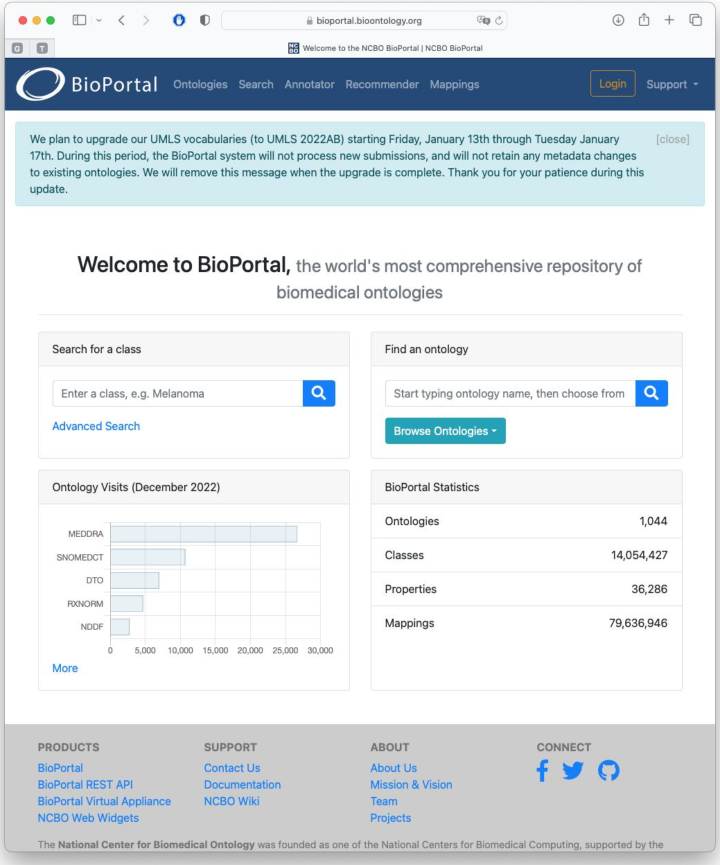
Defining ontologies

- 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 .
```




bioportal.bioontology.org
Welcome to the NCBO BioPortal | NCBO BioPortal
BioPortal
Ontologies
Search
Annotator
Recommender
Mappings
Login
Support

We plan to upgrade our UMLS vocabularies (to UMLS 2022AB) starting Friday, January 13th through Tuesday January 17th. During this period, the BioPortal system will not process new submissions, and will not retain any metadata changes to existing ontologies. We will remove this message when the upgrade is complete. Thank you for your patience during this update. [close]

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

Browse Ontologies

Ontology Visits (December 2022)

BioPortal Statistics

Ontologies	1,044
Classes	14,054,427
Properties	36,286
Mappings	79,636,946

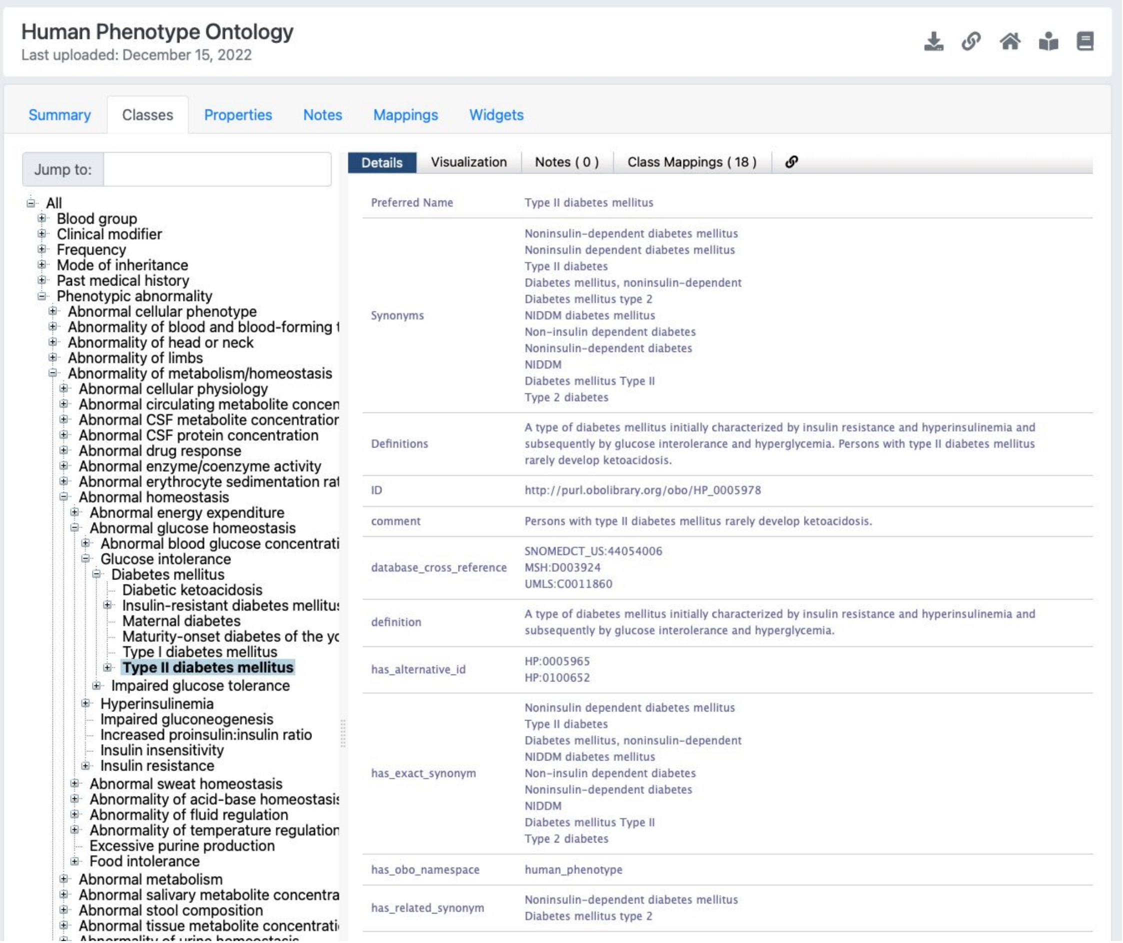
PRODUCTS
- BioPortal
- BioPortal REST API
- BioPortal Virtual Appliance
- NCBO Web Widgets

SUPPORT
- Contact Us
- Documentation
- NCBO Wiki

ABOUT
- About Us
- Mission & Vision
- Team
- Projects

CONNECT

The **National Center for Biomedical Ontology** was founded as one of the National Centers for Biomedical Computing, supported by the



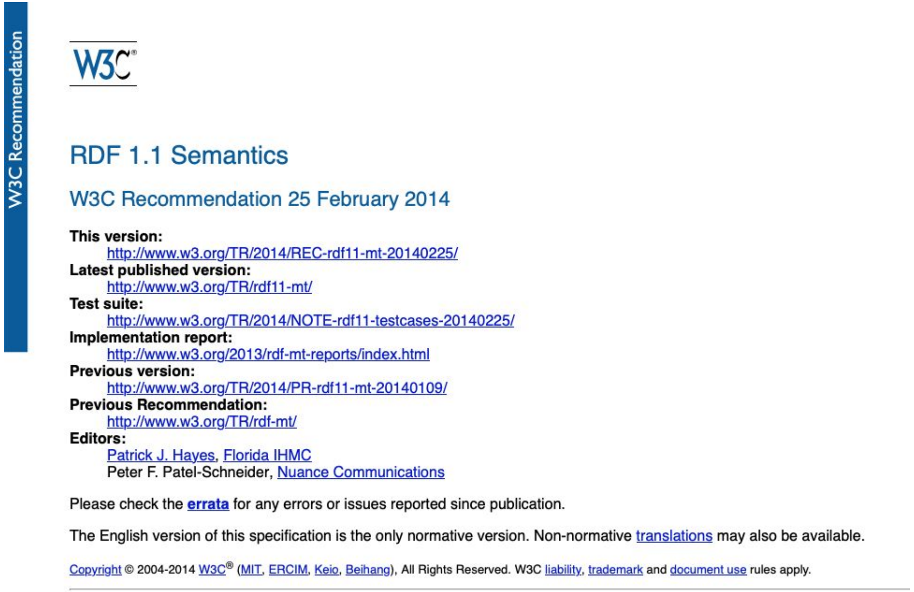
Human Phenotype Ontology

Last uploaded: December 15, 2022

- All
  - Blood group
  - Clinical modifier
  - Frequency
  - Mode of inheritance
  - Past medical history
  - Phenotypic abnormality
    - Abnormal cellular phenotype
    - Abnormality of blood and blood-forming
    - Abnormality of head or neck
    - Abnormality of limbs
    - Abnormality of metabolism/homeostasis
      - Abnormal cellular physiology
      - Abnormal circulating metabolite concen
      - Abnormal CSF metabolite concentratior
      - Abnormal CSF protein concentration
      - Abnormal drug response
      - Abnormal enzyme/coenzyme activity
      - Abnormal erythrocyte sedimentation ra
      - Abnormal homeostasis
        - Abnormal energy expenditure
        - Abnormal glucose homeostasis
          - Abnormal blood glucose concentrati
          - Glucose intolerance
            - Diabetes mellitus
              - Diabetic ketoacidosis
              - Insulin-resistant diabetes mellitu
              - Maternal diabetes
              - Maturity-onset diabetes of the yo
              - Type I diabetes mellitus
              - **Type II diabetes mellitus**
            - Impaired glucose tolerance
          - Hyperinsulinemia
          - Impaired gluconeogenesis
          - Increased proinsulin:insulin ratio
          - Insulin insensitivity
          - Insulin resistance
        - Abnormal sweat homeostasis
        - Abnormality of acid-base homeostasis
        - Abnormality of fluid regulation
        - Abnormality of temperature regulation
        - Excessive purine production
        - Food intolerance
      - Abnormal metabolism
      - Abnormal salivary metabolite concentra
      - Abnormal stool composition
      - Abnormal tissue metabolite concentrati

Preferred Name	Type II diabetes mellitus
Synonyms	Noninsulin-dependent diabetes mellitus Noninsulin dependent diabetes mellitus Type II diabetes Diabetes mellitus, noninsulin-dependent Diabetes mellitus type 2 NIDDM diabetes mellitus Non-insulin dependent diabetes Noninsulin-dependent diabetes NIDDM Diabetes mellitus Type II Type 2 diabetes
Definitions	A type of diabetes mellitus initially characterized by insulin resistance and hyperinsulinemia and subsequently by glucose interolerance and hyperglycemia. Persons with type II diabetes mellitus rarely develop ketoacidosis.
ID	http://purl.obolibrary.org/obo/HP_0005978
comment	Persons with type II diabetes mellitus rarely develop ketoacidosis.
database_cross_reference	SNOMEDCT_US:44054006 MSH:D003924 UMLS:C0011860
definition	A type of diabetes mellitus initially characterized by insulin resistance and hyperinsulinemia and subsequently by glucose interolerance and hyperglycemia.
has_alternative_id	HP:0005965 HP:0100652
has_exact_synonym	Noninsulin dependent diabetes mellitus Type II diabetes Diabetes mellitus, noninsulin-dependent NIDDM diabetes mellitus Non-insulin dependent diabetes Noninsulin-dependent diabetes NIDDM Diabetes mellitus Type II Type 2 diabetes
has_obo_namespace	human_phenotype
has_related_synonym	Noninsulin-dependent diabetes mellitus Diabetes mellitus type 2

Reasoning with RDFS-Entailments



W3C Recommendation

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

Please check the **errata** for any errors or issues reported since publication.

The English version of this specification is the only normative version. Non-normative translations may also be available.

Copyright © 2004-2014 W3C® (MIT, ERCIM, Keio, Beihang), All Rights Reserved. W3C liability, trademark and document use rules apply.

Inference rules to produce new logical facts, or to check for logical soundness (satisfiability)

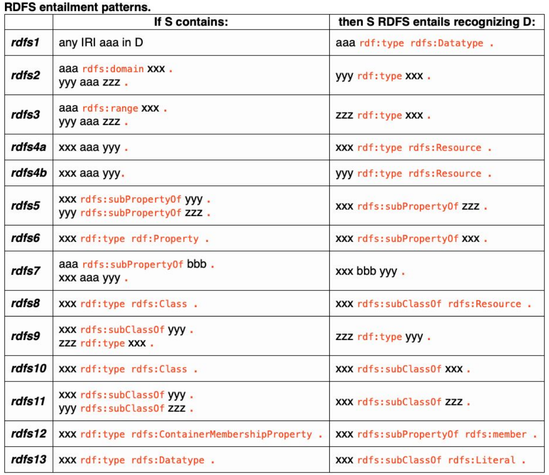
Deduce the multiple **types** of an entity based on **class hierarchies**

Deduce the **types** of entities exploiting the definition of **relations**

… more possibilities with OWL (Web Ontology Language) and Description Logics (DL)

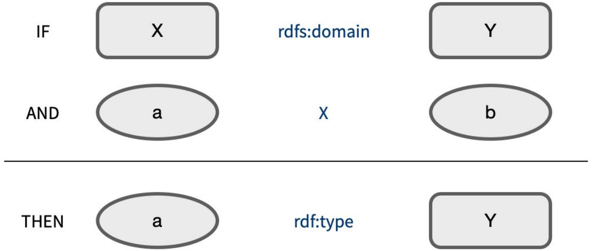
Reasoning with RDFS-Entailments

From the specification of a "link", how to infer the type of a linked entity ?



RDFS entailment patterns.

	If S contains:	then S RDFS entails recognizing D:
rdfs1	any IRI aaa in D	aaa `rdf:type rdfs:Datatype .`
rdfs2	aaa `rdfs:domain` xxx . yyy aaa zzz .	yyy `rdf:type` xxx .
rdfs3	aaa `rdfs:range` xxx . yyy aaa zzz .	zzz `rdf:type` xxx .
rdfs4a	xxx aaa yyy .	xxx `rdf:type rdfs:Resource .`
rdfs4b	xxx aaa yyy.	yyy `rdf:type rdfs:Resource .`
rdfs5	xxx `rdfs:subPropertyOf` yyy . yyy `rdfs:subPropertyOf` zzz .	xxx `rdfs:subPropertyOf` zzz .
rdfs6	xxx `rdf:type rdf:Property .`	xxx `rdfs:subPropertyOf` xxx .
rdfs7	aaa `rdfs:subPropertyOf` bbb . xxx aaa yyy .	xxx bbb yyy .
rdfs8	xxx `rdf:type rdfs:Class .`	xxx `rdfs:subClassOf rdfs:Resource .`
rdfs9	xxx `rdfs:subClassOf` yyy . zzz `rdf:type` xxx .	zzz `rdf:type` yyy .
rdfs10	xxx `rdf:type rdfs:Class .`	xxx `rdfs:subClassOf` xxx .
rdfs11	xxx `rdfs:subClassOf` yyy . yyy `rdfs:subClassOf` zzz .	xxx `rdfs:subClassOf` zzz .
rdfs12	xxx `rdf:type rdfs:ContainerMembershipProperty .`	xxx `rdfs:subPropertyOf rdfs:member .`
rdfs13	xxx `rdf:type rdfs:Datatype .`	xxx `rdfs:subClassOf rdfs:Literal .`

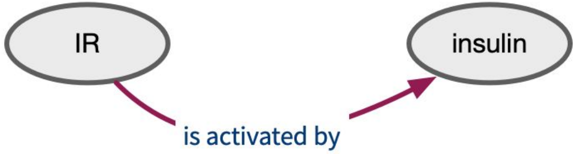


IF X rdfs:domain Y

AND a X b

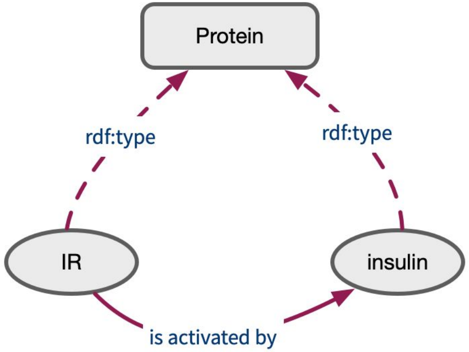
THEN a rdf:type Y

rdfs2



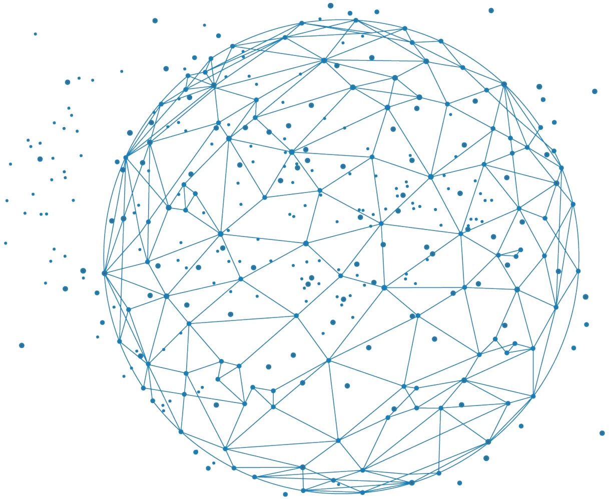
rdfs2, rdfs3

a = IR
b = insulin
X = is activated by
Y = Protein



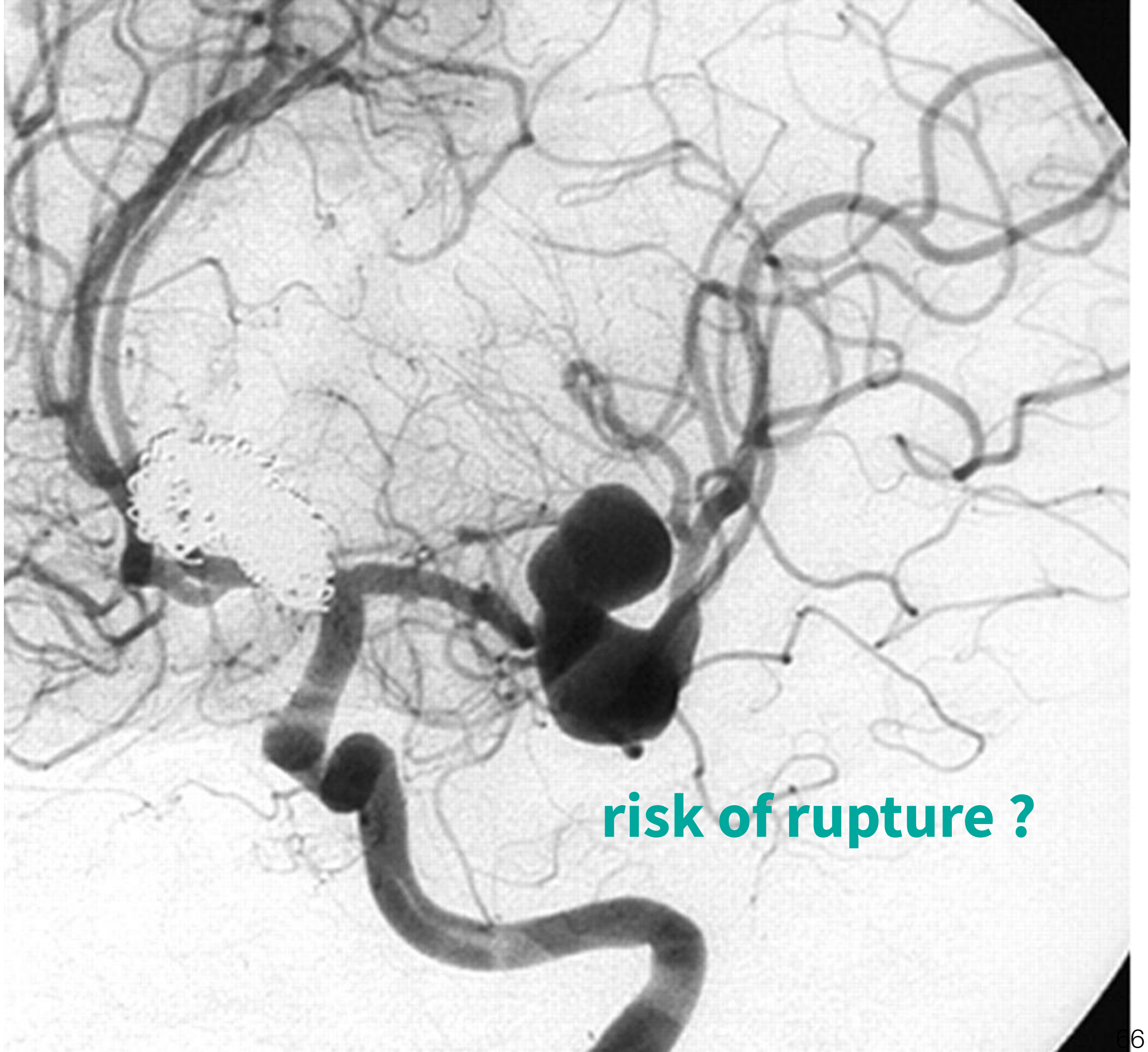
https://www.w3.org/TR/rdf11-mt/

Application with multi-scale data



ICAN cohort : 34 univ. hospitals / 3400 subjects

- 3000 MRIs
- 800 whole genomes (under processing)


risk of rupture ?

CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES

l'institut du thorax

Maxime Folschette

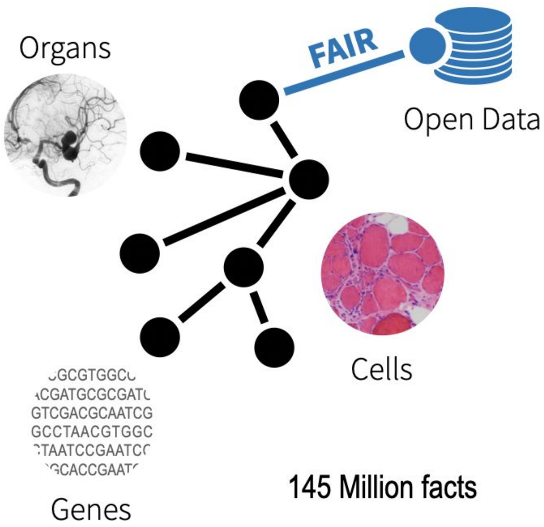
Kirsley Chennen

Bridging imaging-omics-clinical data: **INEX-MED**

LS2N · l'institut du thorax · igbmc · ICUBE · CENTRE HOSPITALIER UNIVERSITAIRE DE NANTES · INSTITUT DE MYOLOGIE · ifb · FLi · FRANCE GÉNOMIQUE

INEX-MED

Intracranial aneurysms (3.2 % world population)
Congenital myopathies (rare diseases)



145 Million facts

Multi-scale « Knowledge Graph »

Select
Combine
Reuse
(Share)

Ontologies
SPARQL queries

Predict
Explain

Statistics
Reasoning
Machine learning

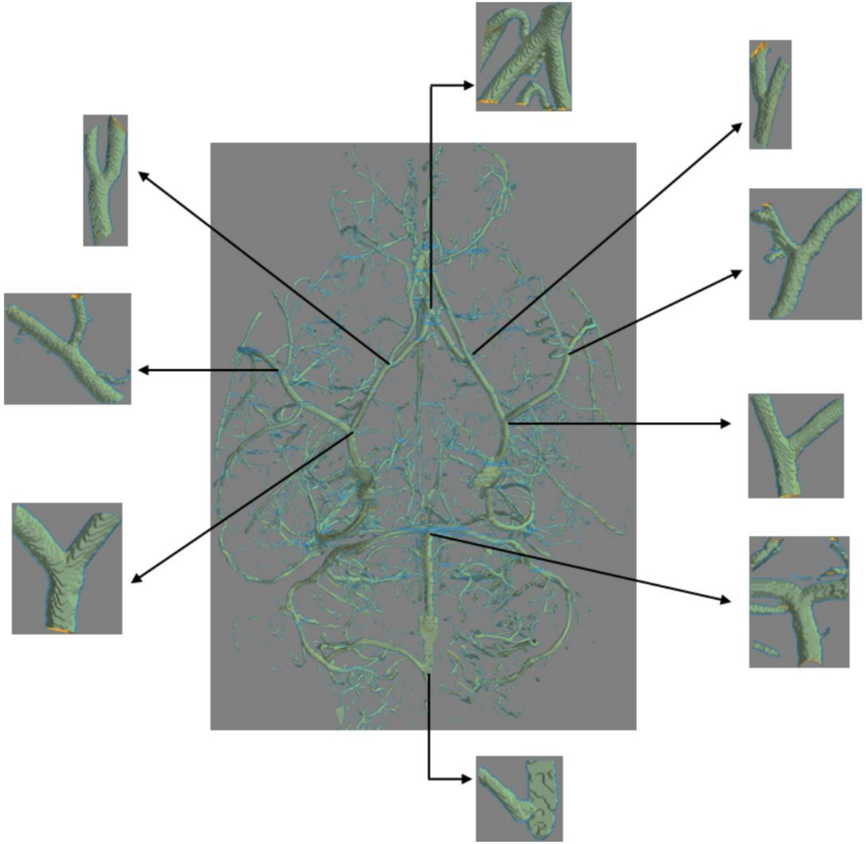
@Nantes

@Strasbourg

Association between imaging phenotypes - « omics » signatures ?
Patients with high/low aneurysm rupture risk ?

Quantifications of cerebral artery bifurcations

Anass Nouri, Florent Autrusseau

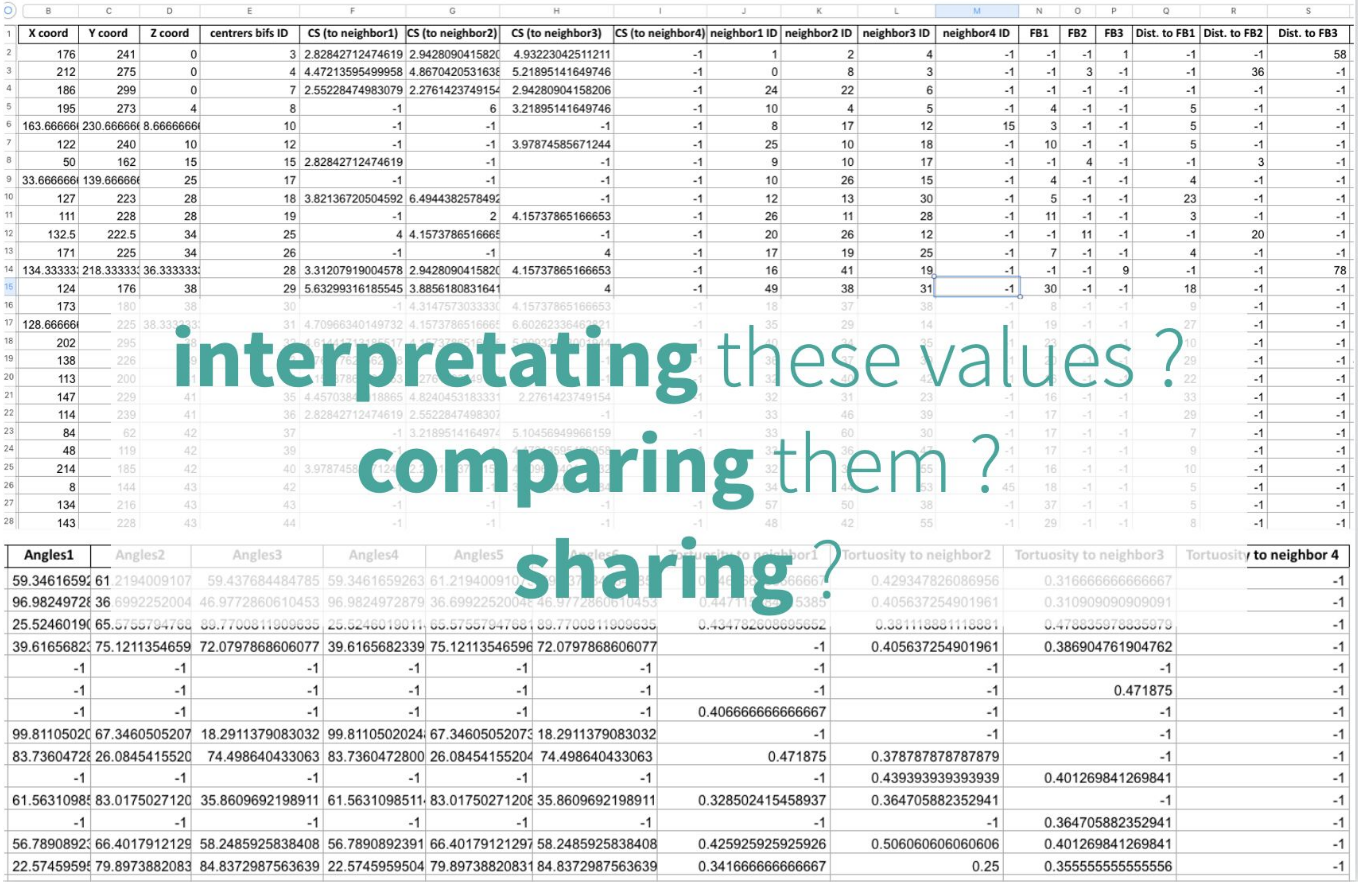


- distances ?
- angles ?
- radius ?
- tortuosity ?

Anass Nouri

Quantifications of cerebral artery bifurcations

Anass Nouri, Florent Autrusseau



	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	X coord	Y coord	Z coord	centrers bifs ID	CS (to neighbor1)	CS (to neighbor2)	CS (to neighbor3)	CS (to neighbor4)	neighbor1 ID	neighbor2 ID	neighbor3 ID	neighbor4 ID	FB1	FB2	FB3	Dist. to FB1	Dist. to FB2	Dist. to FB3
2	176	241	0	3	2.82842712474619	2.9428090415820	4.93223042511211	-1	1	2	4	-1	-1	-1	1	-1	-1	58
3	212	275	0	4	4.47213595499958	4.8670420531638	5.21895141649746	-1	0	8	3	-1	-1	3	-1	-1	36	-1
4	186	299	0	7	2.55228474983079	2.2761423749154	2.94280904158206	-1	24	22	6	-1	-1	-1	-1	-1	-1	-1
5	195	273	4	8	-1	6	3.21895141649746	-1	10	4	5	-1	4	-1	-1	5	-1	-1
6	163.66666	230.66666	8.6666666	10	-1	-1	-1	-1	8	17	12	15	3	-1	-1	5	-1	-1
7	122	240	10	12	-1	-1	3.97874585671244	-1	25	10	18	-1	10	-1	-1	5	-1	-1
8	50	162	15	15	2.82842712474619	-1	-1	-1	9	10	17	-1	-1	4	-1	-1	3	-1
9	33.666666	139.66666	25	17	-1	-1	-1	-1	10	26	15	-1	4	-1	-1	4	-1	-1
10	127	223	28	18	3.82136720504592	6.4944382578492	-1	-1	12	13	30	-1	5	-1	-1	23	-1	-1
11	111	228	28	19	-1	2	4.15737865166653	-1	26	11	28	-1	11	-1	-1	3	-1	-1
12	132.5	222.5	34	25	4	4.1573786516665	-1	-1	20	26	12	-1	-1	11	-1	-1	20	-1
13	171	225	34	26	-1	-1	4	-1	17	19	25	-1	7	-1	-1	4	-1	-1
14	134.33333	218.33333	36.333333	28	3.31207919004578	2.9428090415820	4.15737865166653	-1	16	41	19	-1	-1	-1	9	-1	-1	78
15	124	176	38	29	5.63299316185545	3.8856180831641	4	-1	49	38	31	-1	30	-1	-1	18	-1	-1
16	173	180	38	30	-1	4.3147573033330	4.15737865166653	-1	18	37	38	-1	8	-1	-1	9	-1	-1
17	128.66666	225	38.333333	31	4.70966340149732	4.1573786516665	6.60262336462821	-1	35	29	14	-1	19	-1	-1	27	-1	-1
18	202	295	38	[illegible]	4.61441713185517	[illegible]	[illegible]	-1	40	34	35	-1	23	-1	-1	10	-1	-1
19	138	226	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-1	36	37	[illegible]	-1	[illegible]	-1	-1	29	-1	-1
20	113	200	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-1	32	40	42	-1	[illegible]	-1	-1	22	-1	-1
21	147	229	41	35	[illegible]	4.8240453183331	2.2761423749154	-1	32	31	23	-1	16	-1	-1	33	-1	-1
22	114	239	41	36	2.82842712474619	2.5522847498307	-1	-1	33	46	39	-1	17	-1	-1	29	-1	-1
23	84	62	42	37	-1	3.2189514164974	5.10456949966159	-1	33	60	30	-1	17	-1	-1	7	-1	-1
24	48	119	42	39	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	36	47	-1	17	-1	-1	9	-1	-1
25	214	185	42	40	[illegible]	[illegible]	[illegible]	[illegible]	32	[illegible]	55	-1	16	-1	-1	10	-1	-1
26	8	144	43	42	[illegible]	[illegible]	[illegible]	[illegible]	34	[illegible]	53	45	18	-1	-1	5	-1	-1
27	134	216	43	43	-1	-1	-1	-1	57	50	38	-1	37	-1	-1	5	-1	-1
28	143	228	43	44	-1	-1	-1	-1	48	42	55	-1	29	-1	-1	8	-1	-1

Angles1	Angles2	Angles3	Angles4	Angles5	Angles6	Tortuosity to neighbor1	Tortuosity to neighbor2	Tortuosity to neighbor3	Tortuosity to neighbor 4
59.34616592	61.2194009107	59.437684484785	59.3461659263	61.21940091	[illegible]	[illegible]	0.429347826086956	0.316666666666667	-1
96.98249728	36.6992252004	46.9772860610453	96.9824972879	36.6992252004	46.9772860610453	[illegible]	0.405637254901961	0.310909090909091	-1
25.52460190	65.5755794768	89.7700811909635	25.52460190114	65.5755794768	89.7700811909635	0.434782608695652	0.381118881118881	0.478835978835979	-1
39.61656823	75.1211354659	72.0797868606077	39.6165682339	75.12113546596	72.0797868606077	-1	0.405637254901961	0.386904761904762	-1
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	0.471875	-1
-1	-1	-1	-1	-1	-1	0.406666666666667	-1	-1	-1
99.81105020	67.3460505207	18.2911379083032	99.81105020240	67.34605052073	18.2911379083032	-1	-1	-1	-1
83.73604728	26.0845415520	74.498640433063	83.7360472800	26.08454155204	74.498640433063	0.471875	0.378787878787879	-1	-1
-1	-1	-1	-1	-1	-1	-1	0.439393939393939	0.401269841269841	-1
61.56310985	83.0175027120	35.8609692198911	61.56310985114	83.01750271208	35.8609692198911	0.328502415458937	0.364705882352941	-1	-1
-1	-1	-1	-1	-1	-1	-1	-1	0.364705882352941	-1
56.78908923	66.4017912129	58.2485925838408	56.7890892391	66.40179121297	58.2485925838408	0.425925925925926	0.506060606060606	0.401269841269841	-1
22.57459595	79.8973882083	84.8372987563639	22.5745959504	79.89738820831	84.8372987563639	0.341666666666667	0.25	0.355555555555556	-1

interpretating these values ?

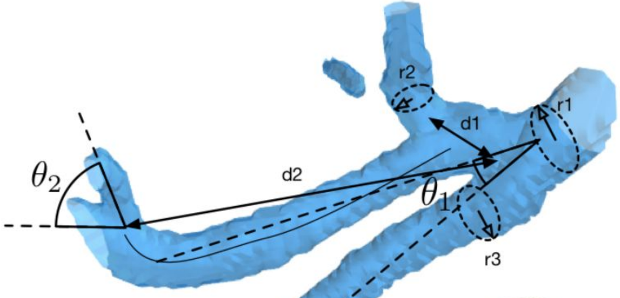
comparing them ?

sharing ?

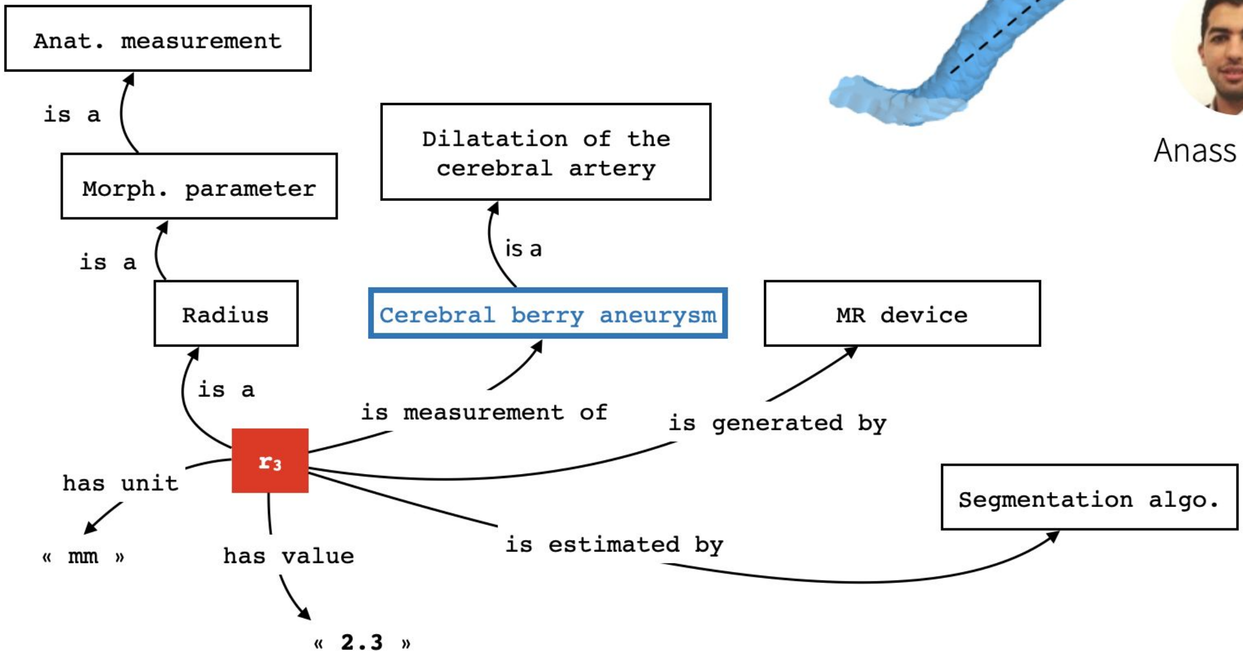
d_1, d_2 : **distances** between **bifurcations**

r_1, r_2, r_3 : **radius** of vessels

θ_1, θ_2 : **angles** of bifurcations



Anass Nouri



Anat. measurement

is a

Morph. parameter

is a

Radius

Dilatation of the cerebral artery

is a

Cerebral berry aneurysm

MR device

Segmentation algo.

is a

is measurement of

is generated by

is estimated by

r_3

has unit

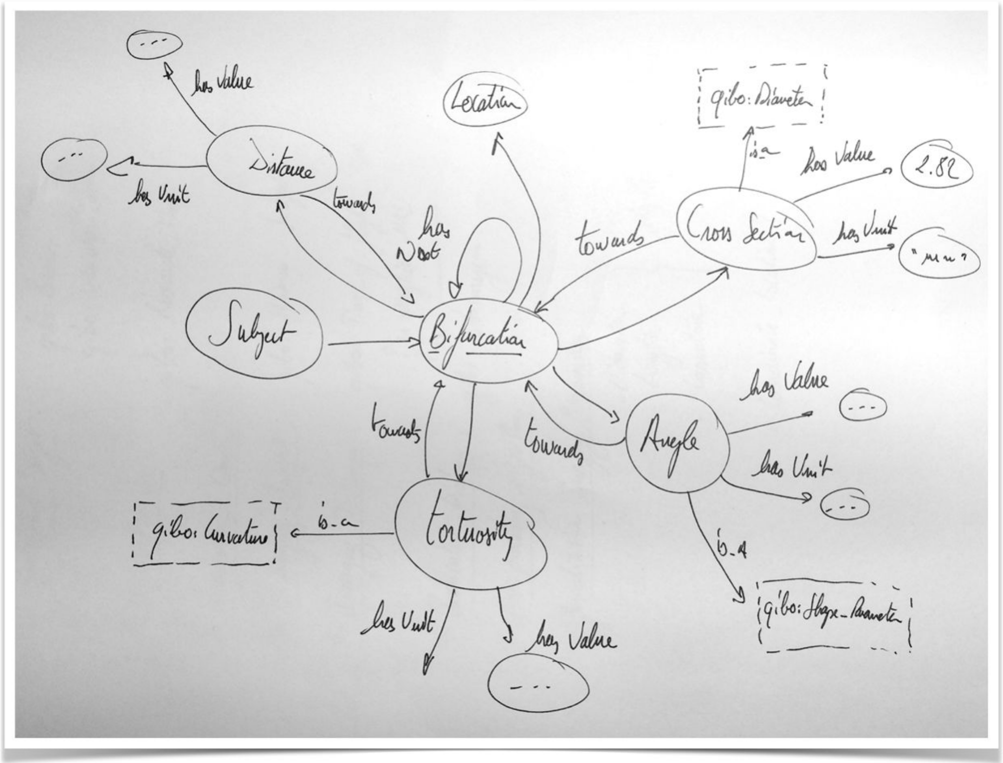
« mm »

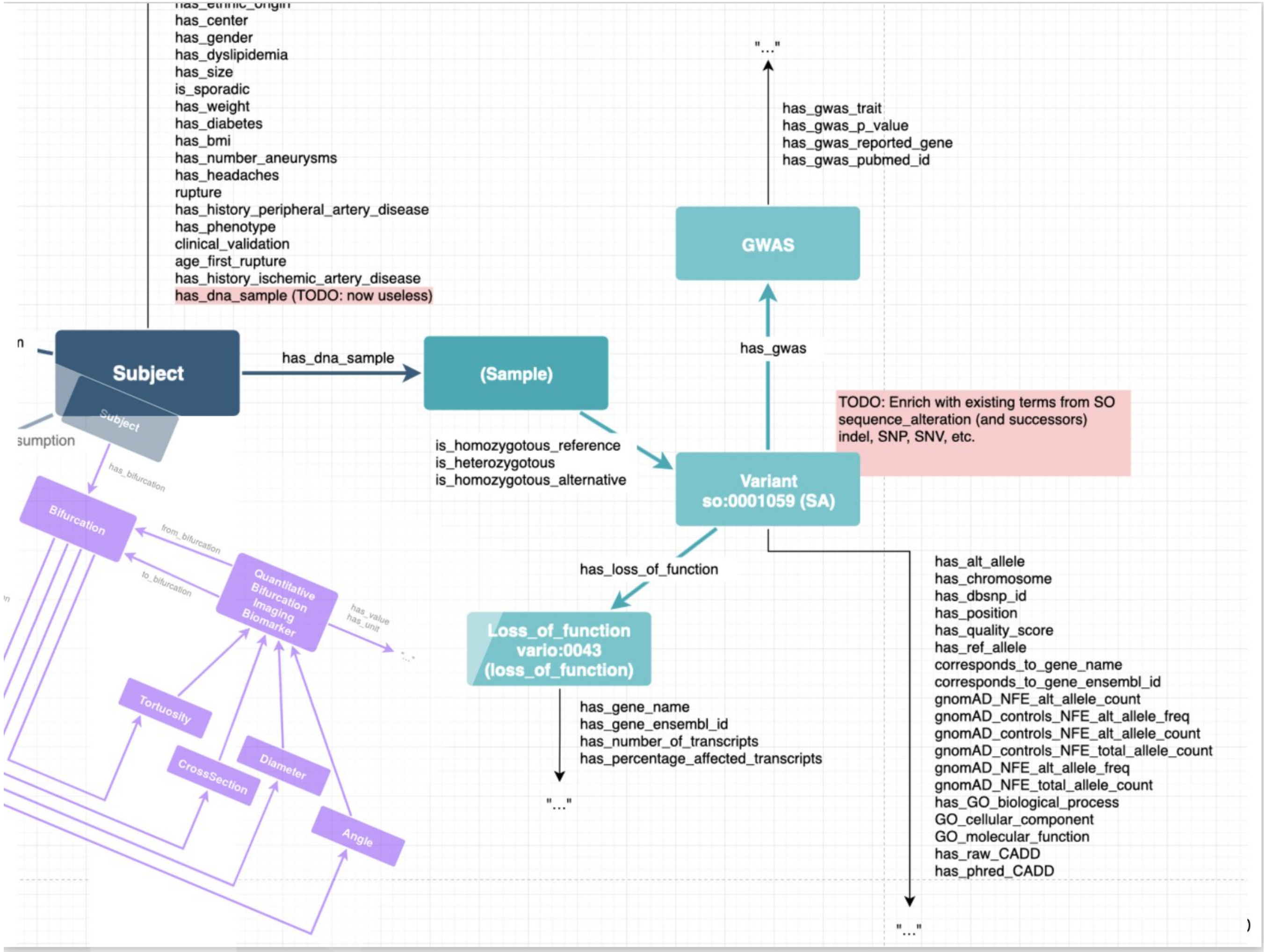
has value

« 2.3 »

BrainHack project

« From MS excel sheets to semantic bioimaging markers : representing and querying cerebral vascular measures with the QIBO ontology »



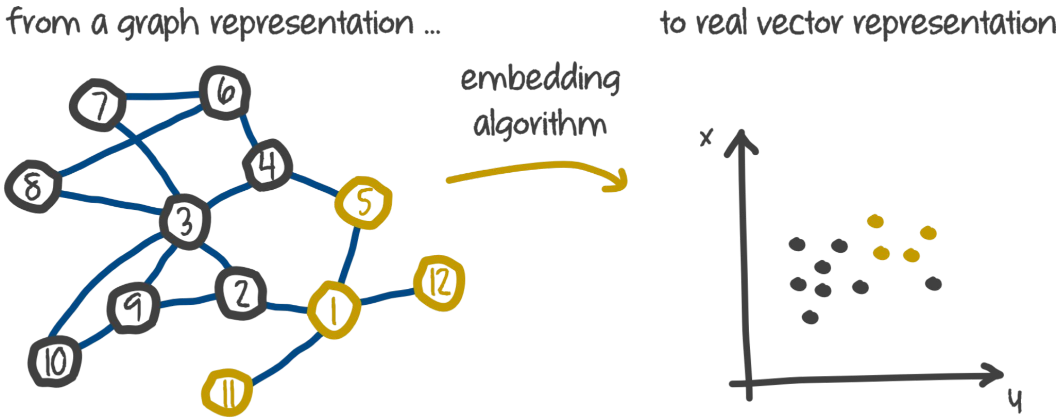

has_ethnic_origin
has_center
has_gender
has_dyslipidemia
has_size
is_sporadic
has_weight
has_diabetes
has_bmi
has_number_aneurysms
has_headaches
rupture
has_history_peripheral_artery_disease
has_phenotype
clinical_validation
age_first_rupture
has_history_ischemic_artery_disease
has_dna_sample (TODO: now useless)
Subject
has_dna_sample
(Sample)
is_homozygotous_reference
is_heterozygotous
is_homozygotous_alternative
Variant
so:0001059 (SA)
has_gwas
GWAS
has_gwas_trait
has_gwas_p_value
has_gwas_reported_gene
has_gwas_pubmed_id
"..."
TODO: Enrich with existing terms from SO sequence_alteration (and successors) indel, SNP, SNV, etc.
has_loss_of_function
Loss_of_function
vario:0043
(loss_of_function)
has_gene_name
has_gene_ensembl_id
has_number_of_transcripts
has_percentage_affected_transcripts
"..."
has_alt_allele
has_chromosome
has_dbsnp_id
has_position
has_quality_score
has_ref_allele
corresponds_to_gene_name
corresponds_to_gene_ensembl_id
gnomAD_NFE_alt_allele_count
gnomAD_controls_NFE_alt_allele_freq
gnomAD_controls_NFE_alt_allele_count
gnomAD_controls_NFE_total_allele_count
gnomAD_NFE_alt_allele_freq
gnomAD_NFE_total_allele_count
has_GO_biological_process
GO_cellular_component
GO_molecular_function
has_raw_CADD
has_phred_CADD
"..."
sumption
Subject
has_bifurcation
Bifurcation
from_bifurcation
to_bifurcation
Quantitative Bifurcation Imaging Biomarker
has_value
has_unit
"..."
Tortuosity
CrossSection
Diameter
Angle

A single query to combine multi-scale criteria

```
SELECT * WHERE {
  # Search variants
  ?subject ican:has_dna_sample ?sample .
  ?sample ican:is_heterozygotous ?variant .
  ?variant ican:has_chromosome "19" ;
           ican:has_position ?position ;
           ican:has_alt_allele ?alt_allele ;
           ican:has_ref_allele ?ref_allele .
  FILTER (?position > 10092337 && ?position < 10106407)
  # dbSNP identifier
  OPTIONAL { ?variant ican:has_dbsnp_id ?dbSNPid . }
  # First aneurysm rupture: age and localisation
  OPTIONAL { ?subject ican:age_first_rupture ?ageRupt . }
  OPTIONAL { ?subject ican:loc_first_rupture ?locRupt . }
}
```

KG Application in ML

Graph embeddings approach


from a graph representation ...
embedding algorithm
to real vector representation
7
6
8
4
5
3
2
9
1
12
10
11
x
y

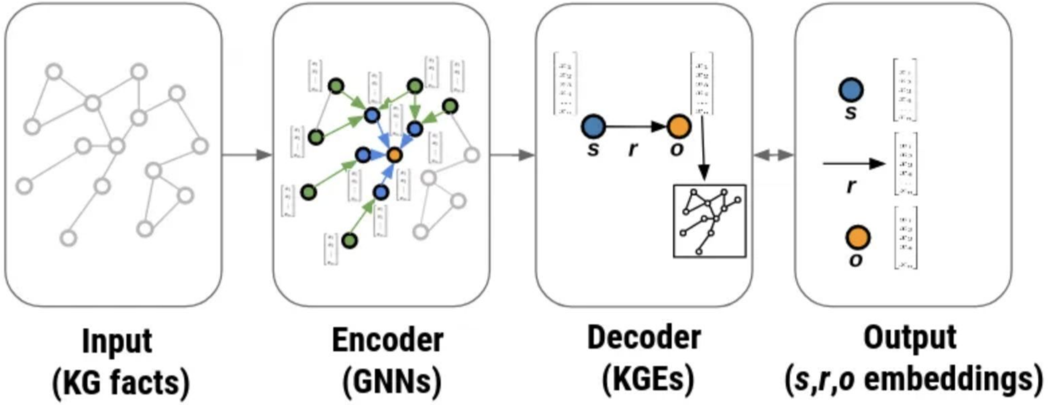
from https://gearons.org

Some GNN Tasks

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

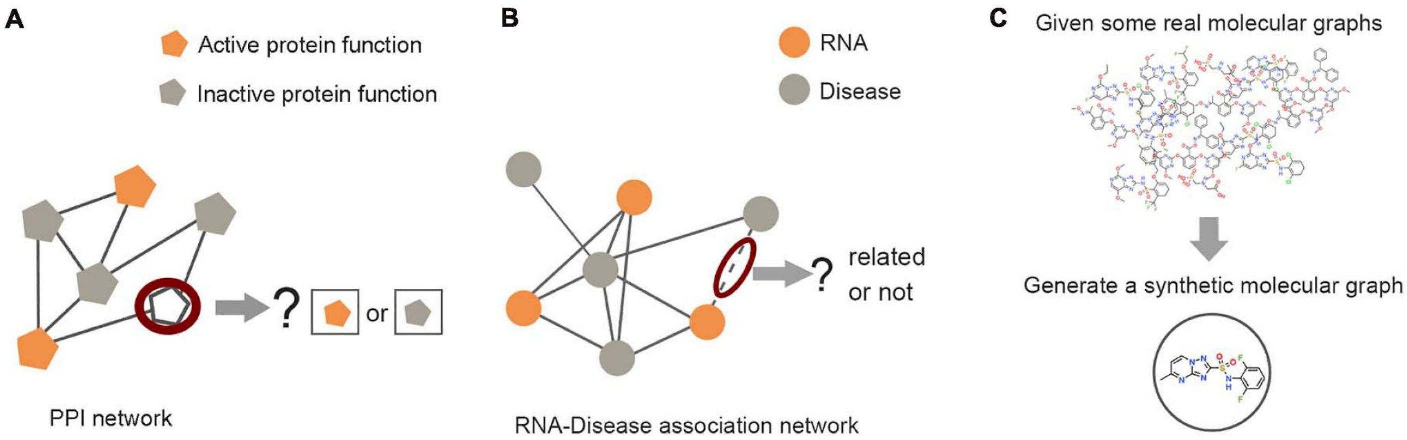
Zhang X-M, Liang L, Liu L and Tang M-J (2021) Graph Neural Networks and Their Current Applications in Bioinformatics. Front. Genet. 12:690049. doi: 10.3389/fgene.2021.690049

Example in practice

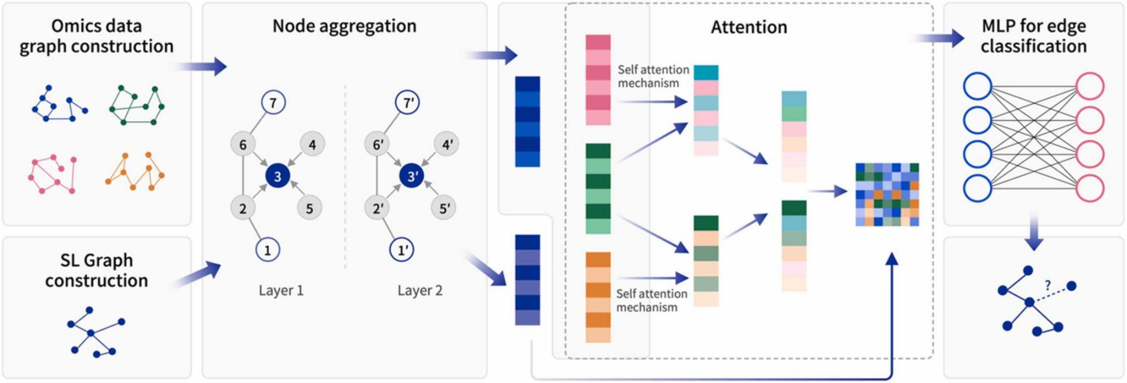


https://towardsdatascience.com/improving-knowledge-graph-embeddings-with-graph-neural-networks-8ffb14804199

Link prediction / node classification in life sciences

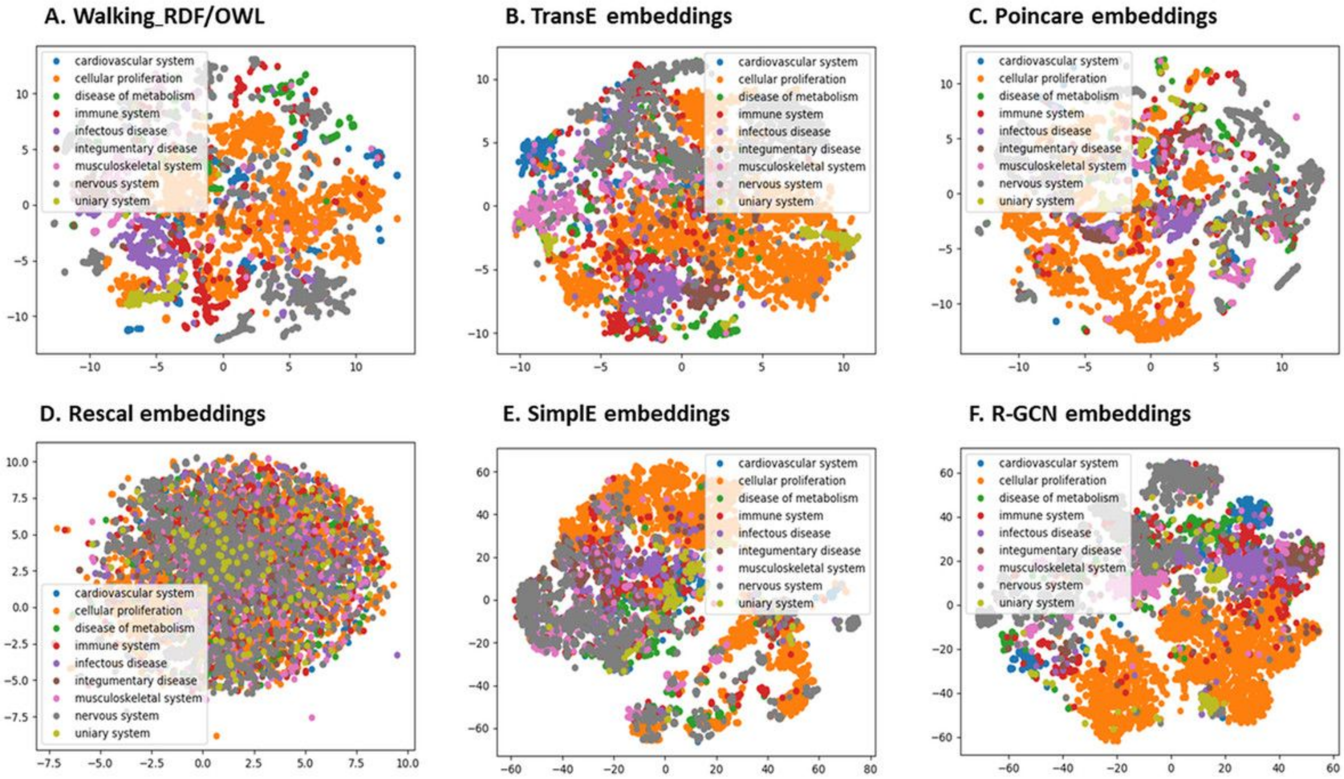

A
Active protein function
Inactive protein function
?
or
PPI network
B
RNA
Disease
?
related
or not
RNA-Disease association network
C
Given some real molecular graphs
Generate a synthetic molecular graph

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


A. Walking_RDF/OWL

B. TransE embeddings

C. Poincare embeddings

D. Rescal embeddings

E. SimplE embeddings

F. R-GCN embeddings

KG embeddings models

[Uniprot Embeddings](#)

[RDF2VEC](#)

[DistMult](#)

[MultiKE](#)

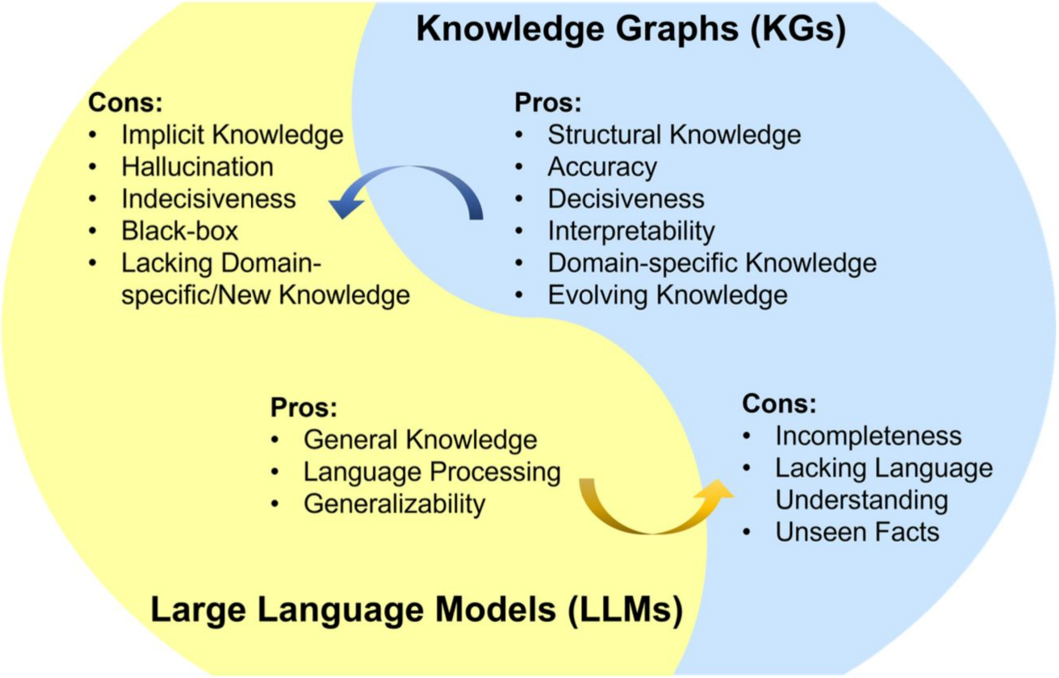
[ComplEx](#)

[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



Knowledge Graphs (KGs)

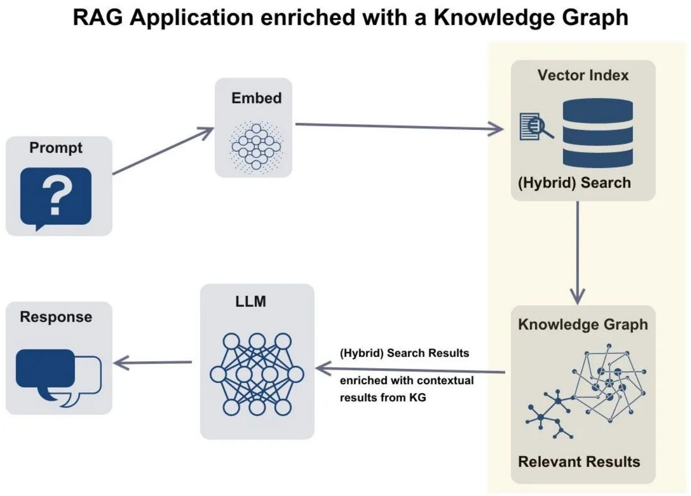
Cons:
- Implicit Knowledge
- Hallucination
- Indecisiveness
- Black-box
- Lacking Domain-specific/New Knowledge

Pros:
- Structural Knowledge
- Accuracy
- Decisiveness
- Interpretability
- Domain-specific Knowledge
- Evolving Knowledge

Pros:
- General Knowledge
- Language Processing
- Generalizability

Cons:
- Incompleteness
- Lacking Language Understanding
- Unseen Facts

Large Language Models (LLMs)



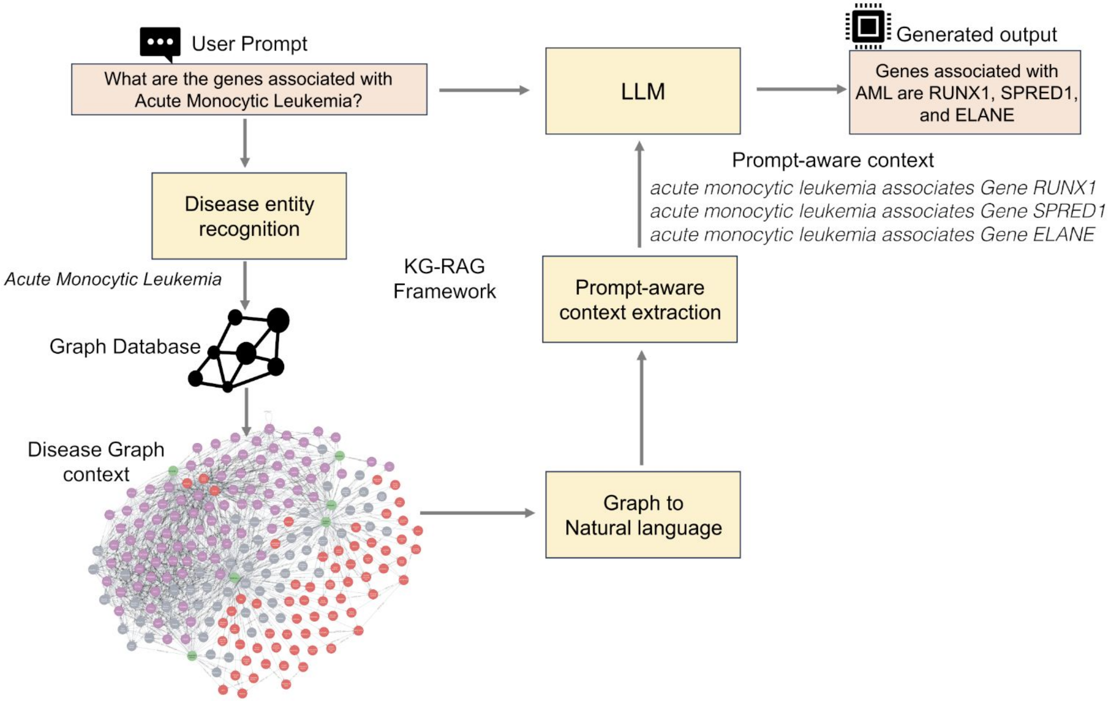
Retrieval Augmented Generation and Knowledge Graphs

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

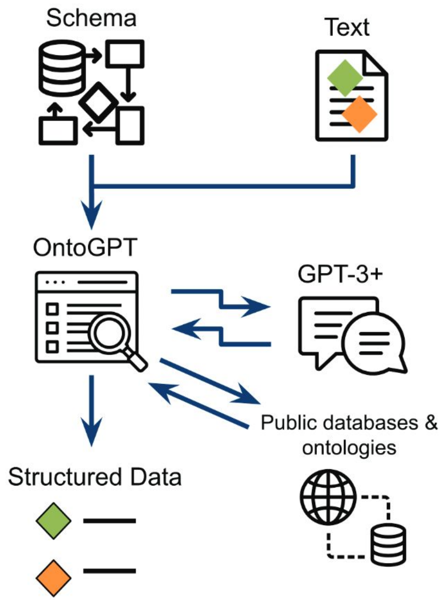
=> Boosting LLMs with External Knowledge: The Case for Knowledge Graphs

[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


User Prompt
What are the genes associated with Acute Monocytic Leukemia?
LLM
Generated output
Genes associated with AML are RUNX1, SPRED1, and ELANE
Prompt-aware context
acute monocytic leukemia associates Gene RUNX1
acute monocytic leukemia associates Gene SPRED1
acute monocytic leukemia associates Gene ELANE
Disease entity recognition
Acute Monocytic Leukemia
KG-RAG Framework
Prompt-aware context extraction
Graph Database
Disease Graph context
Graph to Natural language

KG-RAG. Soman et al. https://arxiv.org/pdf/2311.17330.pdf



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

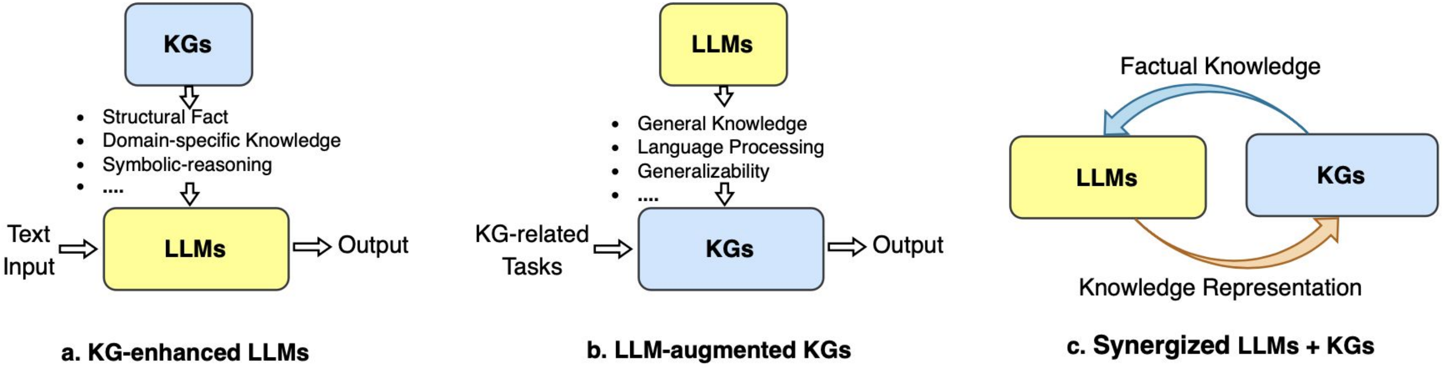
Toward combining KGs and LLMs

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



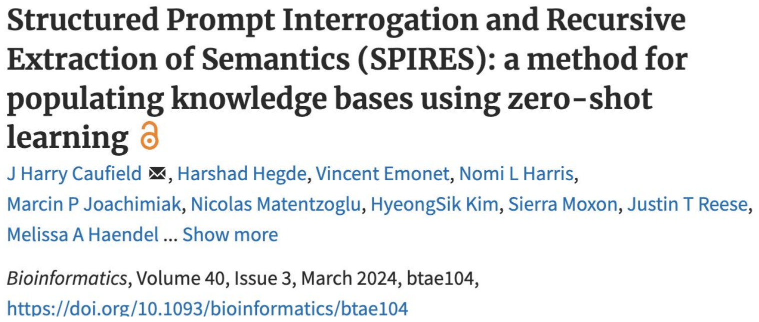
Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2023). Unifying Large Language Models and Knowledge Graphs: A Roadmap. *ArXiv, abs/2306.08302.*

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



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

Biomedical knowledge graph-enhanced prompt generation for large language models
Soman et al. https://arxiv.org/pdf/2311.17330.pdf

Biologically informed variational autoencoders allow predictive modeling of genetic and drug-induced perturbations

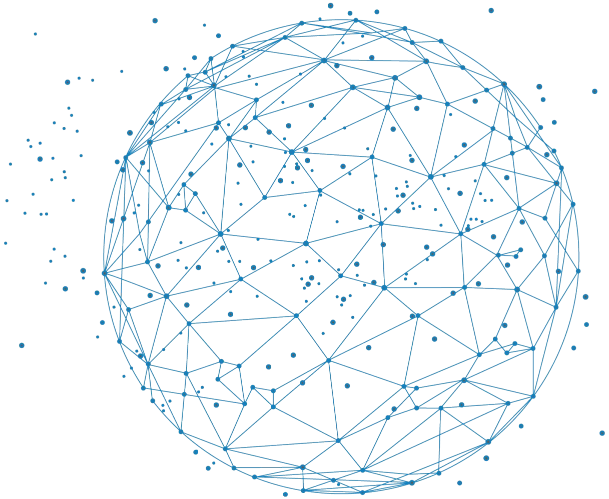
Daria Doncevic [1], Carl Herrmann [1]

Affiliations + expand

PMID: 37326971 PMCID: PMC10301695 DOI: 10.1093/bioinformatics/btad387

Dynamic Retrieval Augmented Generation of Ontologies using Artificial Intelligence (DRAGON-AI)
Sabrina Toro *et al*
https://arxiv.org/abs/2312.10904

Keep in mind



Key concepts

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

Advantages/disadvantages of the approach

Advantages

- Graphs for humans & machines
- Semantic heterogeneity
- Established web technologies
- Technological framework for F - * - I - R principles
- De-centralized: many query-able datasets published on the web (Linked Data Cloud)

Disadvantages

- Semantic heterogeneity (many Life Science ontologies …)
- Reliability of external SPARQL endpoints
- No graphical user interfaces… learn RDF syntaxes and SPARQL query language
- Scalable querying is still a hot research topic …

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