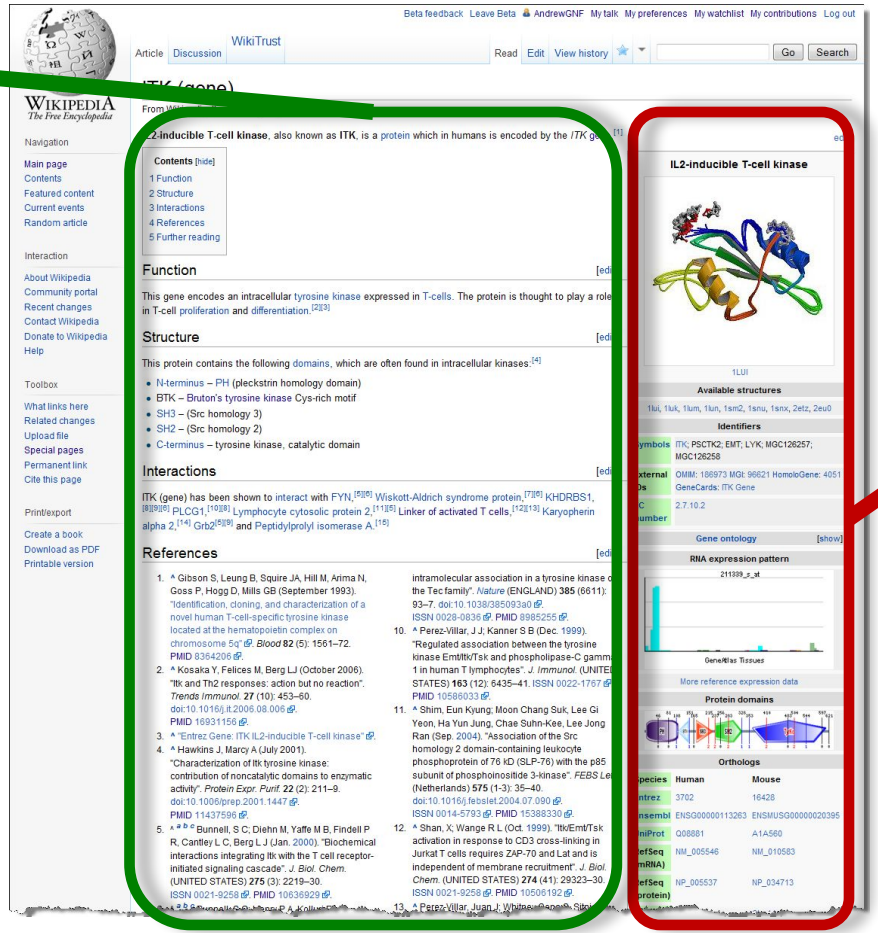
Introducing federated queries with Wikidata

Andra Waagmeester[1], Micelio, Antwerp, Belgium | Email: andra@micel.io, Twitter: @andrawaag
Micelio, Ekeren, Antwerp, Belgium

The Gene Wiki project, circa 2008

Summarized knowledge via crowdsourcing



Data imported from structured databases

Huss, PLoS Biol, 2008

Reelin

From Wikipedia, the free encyclopedia

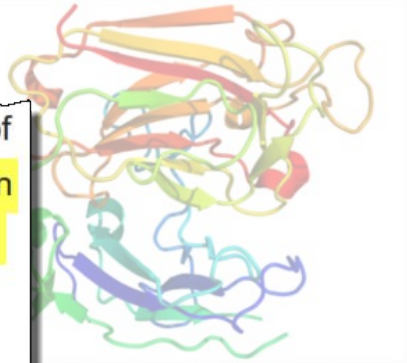
Reelin is a large secreted extracellular matrix glycoprotein that helps regulate processes of neuronal migration and positioning in the developing brain by controlling cell–cell interactions. Besides this important role in early development, reelin continues to work in the adult brain. It modulates synaptic plasticity by [2][3] It also stimulates dendrite[4] migration of neuroblasts generat zones. It is found not only in the tissues.

Reelin has been suggested to be expression of the protein has be bipolar disorder, but the cause of this observation remains uncertain as studies show that psychotropic medication itself affects reelin expression. Moreover, epigenetic hypotheses aimed at explaining the changed levels of reelin expression[6] are controversial.[7][8] Total lack of reelin causes a form of lissencephaly. Reelin may also play a role in Alzheimer's disease, temporal lobe epilepsy and autism.

Reelin's name comes from the abnormal reeling gait of *reeler* mice,[9] which were later found to have a deficiency of this brain protein and were homozygous for mutation of the RELN gene. The

Reelin has been suggested to be implicated in pathogenesis of several brain diseases. The expression of the protein has been found to be significantly lower in schizophrenia and psychotic bipolar disorder, but the cause of this observation remains uncertain as studies show that psychotropic medication itself

Reelin



hic structure of the third reelin repeat domain.[1]

Available structures	
PDB	Ortholog search: PDBe, RCSB
	List of PDB id codes [show]

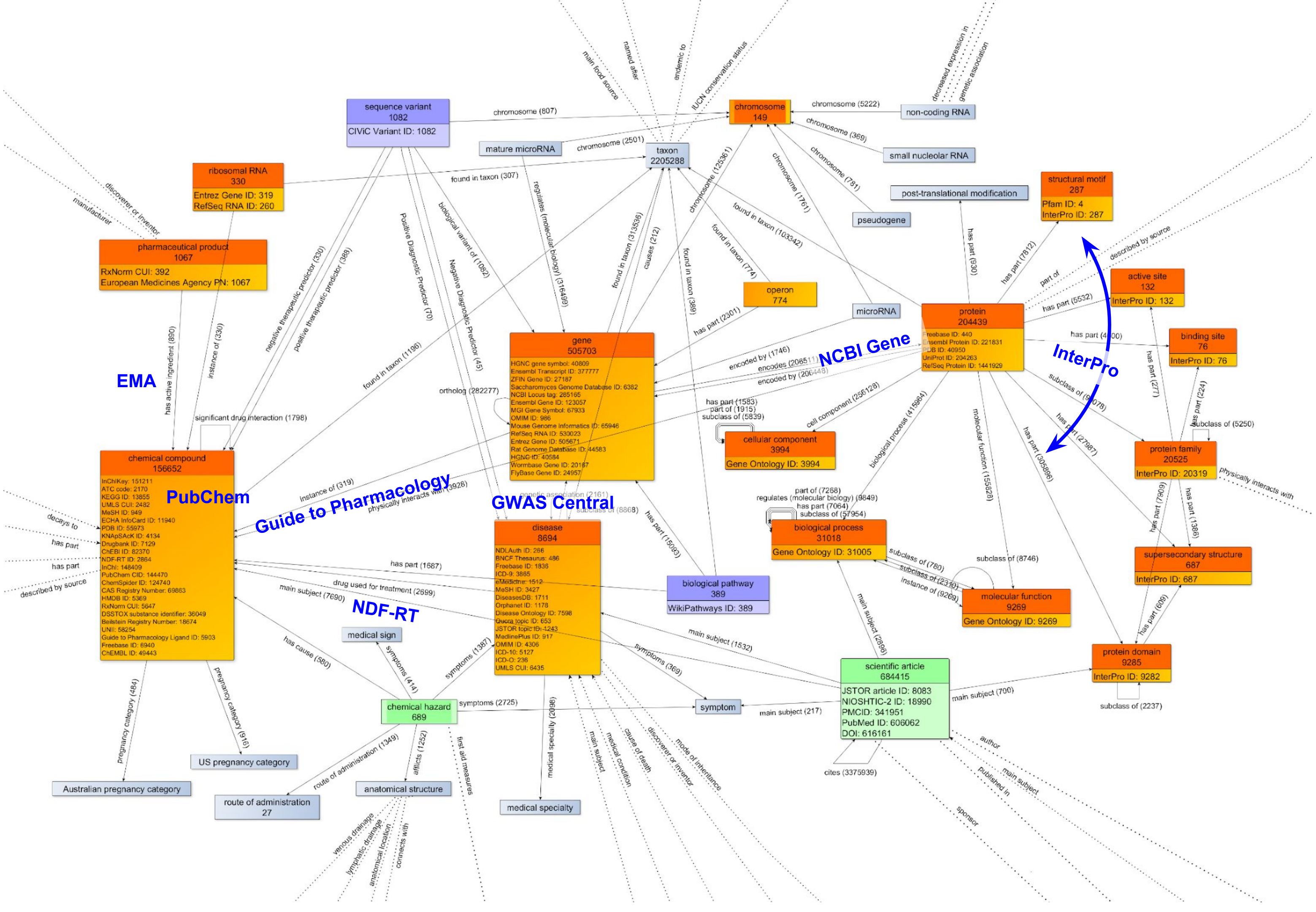
Identifiers	
Symbols	RELN ; LIS2; PRO1598; RL
External	OMIM: 600514 MGI: 103022

is to text

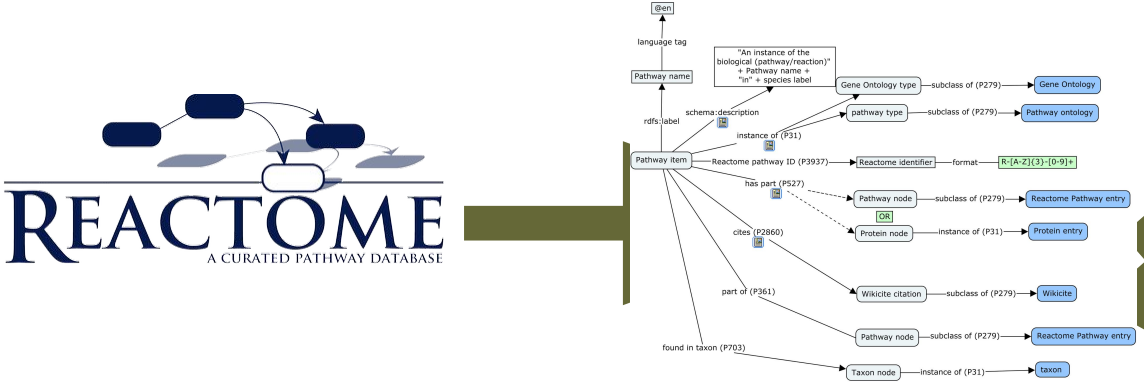
is to data

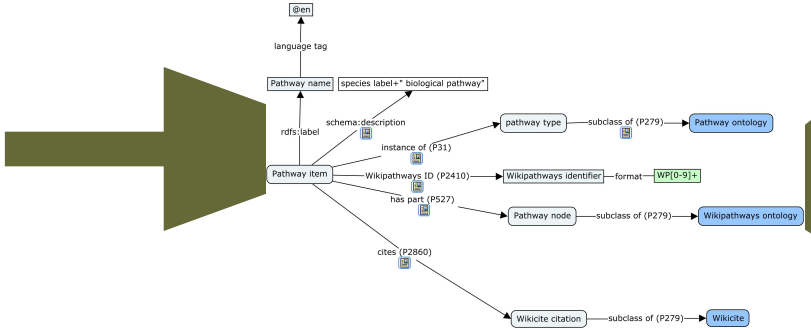
biomedical

> Provide a database of the world's knowledge that anyone can edit
>
> - Denny Vrandečić


EMA
PubChem
Guide to Pharmacology
GWAS Central
NDF-RT
NCBI Gene
InterPro
sequence variant
1082
CIViC Variant ID: 1082
chromosome
149
non-coding RNA
small nucleolar RNA
mature microRNA
taxon
2205288
ribosomal RNA
330
Entrez Gene ID: 319
RefSeq RNA ID: 260
pharmaceutical product
1067
RxNorm CUI: 392
European Medicines Agency PN: 1067
post-translational modification
structural motif
287
Pfam ID: 4
InterPro ID: 287
pseudogene
operon
774
microRNA
protein
204439
active site
132
InterPro ID: 132
binding site
76
InterPro ID: 76
gene
505703
cellular component
3994
Gene Ontology ID: 3994
protein family
20525
InterPro ID: 20319
chemical compound
156652
biological process
31018
Gene Ontology ID: 31005
supersecondary structure
687
InterPro ID: 687
disease
8694
biological pathway
389
WikiPathways ID: 389
molecular function
9269
Gene Ontology ID: 9269
medical sign
protein domain
9285
InterPro ID: 9282
scientific article
684415
JSTOR article ID: 8083
NIOSHTIC-2 ID: 18990
PMCID: 341951
PubMed ID: 606062
DOI: 616161
chemical hazard
689
symptom
US pregnancy category
Australian pregnancy category
route of administration
27
anatomical structure
medical specialty
chromosome (807)
chromosome (5222)
chromosome (369)
chromosome (2501)
found in taxon (307)
significant drug interaction (1798)
has active ingredient (890)
instance of (330)
ortholog (282277)
encoded by (1746)
encodes (206511)
encoded by (206448)
has part (1687)
drug used for treatment (2699)
main subject (7690)
main subject (1532)
main subject (217)
main subject (700)
symptoms (2725)
cites (3375939)
subclass of (2237)
subclass of (8746)
subclass of (5250)

Pathways in Wikidata





Reactome Properties in Wikidata

Properties [edit]

V · T · E	Wikidata properties related to **Reactome** (Q2134522)
Required properties	instance of (P31) · has part (P527) · part of (P361) · Reactome pathway ID (P3937) · found in taxon (P703) · exact match (P2888)
Optional properties	cites (P2860)
Reference properties	stated in (P248) · retrieved (P813) · Reactome pathway ID (P3937)
Primary sources	Reactome (Q2134522)
External Ontologies	Pathway Ontology (Q28864280) · Gene Ontology (Q135085)
Species covered	Homo sapiens (Q15978631)

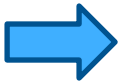
Wikipathways Properties in Wikidata

Properties [edit]

V · T · E	Wikidata properties related to **biological pathway** (Q4915012)
Required properties	instance of (P31) · has part (P527) · WikiPathways ID (P2410) · found in taxon (P703) · exact match (P2888)
Optional properties	connects with (P2789) · cites (P2860)
Reference properties	stated in (P248) · retrieved (P813) · WikiPathways ID (P2410)
Primary sources	WikiPathways (Q7999828)
External Ontologies	Pathway Ontology (Q28864280) · Cell line ontology (Q21039006) · Disease Ontology (Q5282129)
Species covered	Homo sapiens (Q15978631)

Simple data retrieval

"Retrieve genes with GWAS association with asthma"



39 genes

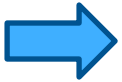
gene	geneLabel		gene	geneLabel		gene	geneLabel		gene	geneLabel
Q5013317	COL22A1		Q18027370	IGSF3		Q18053559	CDHR3		Q14903974	SMAD3
Q14912759	SLC22A5		Q18045382	HPSE2		Q18045669	ATG3		Q18033889	IL1RL1
Q14914243	PSAP		Q18048437	IL33		Q18035037	RAD50		Q17917202	ERBB4
Q14907990	SLC30A8		Q18051900	PYHIN1		Q18036984	FBXL7		Q18027836	IL6R
Q18025002	GAB1		Q17709208	ACO1		Q18033919	XPR1		Q18030185	NOTCH4
Q18035589	C6orf10		Q18027822	IL2RB		Q15326496	RORA		Q18030409	PDE4D
Q18054256	GSDMA		Q18030364	PBX2		Q18042132	GSDMB		Q18045645	IKZF4
Q18058487	C5orf56		Q18037773	ABI3BP		Q18029145	MKLN1		Q18039979	KLHL5
Q18030785	PRKG1		Q18039623	CTNNA3		Q18036729	RAP1GAP2		Q18026947	HLA-DQA1
Q18033424	IL18R1		Q18046350	ZNF665		Q14878303	IL13			

```
SELECT DISTINCT ?gene ?geneLabel where {
  ?gene wdt:P2293 wd:Q35869 .  # gene has genetic association to "asthma"
  ?gene wdt:P31 wd:Q7187 .     # gene is subclass of "gene"
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
```

http://bit.ly/bosc2017_wikidata

Data integration

"Retrieve genes with GWAS association with asthma and gene product is localized to membrane"



22 genes

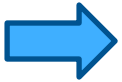
gene	geneLabel	gene	geneLabel	gene	geneLabel	gene	geneLabel
Q14912759	SLC22A5	Q18027370	IGSF3	Q18035037	RAD50	Q18027836	IL6R
Q14914243	PSAP	Q18033424	IL18R1	Q18033919	XPR1	Q18030409	PDE4D
Q14907990	SLC30A8	Q18045382	HPSE2	Q18042132	GSDMB	Q18030185	NOTCH4
Q18035589	C6orf10	Q18027822	IL2RB	Q18036729	RAP1GAP2	Q18026947	HLA-DQA1

```
SELECT DISTINCT ?gene ?geneLabel where {
  ?gene wdt:P2293 wd:Q35869 .  # gene has genetic association to "asthma"

  ?gene wdt:P31 wd:Q7187 .     # gene is subclass of "gene"

  ?gene wdt:P688 ?protein .                # gene encodes a protein
  ?protein wdt:P681 ?cc .                  # protein has a cellular component
  ?cc wdt:P279*|wdt:P361* wd:Q14349455 .   # cell component is 'part of' or 'subclass of' membrane

  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
```

http://bit.ly/bosc2017_wikidata

Computing on provenance

"Retrieve genes with GWAS association with asthma and gene product is localized to membrane (non-IEA)"



15 genes

gene	geneLabel		gene	geneLabel		gene	geneLabel
Q14912759	SLC22A5		Q18045382	HPSE2		Q17917202	ERBB4
Q14914243	PSAP		Q18027822	IL2RB		Q18027836	IL6R
Q14907990	SLC30A8		Q14903974	SMAD3		Q18030409	PDE4D
Q18027370	IGSF3		Q18035037	RAD50		Q18030185	NOTCH4
Q18033424	IL18R1		Q18036729	RAP1GAP2		Q18026947	HLA-DQA1

```
6   ?gene wdt:P31 wd:Q7187 ;      # gene is subclass of "gene"
7         wdt:P688 ?protein ;     # gene encodes a protein
8         rdfs:label ?geneLabel .
9   FILTER (lang(?geneLabel) = "en")
10  ?protein p:P681 ?s .                          # protein's cell component statement
11    ?s ps:P681 ?cp .                            # get statement value
12    FILTER NOT EXISTS {?s pq:P459 wd:Q23190881 .}  # determination method is not IEA
13    ?cp wdt:P279*|wdt:P361* wd:Q14349455 .       # statement value is 'part of' or 'subclass of' membrane
14
```

http://bit.ly/bosc2017_wikidata

Leveraging the Disease Ontology structure

"Retrieve genes with GWAS association with any respiratory disease and gene product is localized to membrane (non-IEA)"

31 genes / 8 diseases

diseaseGALabel	gene_counts	geneList
asthma	15	SMAD3, RAP1GAP2, IL18R1, HPSE2, SLC30A8, SLC22A5, PSAP, ERBB4, HLA-DQA1, IGSF3, IL2RB, IL6R, NOTCH4, PDE4D, RAD50
chronic obstructive pulmonary disease	5	HLA-C, SFTPD, ANXA5, ANXA11, ATP2C2
lung cancer	3	TGM5, VTI1A, PHACTR2
interstitial lung disease	2	DSP, ATP11A
non-small-cell lung carcinoma	2	NALCN, DLST
nasopharynx carcinoma	2	ITGA9, TNFRSF19
adenocarcinoma of the lung	1	BTNL2
pulmonary emphysema	1	BICD1

```
SELECT ?diseaseGALabel (count (DISTINCT ?ger
(group_concat(DISTINCT ?geneLabel; separator
  ?gene wdt:P2293 ?diseaseGA .          # gene
  ?diseaseGA wdt:P279* wd:Q3286546 .    # to a

  ?gene wdt:P31 wd:Q7187 ; wdt:P688 ?protein ;    # gene is subclass of "gene" and encodes protein
        rdfs:label ?geneLabel .
  FILTER (lang(?geneLabel) = "en")
  ?protein p:P681 ?s .                            # protein's cell component statement
    ?s ps:P681 ?cp .                              # get statement value
```

http://bit.ly/bosc2017_wikidata

Opportunistic integration

"Retrieve genes with GWAS association with any respiratory disease and gene product is localized to membrane (non-IEA) and show causative chemical hazards"

4 diseases / 6 chemical hazards

diseaseGALabel	exposureLabel
lung cancer	arsenic pentoxide exposure
lung cancer	HN1 exposure
lung cancer	mechlorethamine exposure
lung cancer	HN3 exposure
asthma	Phenacyl chloride exposure
pulmonary emphysema	phosgene exposure

```
12
13    ?exposure wdt:P1542 ?diseaseGA .  # something causes disease
14    ?exposure wdt:P279 wd:Q21167512 . # and that something is a chemical hazard
15
16    SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
17  }
```

http://bit.ly/bosc2017_wikidata

... and show associated pathways

"Retrieve genes with GWAS association with any respiratory disease and gene product is localized to membrane (non-IEA), show causative chemical hazards and show pathways where they have a role."

16 genes / 59 pathways

gene	pathway
SMAD3	Androgen receptor signaling pathway
SMAD3	TGF-beta Receptor Signaling
SMAD3	mechlorethamine exposure
HLA-C	Allograft Rejection
SFTPD	Regulation of toll-like receptor signaling pathway
....	.....

```
12
13    ?pathway wdt:P31 wd:Q4915012 ;                # instance of a biological pathway
14             wdt:P527 ?gene .
15
16    SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
17  }
```

http://bit.ly/bosc2017_wikidata

... and show remote pathway annotations

"Retrieve genes with GWAS association with any respiratory disease and gene product is localized to membrane (non-IEA), show causative chemical hazards, show pathways where they have a role and show pathway annotations in WikiPathways"

16 genes / 59 pathways

Remote pathway annotations	
Pathway ontology	55
Cell line ontology	12
Disease ontology	11
….	…..

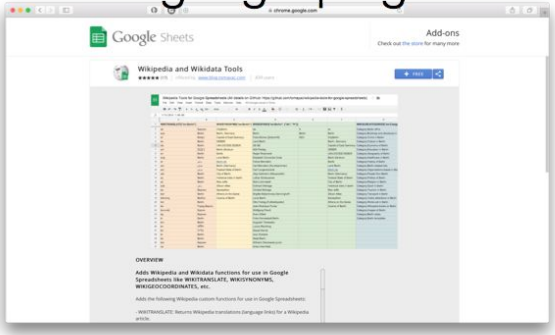
```
        wdt:P2888 ?source_pathway .
  SERVICE <http://sparql.wikipathways.org/> {
    ?wp_pathway dc:identifier ?source_pathway ;
                wp:ontologyTag ?pwTag .
}
```

wikidata

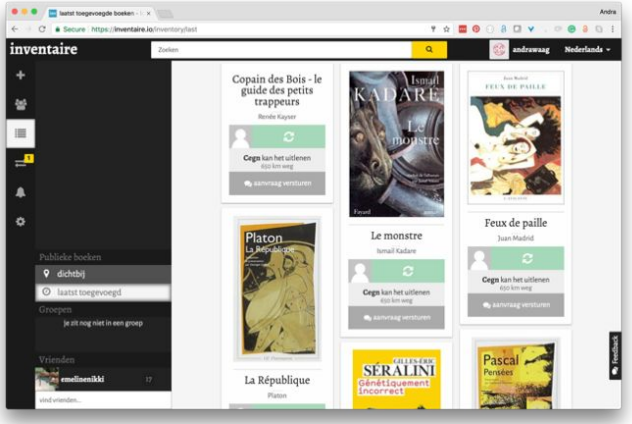
Tools using Wikidata

http://www.wikigenomes.org

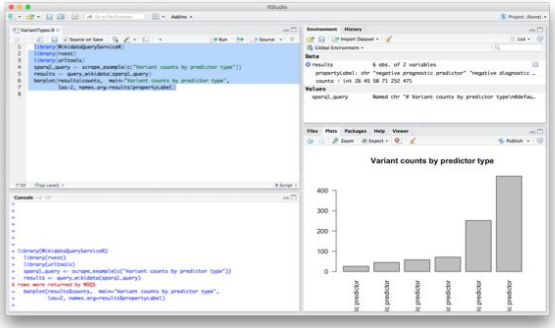
Wikipedia and Wikidata google plugin



http://inventaire.io



Cytoscape



R plugins

Availability

www.wikidata.org/wiki/User:Pathwaybot

- www.wikidata.org/wiki/User:Pathwaybot/data_model - Semantic models
- www.wikidata.org/wiki/User:Pathwaybot/query_examples – Pathway examples
- www.wikidata.org/wiki/User:ProteinBoxBot/SPARQL_Examples– Genewiki example

github.com/SuLab/GeneWikiCentral

- github.com/SuLab/**wikidataintegrator** – python module for Wikidata
- github.com/SuLab/**scheduled-bots** – bot automation framework
- github.com/SuLab/**Genewiki-ShEx** – data models

Structure of the federated query

```
SELECT * WHERE {
    ?localitem ?localproperty ?shareditem .
    SERVICE <http://example.remote.sparql> {
        ?shareditem ?remoteproperty ?remoteitem .
    }
}
```

Local pattern

Remote pattern

A federated query contains query patterns for both the local endpoint (green box) and a remote endpoint (blue box). The address of the remote SPARQL endpoint is expressed with the SERVICE keyword

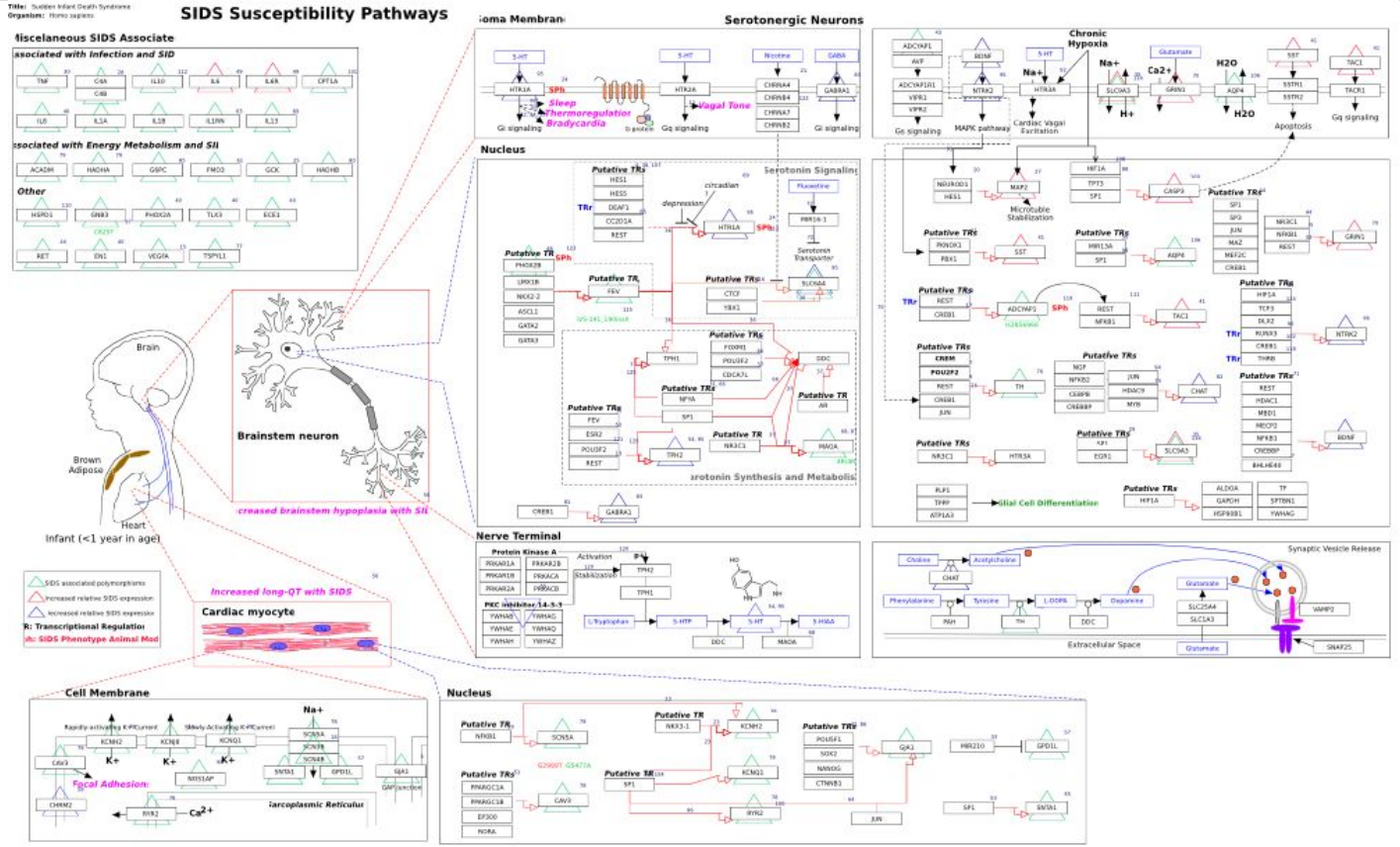
Why federate

- Incorporate more fine-grained data not captured in Wikidata
- Link to other (non-public licenses)
- Link to volatile data

Wikidata Query Service support federated queries

- From Wikidata to an external SPARQL endpoint (Wikipathways)
- From a remote SPARQL endpoint to Wikidata
- From a local SPARQL endpoint to Wikidata
- From a local Wikibase (with WDQS) to Wikidata

From Wikidata to an external SPARQL endpoint (Wikipathways)



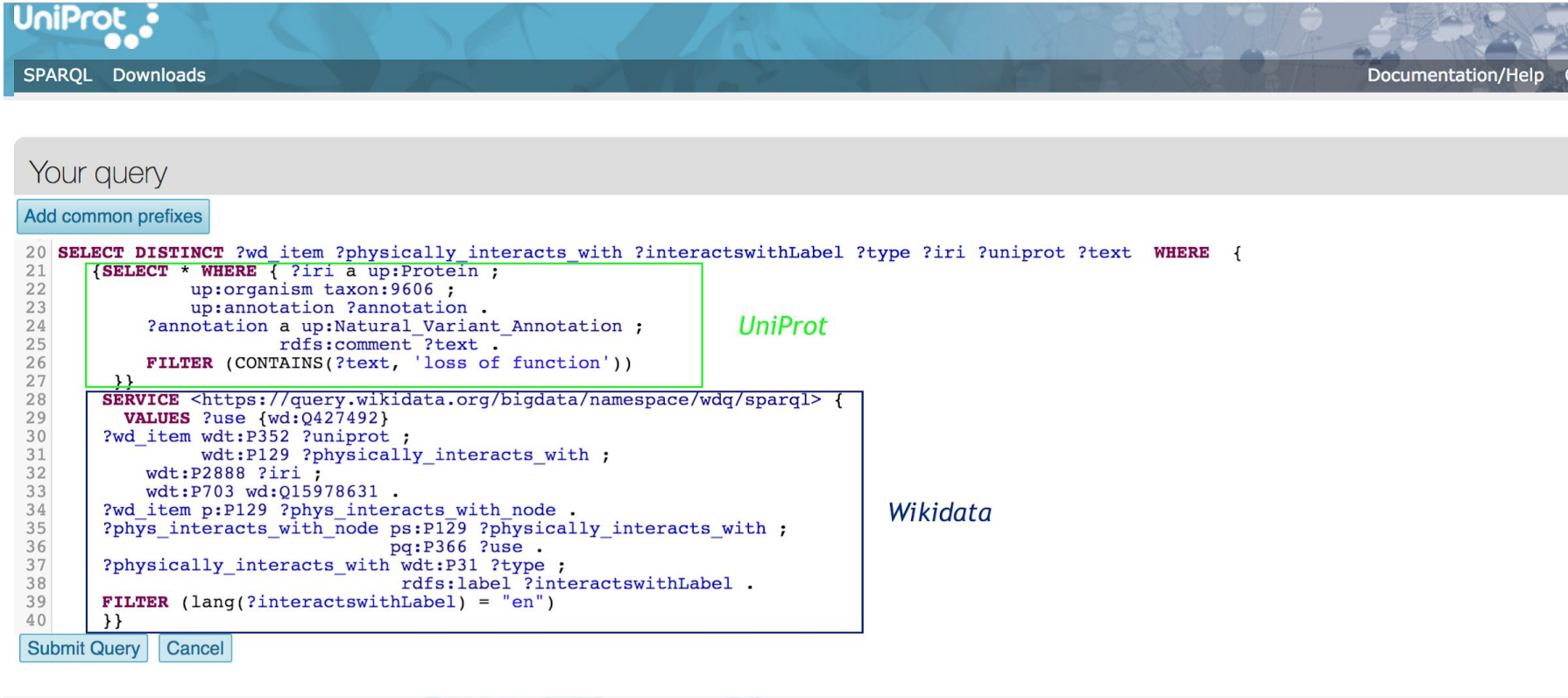
```
PREFIX wp:       <http://vocabularies.wikipathways.org/wp#>
PREFIX dcterms:   <http://purl.org/dc/terms/>
PREFIX dc:       <http://purl.org/dc/elements/1.1/>
SELECT DISTINCT ?metabolite1Label ?metabolite2Label ?mass1 ?mass2 WITH {
  SELECT ?metabolite1 ?metabolite2 WHERE {
    ?pathwayItem wdt:P2410 "WP706";
                 wdt:P2888 ?pwIri.
    SERVICE <http://sparql.wikipathways.org/> {
      ?pathway dc:identifier ?pwIri.
      ?interaction rdf:type wp:Interaction;
                   wp:participants ?wpmb1, ?wpmb2;
                   dcterms:isPartOf ?pathway.
      FILTER (?wpmb1 != ?wpmb2)
      ?wpmb1 wp:bdbWikidata ?metabolite1.
      ?wpmb2 wp:bdbWikidata ?metabolite2.
    }
  }
} AS %metabolites WHERE {
  INCLUDE %metabolites.
  ?metabolite1 wdt:P2067 ?mass1.
  ?metabolite2 wdt:P2067 ?mass2.
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
```

Wikidata

Wikipathways

Wikidata

From a remote SPARQL endpoint to Wikidata



We use ShEx to describe contents of a remote endpoint (http://shex.io)

- Draft Shape Expression for the Europeana Sparql endpoint:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX edm: <http://www.europeana.eu/schemas/edm/>
PREFIX rdvocab: <http://rdvocab.info/ElementsGr2/>
PREFIX dce: <http://purl.org/dc/elements/1.1/>
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX dbp: <http://dbpedia.org/resource/>
PREFIX oaterms: <http://www.openarchives.org/ore/terms/>
PREFIX bibgalicia: <http://www.galiciana.bibliotecadegalicia.xunta.es/aut/>
PREFIX gtaa: <http://data.beeldengeluid.nl/gtaa/>
PREFIX owl: <http://www.w3.org/2002/07/owl#>

<aggregation> {
    edm:dataProvider LITERAL* ;
    edm:rights <licenses>? ;
    edm:isShownBy IRI* ;
}

<concept> {
    rdf:type skos:Concept ;
    skos:prefLabel LITERAL* ;
    skos:broader  [dbp:~]*  ;
    skos:exactMatch IRI* ;
    skos:closeMatch IRI* ;
    skos:related [dbp:~ bibgalicia:~]* ;
    skos:notation LITERAL? ;
```