Wikidata as a infrastructural backbone of scientific knowledge

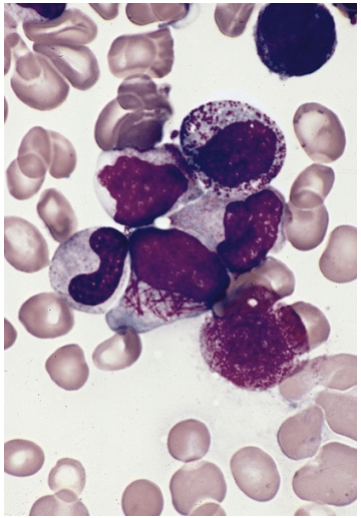
Andra Waagmeester,
Micelio, Antwerp, Belgium | Email: andra@micelio.be, Twitter: @andrawaag

DOI: https://doi.org/10.6084/m9.figshare.7958693

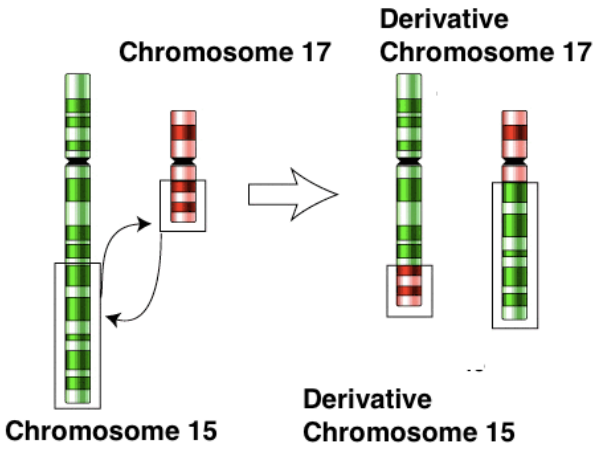
Folk tale of the stone soup

Acute promyelocytic leukemia

Cancer of the white blood cells



Chromosomal translocation on the RARα gene



VOLUME 32 · NUMBER 30 · OCTOBER 20 2014

JOURNAL OF CLINICAL ONCOLOGY

ORIGINAL REPORT

Randomized Phase III Trial of Retinoic Acid and Arsenic Trioxide Versus Retinoic Acid and Chemotherapy in Patients With Acute Promyelocytic Leukemia: Health-Related Quality-of-Life Outcomes

Fabio Efficace, Franco Mandelli, Giuseppe Avvisati, Francesco Cottone, Felicetto Ferrara, Eros Di Bona, Giorgina Specchia, Massimo Breccia, Alessandro Levis, Simona Sica, Olimpia Finizio, Maria Grazia Kropp, Giuseppe Fioritoni, Elisa Cerqui, Marco Vignetti, Sergio Amadori, Richard F. Schlenk, Uwe Platzbecker, and Francesco Lo-Coco

Fabio Efficace, Franco Mandelli, Francesco Cottone, and Marco Vignetti, Gruppo Italiano Malattie Ematologiche dell'Adulto; Giuseppe Avvisati, Università Campus Biomedico; Massimo Breccia, Università "La Sapienza,"; Simona Sica, Università Cattolica Sacro Cuore; Sergio Amadori and Francesco Lo-Coco, Università Tor Vergata; Francesco Lo-Coco, Fondazione Santa Lucia, Roma; Felicetto Ferrara, Ospedale Cardarelli; Olimpia Finizio, Ospedale Cardarelli, Napoli; Eros Di Bona, Ospedale San Bortolo, Vicenza; Giorgina Specchia, Università di Bari, Bari; Alessandro Levis, Ospedale SS Antonio e Biagio, Alessandria; Maria Grazia Kropp, Azienda Ospedaliera Pugliese Ciaccio, Catanzaro; Giuseppe Fioritoni, Ospedale Civile, Pescara; Elisa Cerqui, Spedali Civili, Brescia, Italy; Richard F. Schlenk, University of Ulm, Ulm; and Uwe Platzbecker, Universitatsklinikum Carl Gustav Carus, Dresden, Germany.

Published online ahead of print at

A B S T R A C T

Purpose
A randomized clinical trial compared efficacy and toxicity of standard all-*trans*-retinoic acid (ATRA) plus chemotherapy versus ATRA plus arsenic trioxide in patients with newly diagnosed, low- or intermediate-risk acute promyelocytic leukemia (APL). Here, we report health-related quality-of-life (HRQOL) results.

Patients and Methods
HRQOL was a secondary end point of this trial. The European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire–Core 30 was used to assess HRQOL at end of induction and after consolidation therapy. All analyses were based on 156 patients who received at least one dose of treatment, with groups defined according to randomly assigned treatment. Primary analysis was performed, estimating mean HRQOL score over time and differences between treatment arms using a linear mixed model.

Effects of arsenic trioxide known for decades in China

Original papers were published in the Chinese language and in journals that are obscure even to most Chinese readers

SCIENCE CHINA
Life Sciences

• REVIEW •

June 2013 Vol.56 No.6: 495–502
doi: 10.1007/s11427-013-4487-z

A drug from poison: how the therapeutic effect of arsenic trioxide on acute promyelocytic leukemia was discovered

RAO Yi[1*], LI RunHong[2] & ZHANG DaQing[2]

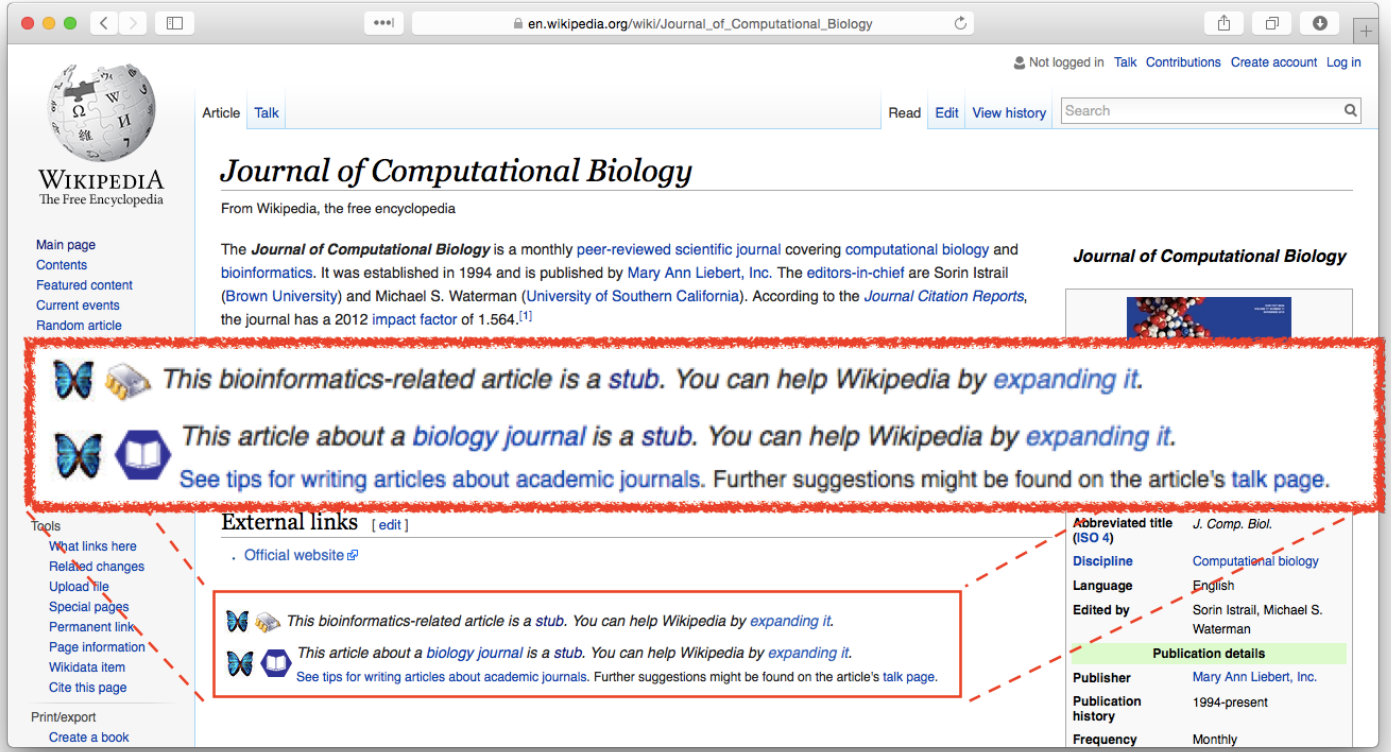
[1]*Peking-Tsinghua Center for Life Sciences at Peking University School of Life Sciences, Beijing 100871, China;*
[2]*Peking University Health Sciences Center, Beijing 100871, China*

Received March 27, 2013; accepted April 5, 2013; published online May 3, 2013

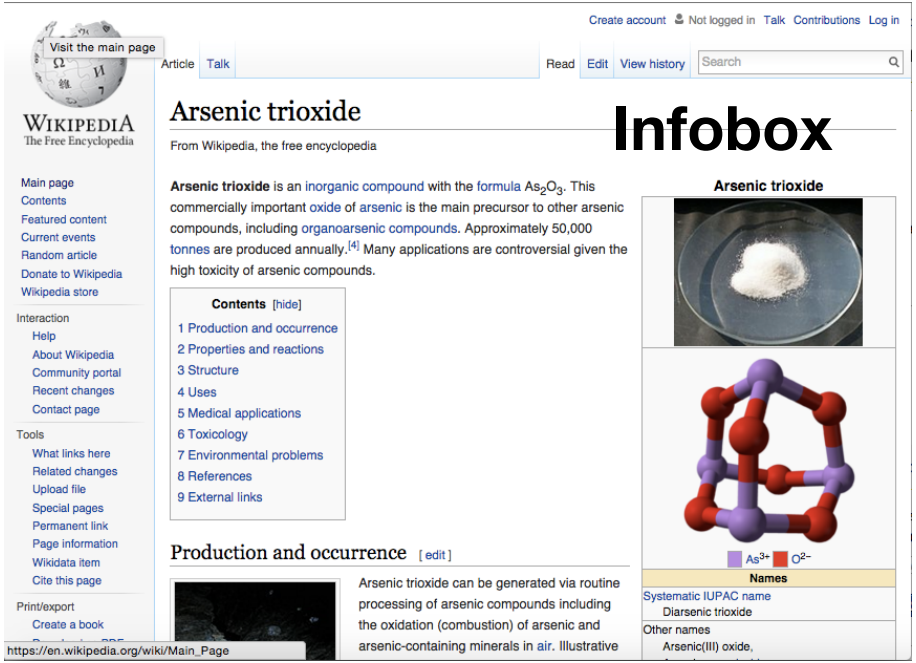
Findings support the use of retinoic acid plus arsenic trioxide as prefered first-line treatment

Source: http://jco.ascopubs.org/content/early/2014/09/17/JCO.2014.55.3453

Een Wikipedia artikel wordt net zo bereid als de stenen soep



Multilingual Wikipedia & Infoboxes

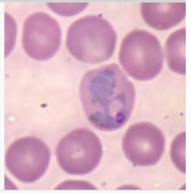


Language independent info boxes

Dutch

nl.m.wikipedia.org

Malaria



Rode bloedcel geïnfecteerd met *P. vivax*

Coderingen

ICD-10	B50
ICD-9	084
OMIM	248310

Greek

el.m.wikipedia.org

Ελονοσία

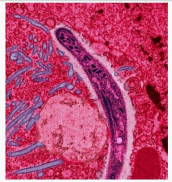
Ταξινόμηση και εξωτερικές πηγές

Ταξινόμηση ICD-10	B50
Ταξινόμηση ICD-9	084
OMIM	248310
DiseasesDB	7728
MedlinePlus	000621
eMedicine	med/1385 emerg/305 ped/1357

English

en.m.wikipedia.org

Malaria



A *Plasmodium* from the saliva of a female mosquito moving across a mosquito cell

Classification and external resources

Specialty	Infectious disease
ICD-10	B50-B54
ICD-9-CM	084

ICD-10 B50

ICD-10 B50-B54

Ταξινόμηση ICD-10 B50

Dutch

nl.m.wikipedia.org

	Papiaments
Hoofdstad	Oranjestad
Regeringsvorm	Constitutionele monarchie
Staatshoofd	Koning Willem-Alexander Fredis Refunjol (gouverneur)
Regeringsleider	Mike Eman (Arubaanse Volkspartij)
Religie	Katholiek 82%, protestant 8%

103,400 [2] (197th)

Greek

el.m.wikipedia.org

Πολίτευμα	Συνταγματική Μοναρχία
Μονάρχης Κυβερνήτης Πρωθυπουργός	Γουλιέλμος-Αλέξανδρος Φρέντις Ρεφουνιόλ Μάικ Έμαν
Πλήρης αυτονομία από το Βασίλειο των Κάτω Χωρών	1η Ιανουαρίου 1986
Σύνταγμα	
Έκταση	
• Σύνολο	180 km² (213η)
Ακτογραμμή	68,5 km
Πληθυσμός	
• Εκτίμηση 2014	107.394[1] (196η)
• Απογραφή 2000	103.065
• Πυκνότητα	556,4 κατ./km² (21
Α.Ε.Π. (PPP)	
• Ολικό (2005)	2,258 δισ. $[2]

• Εκτίμηση 2014 107.394[1] (196η)

English

pap.m.wikipedia.org

Forma di gobernacion	Democracia p Monarkia con
- Rei	Willem-Alexan
- Gobernador	Fredis Refunjo
- Prome Minister	Mike Eman
Pais den Reino di Hulanda	
Status aparte	1 januari di 19
Area	
- Total	193 km² (n/a)

101.484 (2010)[2]

110.663 (2014)[3]

(614,8/km² (2014))

The Gene Wiki project, circa 2008

Summarized knowledge via crowdsourcing

Data imported from structured databases

Huss, PLoS Biol, 2008

Reelin

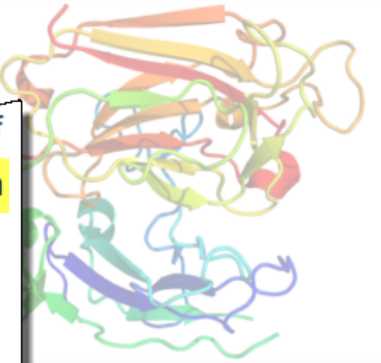
From Wikipedia, the free encyclopedia

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 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 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



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

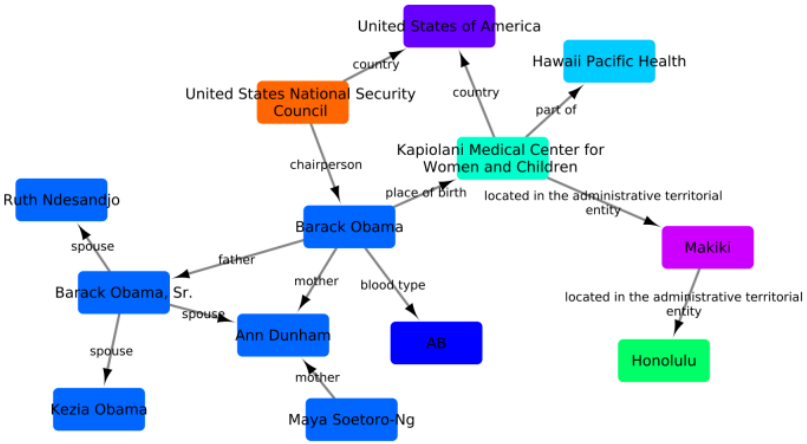
Wikidata is to data as Wikipedia is to text

Wikidata is a collaboratively edited knowledge base operated by the Wikimedia Foundation

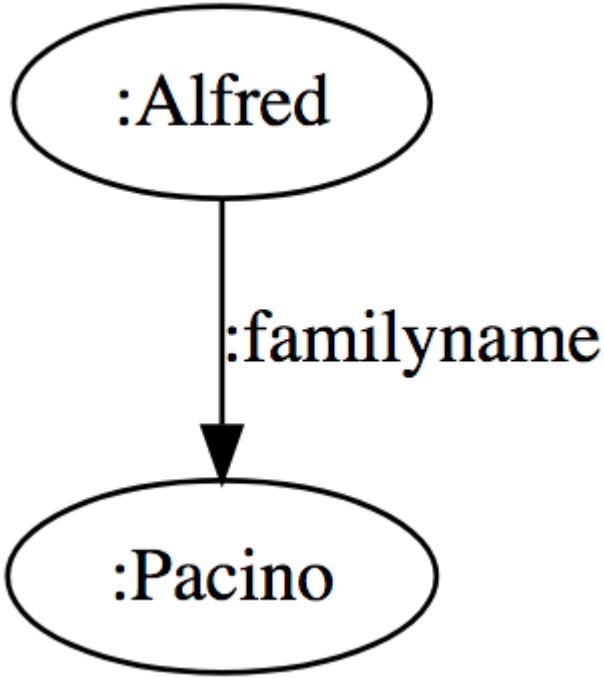
- Completely free, even for commercial usage (CC0)
- Anybody can contribute
- Covers all domains of knowledge
- Extensive item history, talk pages, projects, users
- Integration with the semantic web
- High performance query engine (SPARQL)

- Stable! Long term support not dictated by funding cy
- Actively developed
- Already has large number of active users, editors, contributors!

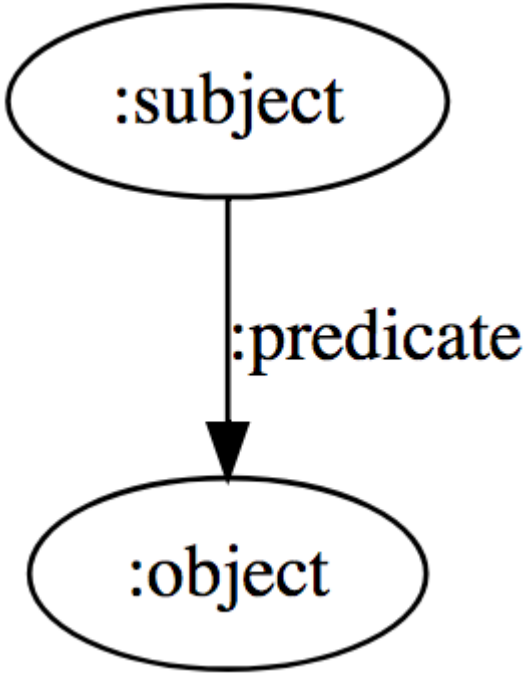
A giant graph of knowledge!



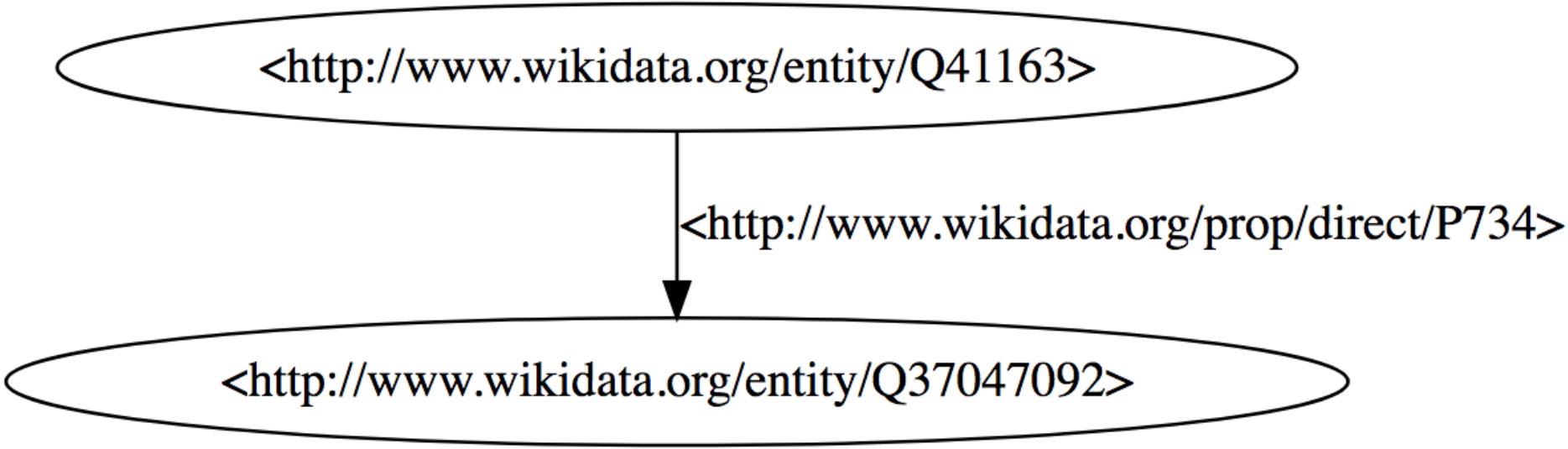
Linked data



Linked data



Linked Data


<http://www.wikidata.org/entity/Q41163>
<http://www.wikidata.org/prop/direct/P734>
<http://www.wikidata.org/entity/Q37047092>

Al Pacino (Q41163)

American film and stage actor and director edit

Alfredo James Steven Drew Duncan Pacino

In more languages

Language	Label	Description	Also known as
English	Al Pacino	American film and stage actor and director	Alfredo James Steven Drew Du…
German	Al Pacino	US-amerikanischer Schauspieler, Filmregisseur und Produzent italienischer Abstammung	Alfredo James Pacino Pacino
French	Al Pacino	acteur américain	Alfredo James Pacino
Dutch	Al Pacino	Amerikaans acteur	
Sranan Tongo	Al Pacino	No description defined	Alfredo Pacino

All entered languages

Statements

instance of — human edit

2 references

+ add value

image edit

Al Pacino.jpg
349 × 500; 48 KB

point in time 2004

Wikipedia (97 entries) edit

af	Al Pacino
an	Al Pacino
ar	آل باتشينو
arz	ال باتشينو
ast	Al Pacino
azb	آل پاچینو
az	Al Paçino
be_x_old	Аль Пачына
be	Аль Пачына
bg	Ал Пачино
bn	আল পাচিনো
bs	Al Pacino
ca	Al Pacino
ckb	ئاڵ پاچینۆ
co	Al Pacino
csb	Al Pacino
cs	Al Pacino
cy	Al Pacino
da	Al Pacino
de	Al Pacino
el	Αλ Πατσίνο
eml	Al Pacino
en	Al Pacino
eo	Al Pacino
es	Al Pacino
et	Al Pacino
eu	Al Pacino
fa	آل پاچینو
fi	Al Pacino
fr	Al Pacino
ga	Al Pacino
gd	Al Pacino

family name (P734)

surname or last name of a person

last name | has surname | surname

edit

In more languages

Language	Label	Description	Also known as
English	family name	surname or last name of a person	last name has surname surname
German	Familienname	Ergänzung zum Vornamen	Nachname Name Zuname
French	nom de famille	nom de famille d'une personne	patronyme
Dutch	achternaam	laatste naam van een persoon	familienaam
Sranan Tongo	No label defined	No description defined	

All entered languages

Data type

Item

Statements

instance of

Wikidata property for items about people edit

0 references

+ add reference

Wikidata property to indicate a name edit

0 references

+ add reference

Wikidata property encoding a vCard value edit

0 references

Pacino (Q37047092)

family name edit

In more languages

Language	Label	Description	Also known as
English	Pacino	family name	
German	Pacino	Familienname	
French	Pacino	nom de famille	
Dutch	Pacino	achternaam	
Sranan Tongo	Pacino	No description defined	

All entered languages

Statements

instance of	family name 1 reference	edit + add value
native label	Pacino (multiple languages) 0 references	edit + add reference + add value
writing system	Latin script 0 references	edit + add reference + add value

Wikipedia (0 entries) edit

Wikibooks (0 entries) edit

Wikinews (0 entries) edit

Wikiquote (0 entries) edit

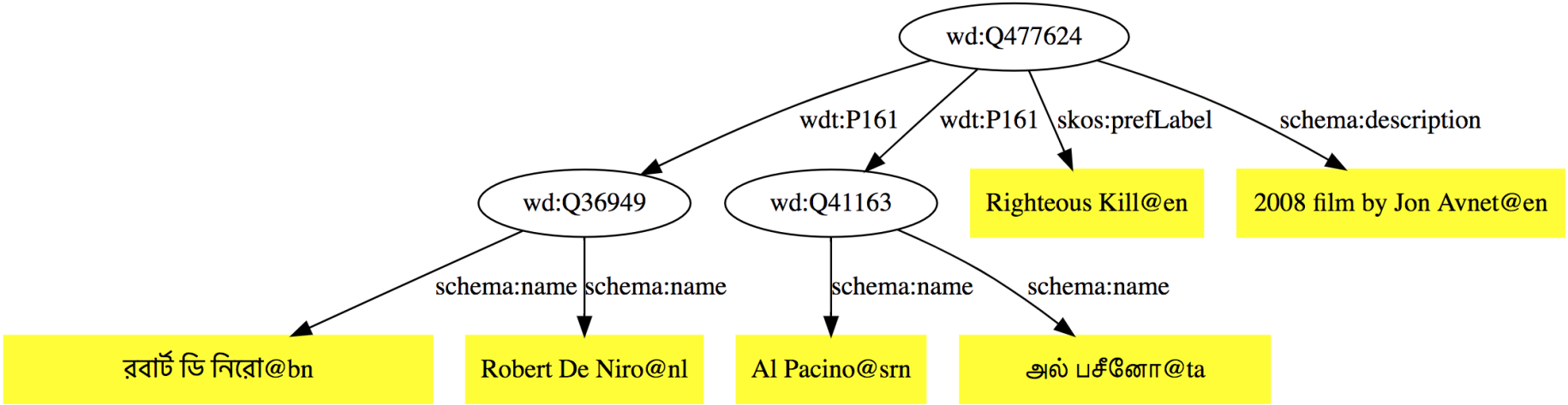
Wikisource (0 entries) edit

Wikiversity (0 entries) edit

Wikivoyage (0 entries) edit

Wiktionary (0 entries) edit

Other sites (0 entries) edit


wd:Q477624
wdt:P161
wdt:P161
skos:prefLabel
schema:description
wd:Q36949
wd:Q41163
Righteous Kill@en
2008 film by Jon Avnet@en
schema:name
schema:name
schema:name
schema:name
রবার্ট ডি নিরো@bn
Robert De Niro@nl
Al Pacino@srn
அல் பச்சீனோ@ta

Reelin

From Wikipedia, the free encyclopedia

Reelin (RELN)[5] 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 enhancing the induction and maintenance of long-term potentiation.[6][7] It also stimulates dendrite[8] and dendritic spine[9] development and regulates the continuing migration of neuroblasts generated in adult neurogenesis sites like subventricular and subgranular zones. It is found not only in the brain, but also in the spinal cord, blood, and other body organs and tissues.[*citation needed*]

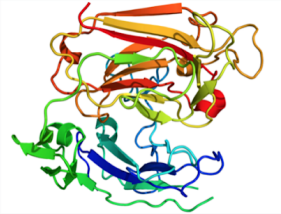
Reelin has been suggested to be implicated in pathogenesis of several brain diseases. The expression of the protein has been found to be significantly lower in schizophrenia and psychotic bipolar disorder,[10] but the cause of this observation remains uncertain as studies show that psychotropic medication itself affects reelin expression. Moreover, epigenetic hypotheses aimed at explaining the changed levels of reelin expression[11] are controversial.[12][13] Total lack of reelin causes a form of lissencephaly. Reelin may also play a role in Alzheimer's disease, temporal lobe epilepsy and autism.[*citation needed*]

Reelin's name comes from the abnormal reeling gait of *reeler* mice,[14] which were later found to have a deficiency of this brain protein and were homozygous for mutation of the RELN gene. The primary phenotype associated with loss of reelin function is a failure of neuronal positioning throughout the developing central nervous system (CNS). The mice heterozygous for the reelin gene, while having little neuroanatomical defects, display the endophenotypic traits linked to psychotic disorders.[15]

Contents [hide]

1 Discovery
2 Tissue distribution and secretion
3 Structure
4 Function
4.1 During development
4.2 In adults
5 Evolutionary significance
6 Mechanism of action

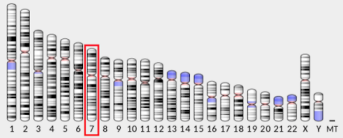
RELN



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

Identifiers	
Aliases	RELN, LIS2, PRO1598, RL, reelin, ETL7
External IDs	OMIM: 600514 MGI: 103022 HomoloGene: 3699 GeneCards: RELN

Gene location (Human) [hide]



Chr. Chromosome 7 (human)[1]

Reelin - Wikidata

Secure https://www.wikidata.org/wiki/Q13561329

English Andrawaag 21 99+ Talk Preferences Beta Watchlist Contributions Log out

WIKIDATA

Item Discussion Read View history More Search Wikidata

Reelin (Q13561329)

mammalian protein found in Homo sapiens edit

RELN | reelin | uniprot:P78509

In more languages

Main page
Community portal
Project chat
Create a new item
Recent changes
Random item
Query Service
Nearby
Help
Donate

Tools

What links here
Related changes
Special pages
Permanent link
Page information
Concept URI
Cite this page
Import interwiki

Statements

instance of	protein 1 reference	edit
		+ add value
subclass of	protein 1 reference	edit
	Reelin 1 reference	edit
		+ add value
image	2DDU.png 1 reference	edit

Retinoic acid receptor alpha (Q254943)

mammalian protein found in Homo sapiens

Nuclear receptor subfamily 1 group B member 1 | RARA

Statements

molecular function

molecular function (P680)

represents gene ontology function annotations

retinoic acid binding

edit

determination method IDA

1 reference

retrieved	3 January 2017
stated in	A human retinoic acid receptor which belongs to the family of nuclear receptors
	UniProt-GOA
curator	British Heart Foundation
reference URL	http://www.ebi.ac.uk/QuickGO/GAnnotation?protein=P10276
determination method	IDA

+ add reference

transcription corepressor activity

edit

determination method IDA

1 reference

retinoic acid binding (Q14901431)

Interacting selectively and non-covalently with retinoic acid, 3,7

GO:0001972

Statements

subclass of	retinoid binding
	1 reference

IDA (Q23174122)

Gene Ontology evidence code

Inferred from Direct Assay

Statements

instance of	Gene Ontology Evidence code
	manual assertion

A human retinoic acid receptor which belongs to the family of nuclear receptors (Q24339631)

Statements

instance of	scientific article

Identifiers

PubMed ID	2825025

British Heart Foundation (Q4970039)

Statements

instance of	organization
official website	http://www.bhf.org.uk/

Identifiers

GRID ID	grid.452924.c

Wikipedia (7 entries) edit

ar	مستقبل حمض الريتينويك ألفا
en	Retinoic acid receptor alpha
es	Receptor de ácido retinoico alfa
sh	Receptor retinoinske kiseline alfa
sr	Receptor retinoinske kiseline alfa
uk	RARA
zh	视黄酸受体α

Item Discussion Read View history Search W

Revision history of "Retinoic acid receptor alpha" (Q254943)

View logs for this item

Search for revisions

From year (and earlier): 2017 From month (and earlier): all Tag filter: Show

Diff selection: Mark the radio boxes of the revisions to compare and hit enter or the button at the bottom.

Legend: **(cur)** = difference with latest revision, **(prev)** = difference with preceding revision, **m** = minor edit.

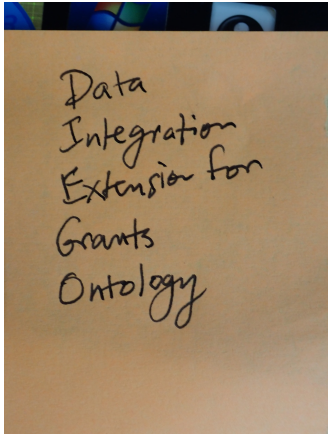
Compare selected revisions

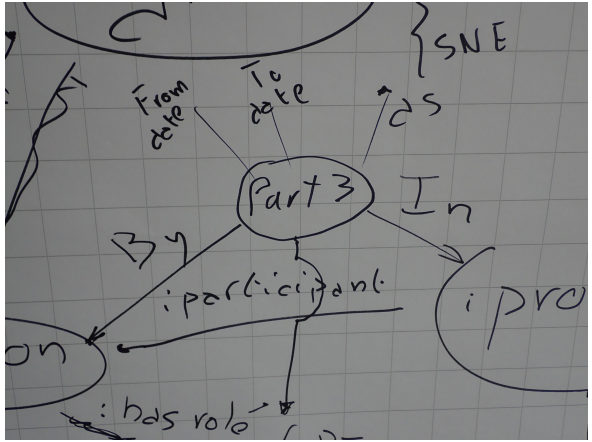
Select: All, None, Invert

- (cur | prev) 20:13, 21 March 2017 ProteinBoxBot (talk | contribs) . . (454,236 bytes) (-440) . . *(Updated item: replace thumbnail gene atlas image with fs)* (undo)
- (cur | prev) 08:00, 28 January 2017 Edoderoobot (talk | contribs) . . (454,676 bytes) (+67) . . *(Updated item: #proteine)* (undo) (restore)
- (cur | prev) 12:06, 4 January 2017 ProteinBoxBot (talk | contribs) . . (454,609 bytes) **(+165,607)** . . *(Updated item: update GO terms)* (undo) (restore)
- (cur | prev) 03:57, 3 January 2017 ProteinBoxBot (talk | contribs) . . (289,002 bytes) **(+1,584)** . . *(Updated item)* (undo) (restore)
- (cur | prev) 09:07, 17 September 2016 Okkn (talk | contribs) . . (287,418 bytes) (-2) . . *(Changed claim: subclass of (P279): Retinoic acid receptor (Q2838685))* (undo | thank) (restore)
- (cur | prev) 15:18, 16 September 2016 ProteinBoxBot (talk | contribs) . . (287,420 bytes) (-292) . . *(Updated item)* (undo) (restore)
- (cur | prev) 12:03, 17 August 2016 ProteinBoxBot (talk | contribs) . . (287,712 bytes) (0) . . *(Updated item)* (undo) (restore)
- (cur | prev) 04:50, 9 August 2016 ProteinBoxBot (talk | contribs) . . (287,712 bytes) **(+11,503)** . . *(Updated item)* (undo) (restore)

Getting data in..

Community engagement and model discussion

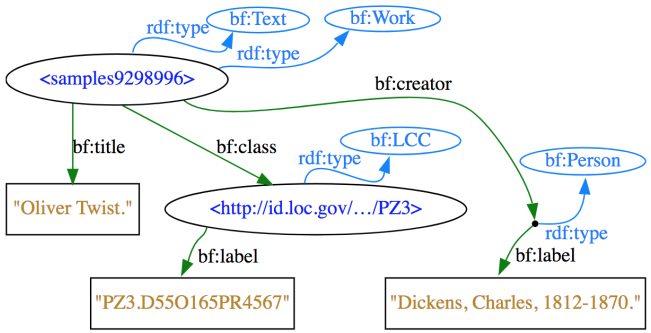




Formally capture and describe model and community consensus

Model development

- Legacy review – develop punch lists for existing data issues that needs fixing
- Documentation – terse, human-readable representation helping contributers and maintainers quickly grok the model
- Client pre-submission – submitters test their data before submission to make sure they're saying what they want to say and that the receiving schema can accommodate all of their data
- Server pre-ingestion – submission process checks data as it comes in and either rejects or warns about non-conformant data



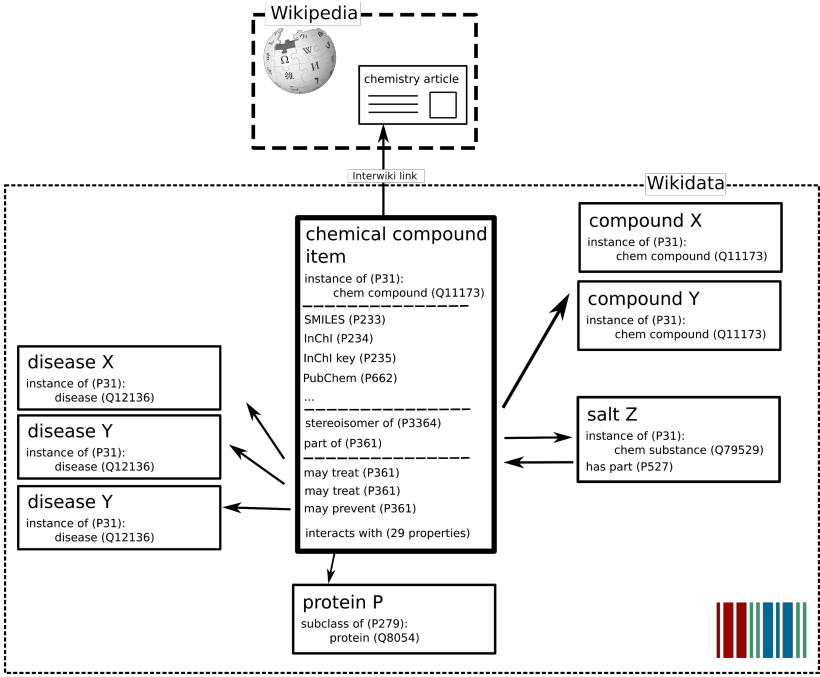
Data (Turtle)

```
<samples9298996>
  rdf:type bf:Text ;
  rdf:type bf:Work ;
  bf:title "Oliver Twist." ;
  bf:class <id.loc.gov/…/PZ3> ;
  bf:creator [
    rdf:type bf:Person ;
    bf:label "Dickens, Charles, 1812-1870." ;
  ] .

<id.loc.gov/…/PZ3>
  rdf:type bf:LCC ;
  bf:label "PZ3.D55O165PR4567" .
```

Seeding with data

- Model structure of items (genes, drugs, diseases, .. etc) & relationships between items
- Import data from many sources and ontologies
- Linked to many identifiers from external databases
- Architecture for maintaining data from external sources



SuLab / **WikidataIntegrator**

Watch 7 | Star 16 | Fork 6

Code | Issues 4 | Pull requests 1 | Projects 0 | Pulse | Graphs

A Wikidata Python module integrating the MediaWiki API and the Wikidata SPARQL endpoint

397 commits | **2** branches | **1** release | **7** contributors | MIT

Branch: **master** | New pull request | Find file | Clone or download

sebotic fixed an omission where new items don't get created when domain not s… ··· Latest commit `2f5d2fd` 22 hours ago

doc	Wikidata to Wikipedia mapping prototype for diseases added.	2 years ago
wikidataintegrator	fixed an omission where new items don't get created when domain not s…	22 hours ago

Jenkins

Jenkins › Running ›

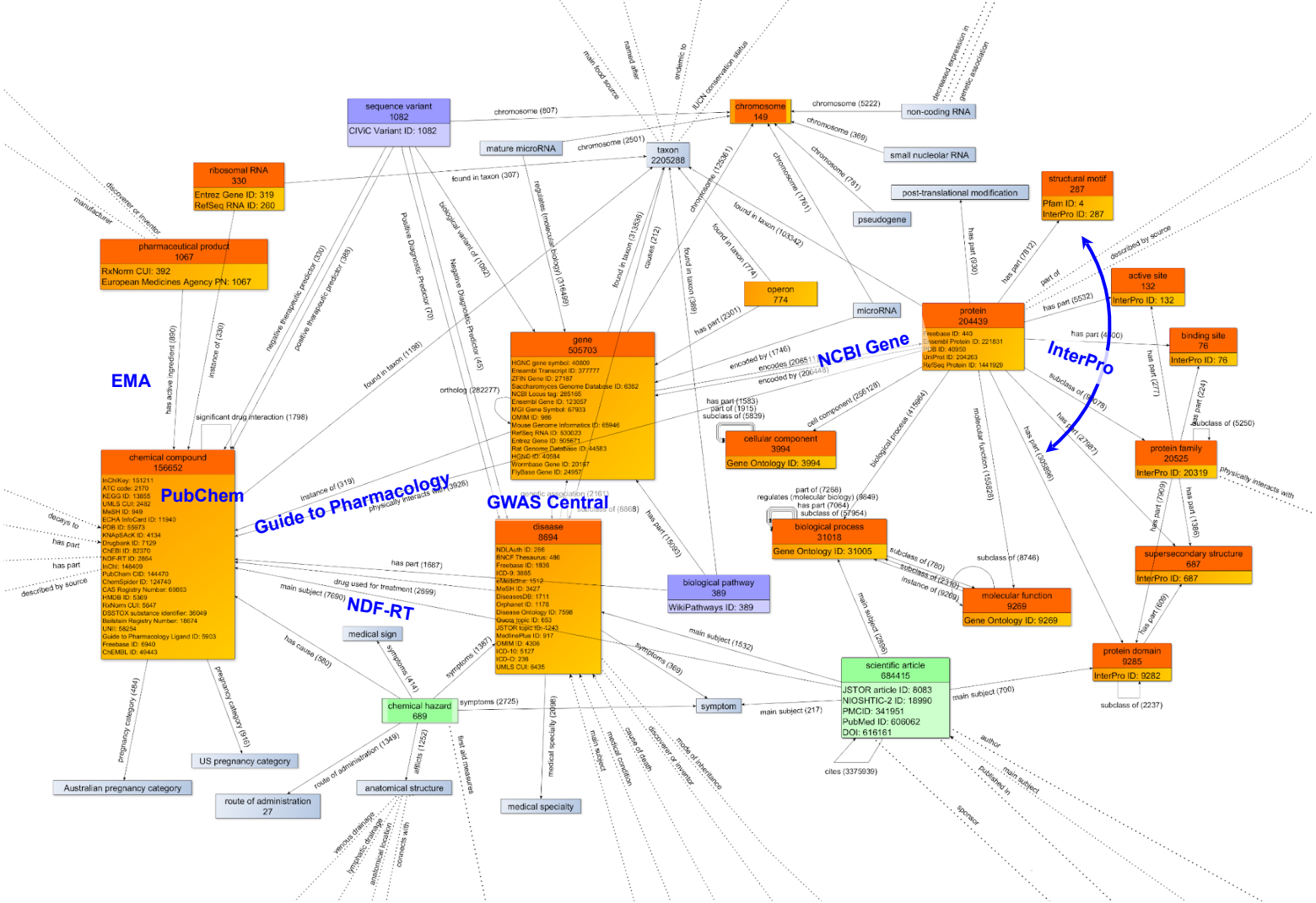
- New Item
- People
- Build History
- Edit View
- Delete View
- Manage Jenkins
- My Views
- Credentials

Build Queue

Running Bots

All | **Running** | +

S	Name	Last Success ↑	Last Failure
	ProteinBot_homo_sapiens	1 day 21 hr - #12	N/A
	GOBot_bigmem	2 days 15 hr - #15	9 days 15 hr - #14
	GeneBot_Homo_sapiens	2 days 19 hr - #25	2 days 20 hr - #24
	Disease_Ontology	2 days 23 hr - #11	4 days 13 hr - #8
	GeneDiseaseBot	2 days 23 hr - #9	1 mo 6 days - #2


EMA
PubChem
Guide to Pharmacology
GWAS Central
NDF-RT
NCBI Gene
InterPro
sequence variant 1082
CIVIC Variant ID: 1082
chromosome 149
non-coding RNA
mature microRNA
taxon 2205288
small nucleolar RNA
ribosomal RNA 330
Entrez Gene ID: 319
RefSeq RNA ID: 260
post-translational modification
structural motif 287
Pfam ID: 4
InterPro ID: 287
pseudogene
pharmaceutical product 1067
RxNorm CUI: 392
European Medicines Agency PN: 1067
operon 774
microRNA
protein 204439
active site 132
InterPro ID: 132
binding site 76
InterPro ID: 76
gene 505703
chemical compound 156652
cellular component 3994
Gene Ontology ID: 3994
protein family 20525
InterPro ID: 20319
disease 8694
biological process 31018
Gene Ontology ID: 31005
supersecondary structure 687
InterPro ID: 687
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

Use..

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
Q1802737			Q1803503			Q1803018	

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

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 ?gen
(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

```
?exposure wdt:P1542 ?diseaseGA .  # something causes disease
?exposure wdt:P279 wd:Q21167512 . # and that something is a chemical hazard

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

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."

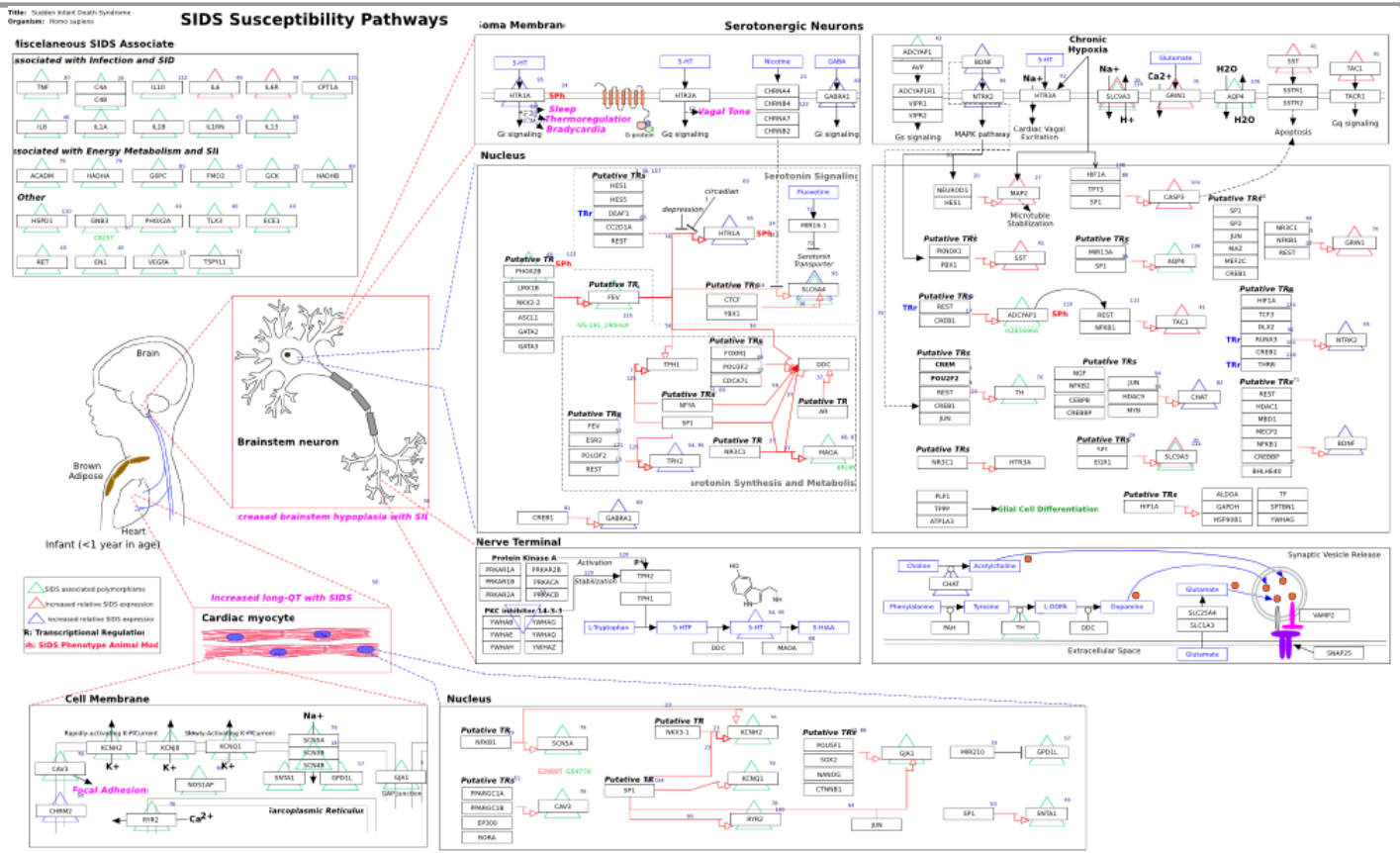
16 genes / 59 pathways

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

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

Try me....

From a remote SPARQL endpoint to Wikidata

UniProt

SPARQL Downloads Documentation/Help Contact

Your query

Add common prefixes

```
SELECT DISTINCT ?wd_item ?physically_interacts_with ?interactswithLabel ?type ?iri ?uniprot ?text  WHERE  {
    {SELECT * WHERE { ?iri a up:Protein ;
            up:organism taxon:9606 ;
            up:annotation ?annotation .
        ?annotation a up:Natural_Variant_Annotation ;
                    rdfs:comment ?text .
        FILTER (CONTAINS(?text, 'loss of function'))
    }}
    SERVICE <https://query.wikidata.org/bigdata/namespace/wdq/sparql> {
      VALUES ?use {wd:Q427492}
    ?wd_item wdt:P352 ?uniprot ;
             wdt:P129 ?physically_interacts_with ;
        wdt:P2888 ?iri ;
        wdt:P703 wd:Q15978631 .
    ?wd_item p:P129 ?phys_interacts_with_node .
    ?phys_interacts_with_node ps:P129 ?physically_interacts_with ;
                              pq:P366 ?use .
    ?physically_interacts_with wdt:P31 ?type ;
                               rdfs:label ?interactswithLabel .
    FILTER (lang(?interactswithLabel) = "en")
    }}
```

UniProt

Wikidata

Submit Query Cancel

© 2002– 2017 UniProt Consortium | License & Disclaimer

EMBL-EBI PIR SIB

Cellosaurus
@Cellosaurus

Following

ITS DONE!! All 107'576 #Cellosaurus #celllines are now in #Wikidata. There is still some work to be carried out in the next weeks in term of disease mappings but its a major step. Thanks again to Lelia Debornes who wrote the bot and to @andrawaag who was of tremendous help

6:09 PM - 16 Jun 2018

Example query Cellosaurus en Pathways

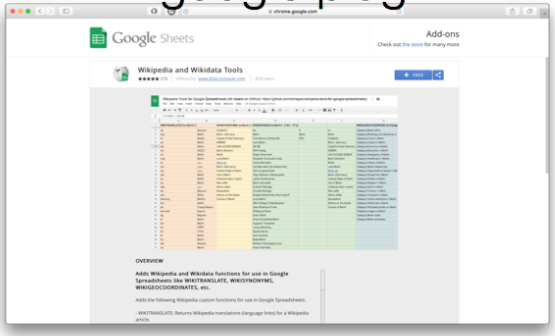
Which papers do Cellines and Pathway Share

Tooling..

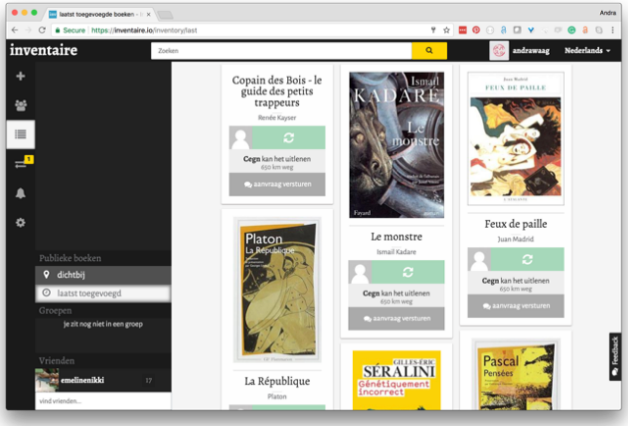
Tools using Wikidata

http://www.wikigenomes.org

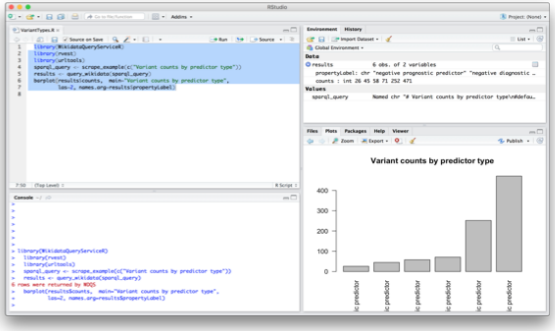
Wikipedia and Wikidata google plugin



http://inventaire.io



Cytoscape



R plugins

Chlambase.org and Wikigenomes.org

WikiGenomes

Organism Search: choose a new organism

Authorize to Edit Wikidata

Helicobacter pylori 26695

Organism | Annotation Landscape

Proteins	Genes
1422	1415

NCBI Taxonomy ID:
85962

Wikidata Item ID:
Q21065231

NC_000915.1 | NC_000915.1:67020..77817 (10.8 Kb) | Go

Select tracks

HP0066 ATP-binding protein HP0066
HP0070 urease accessory protein UreE HP0070
HP0073 urease subunit
HP0064 hypothetical protein HP0064
HP0068 urease accessory protein UreG HP0068
HP0072 urease subunit beta HP0072
HP0063 hypothetical protein HP0063
HP0065 hypothetical protein HP0065
HP0067 urease accessory protein UreH HP0067
HP0071 acid-activated urea channel HP0071
HP0069 urease accessory protein UreF HP0069

Operons

Gene Search: choose a gene

Gene | Protein

Gene Name:
urease accessory protein UreH HP0067

Entrez ID:
899986

Wikidata ID:
Q21629448

NCBI Locus Tag:
HP0067

Molecular Function (2) | Biological Process (2) | Cellular Component (1) | InterPro Domains (0) | Enzyme Activity (0) | Operon (0) | Genomic Position

Molecular Function	Gene Ontology ID	Evidence
protein binding	GO:0005515	IPI
nickel cation binding	GO:0016151	IEA

Availability

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
- github.com/SuLab/**Genewiki-ShEx** – data models
- github.com/SuLab/**SPARQL_Examples** – SPARQL-query examples data models
- github.com/SuLab/**wikibase-jenkins** – Wikibase instance with Jenkins

MIT License

Conclusions

- Wikidata is a public linked data resource
- Wikidata is a community of communities
- Wikidata applies a CC0 license (public data)
- Integration with non-CC0 data through captured links and federated SPARQL queries
- Metadata on genes, proteins, diseases and drugs are regularly synchronised (Gene wiki)
- No prior permission or vetting needed to add your content to Wikidata

DRUGBANK
NCBI
DISEASE ONTOLOGY
PubChem
WIKIDATA

Demo

- Translations of English disease terms in South African languages