Gene Wiki: Making data FAIR on Wikidata

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

Micelio, Ekeren, Antwerp, Belgium

Structure

- Introduction
  - Wikidata
  - Wikibase
  - Gene Wiki
- Bits and Bolts
  - How to get your data into Wikidata/Wikibase
  - Using Wikidata/Wikibase
- Is Wikidata FAIR?
  - Where is it located in the research life cycle
  - How FAIR is Wikidata/Wikibase/Genewiki
  - interoperate/collaborate/engage
  - one key change to successfully implement FAIR data infrastructure

Wikidata and its sisters

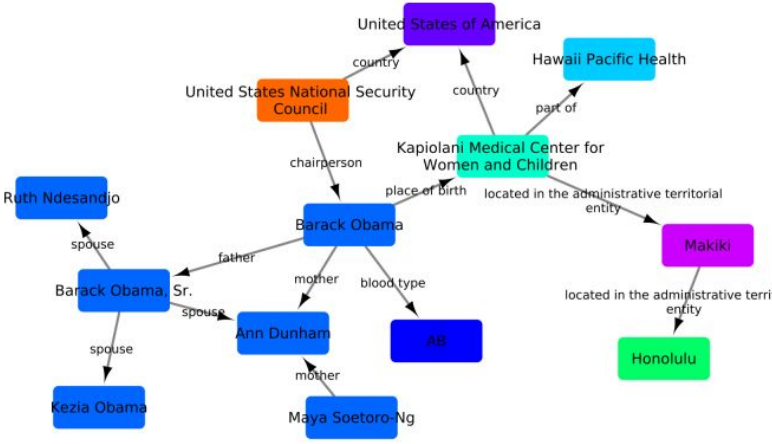
Wikidata is to data as Wikipedia is to text

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 cycles
- Actively developed
- Already has large number of active users, editors, contributors!

A giant graph of knowledge!



FAIR on Wikidata The Wikimedia infrastructure

Infrastructure	Resource	Content	Where?
WIKIBASE	WIKIDATA	Data	https://www.wikidata.org https://www.wikibase.org
MediaWiki	WIKIPEDIA The Free Encyclopedia	Text	https://<lang>.wikipedia.org https://releases.wikimedia.org/mediawiki/
	ShEx	**Under developement** Schema	https://shex.io https://wikidata-shex.wmflabs.org/

FAIR on Wikidata: Providing identifiers

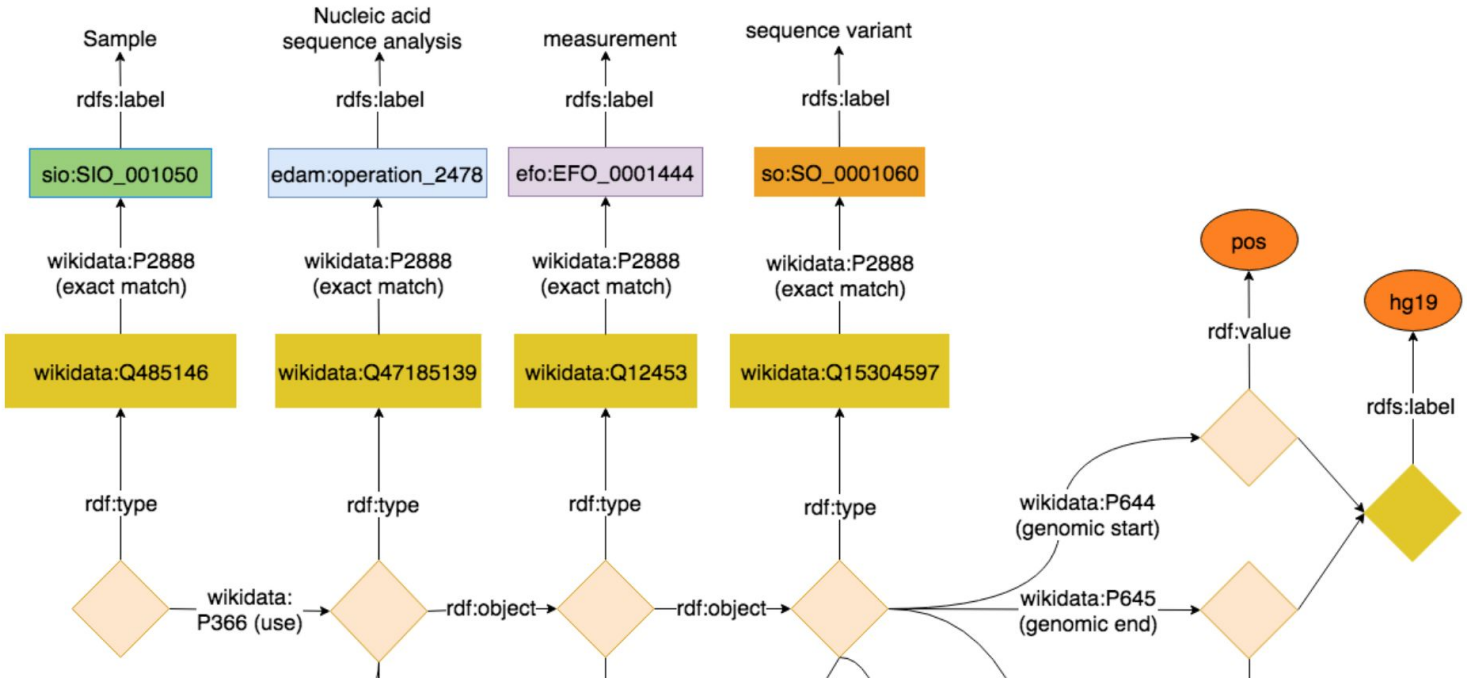
Wikidata as an intuitive resource towards semantic data modeling in data FAIRification

Annika Jacobsen[1][0000−0003−4818−2360], Andra Waagmeester[2][0000−0001−9773−4008], Rajaram Kaliyaperumal[1][0000−0002−1215−167X], Gregory S. Stupp[3][0000−0002−0644−7212], Lynn M. Schriml[4][0000−0001−8910−9851], Mark Thompson[1][0000−0002−7633−1442], Andrew I. Su[3][0000−0002−9859−4104], and Marco Roos[1][0000−0002−8691−772X]

[1] Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
[2] Micelio, Antwerp - Ekeren, Belgium
[3] The Scripps Research Institute, San Diego CA 92037, USA
[4] University of Maryland School of Medicine, Baltimore, MD, USA

http://ceur-ws.org/Vol-2275/short1.pdf

FAIR on Wikidata: Providing identifiers



Reelin - Wikipedia

Andra

Secure | https://en.wikipedia.org/wiki/Reelin

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

Article Talk | Read Edit View history More | Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

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

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

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

Print/export
Create a book
Download as PDF
Printable version

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*]

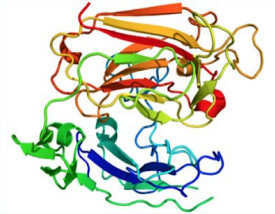
Reelin has been suggested to be implicated in pathogenesis of several brain diseases. The expression of the protein has been found to be significantly lower in schizophrenia and psychotic bipolar disorder,[10] but the cause of this observation remains uncertain as studies show that psychotropic medication 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

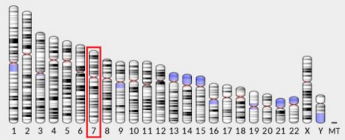
RELN



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

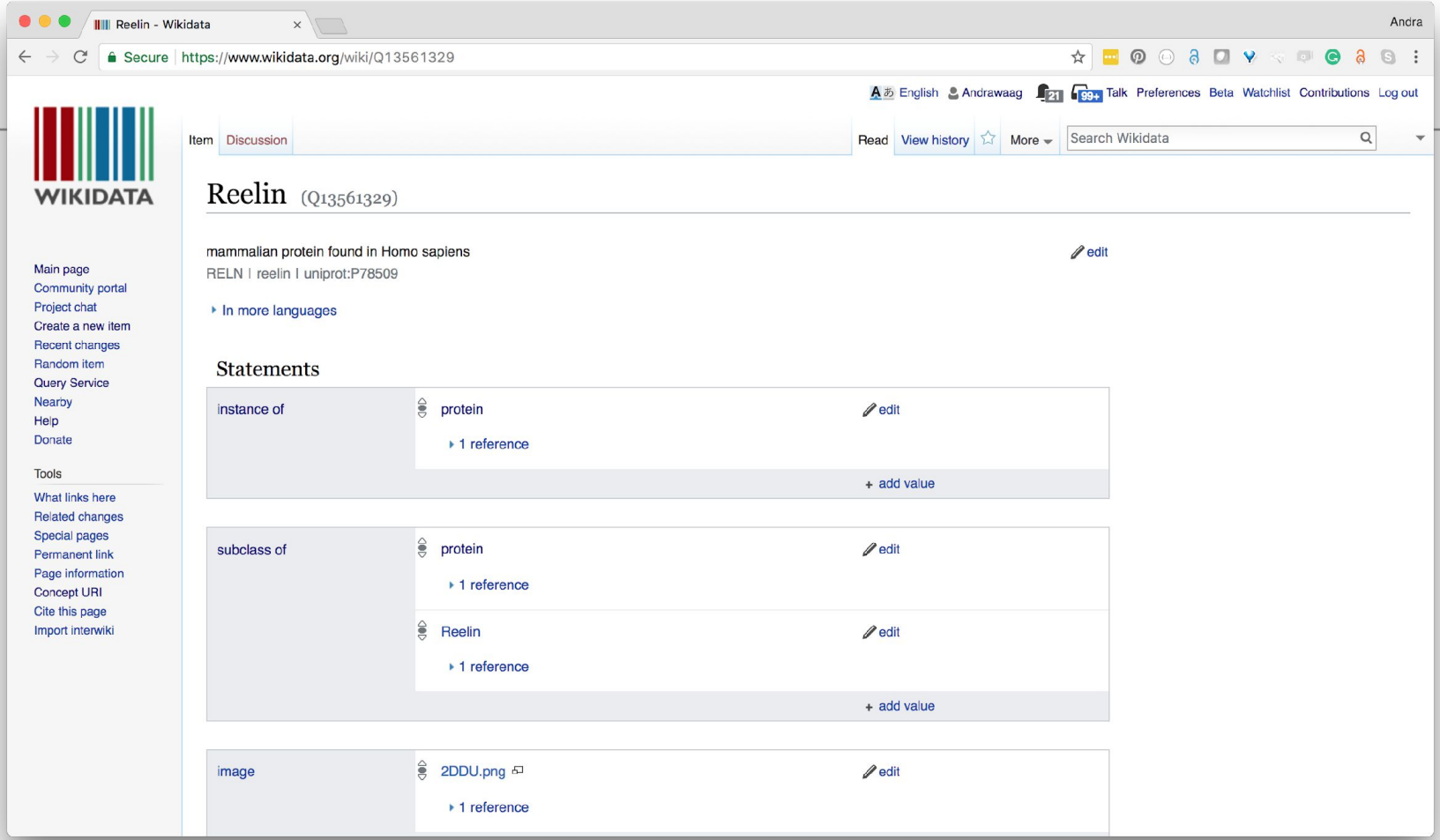
Identifiers	
Aliases	RELN, LIS2, PRO1598, RL, reelin, ETL7
External IDs	OMIM: 600514 MGI: 103022 HomoloGene: 3699 GeneCards: RELN

Gene location (Human) [hide]



Chr. Chromosome 7 (human)[1]

https://en.wikipedia.org/wiki/Reelin#Psychotropic_medication


Reelin - Wikidata
Andra
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
edit
1 reference
add value
subclass of
protein
edit
1 reference
Reelin
edit
1 reference
add value
image
2DDU.png
edit
1 reference

Retinoic acid receptor alpha (Q254943)

mammalian protein found in Homo sapiens

Nuclear receptor subfamily 1 group B member 1 | RARA

Statements

molecular function

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

molecular function (P680)

represents gene ontology function annotations

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

Wikidata Query Service

https://query.wikidata.org/#%23A%20network%20of%20Drug-disease%20interactions%20on%2...

Wikidata Query Service | Examples | Help | More tools | English

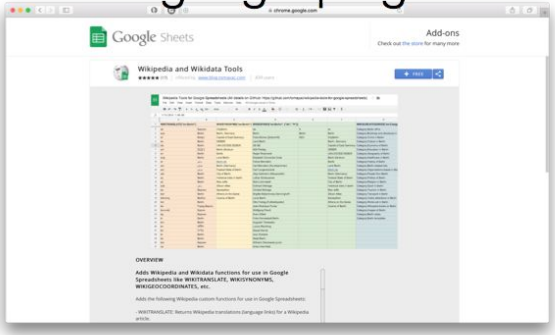
```
#A network of Drug-disease interactions on infectious diseases (Source: Disease Ontology, NDF-RT and ChEMBL)
#added before 2016-10
#defaultView:Graph
SELECT DISTINCT ?item ?itemLabel ?rgb ?link
WHERE
{
  VALUES ?toggle { true false }
  ?disease wdt:P699 ?doid;
           wdt:P279+ wd:Q18123741;
           wdt:P2176 ?drug.
  ?drug rdfs:label ?drugLabel.
        FILTER(LANG(?drugLabel) = "en").
  ?disease rdfs:label ?diseaseLabel.
        FILTER(LANG(?diseaseLabel) = "en").
  BIND(IF(?toggle,?disease,?drug) AS ?item).
  BIND(IF(?toggle,?diseaseLabel,?drugLabel) AS ?itemLabel).
  BIND(IF(?toggle,"FFA500","7FFF00") AS ?rgb).
  BIND(IF(?toggle,"",?disease) AS ?link).
}
```

Graph | 1177 results in 346 ms | Code | Download | Link

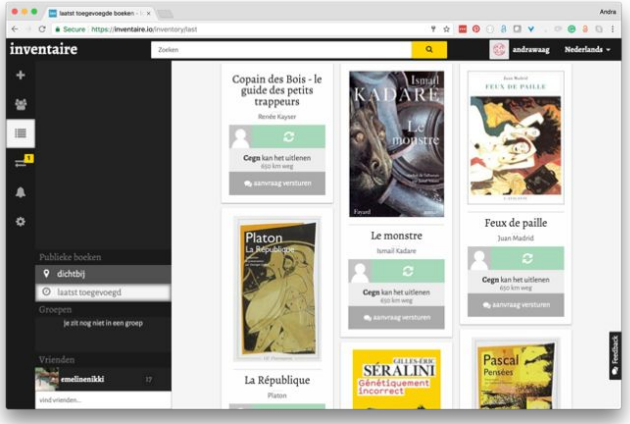
Tools using Wikidata

http://www.wikigenomes.org

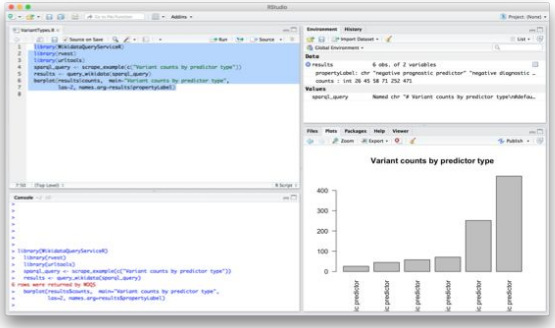
Wikipedia and Wikidata google plugin



http://inventaire.io



Cytoscape

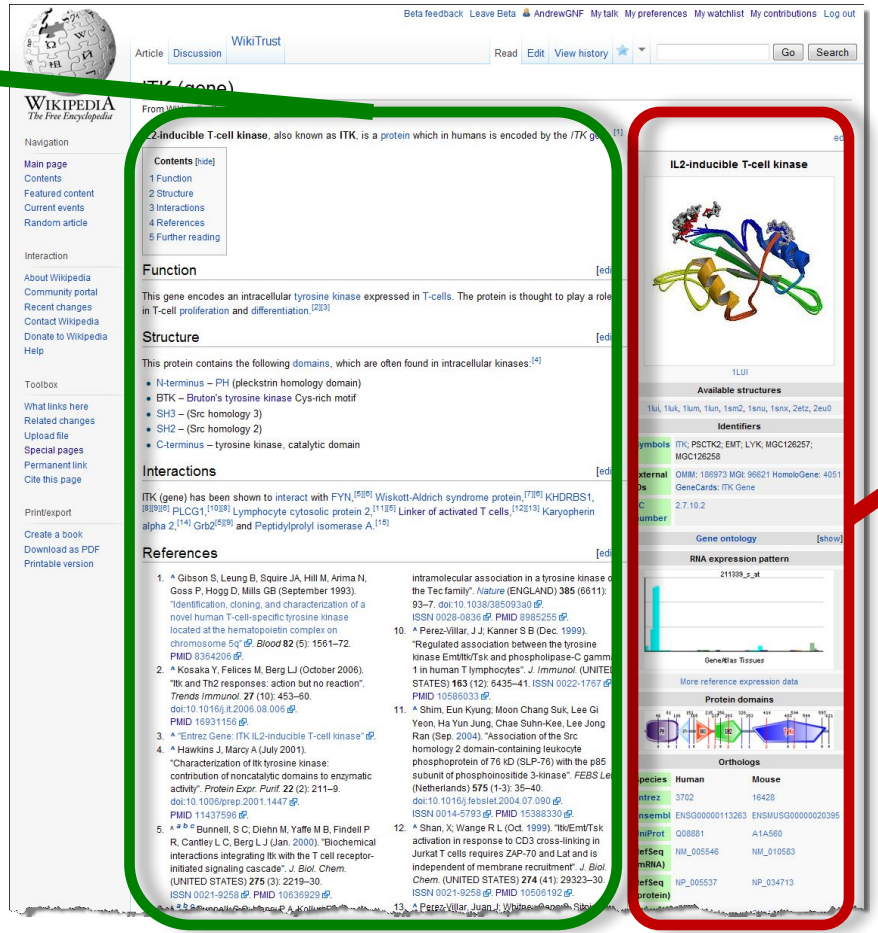


R plugins

Wikidata and its sisters

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

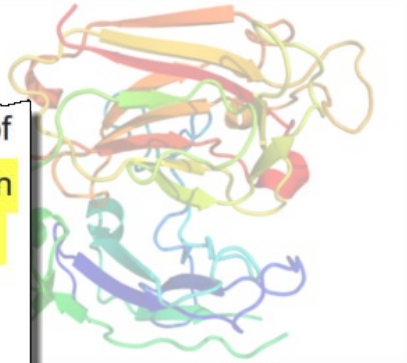
Reelin is a large secreted extracellular matrix glycoprotein that helps regulate processes of neuronal migration and positioning in the developing brain by controlling cell–cell interactions. Besides this important role in early development, reelin continues to work in the adult brain. It modulates synaptic 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

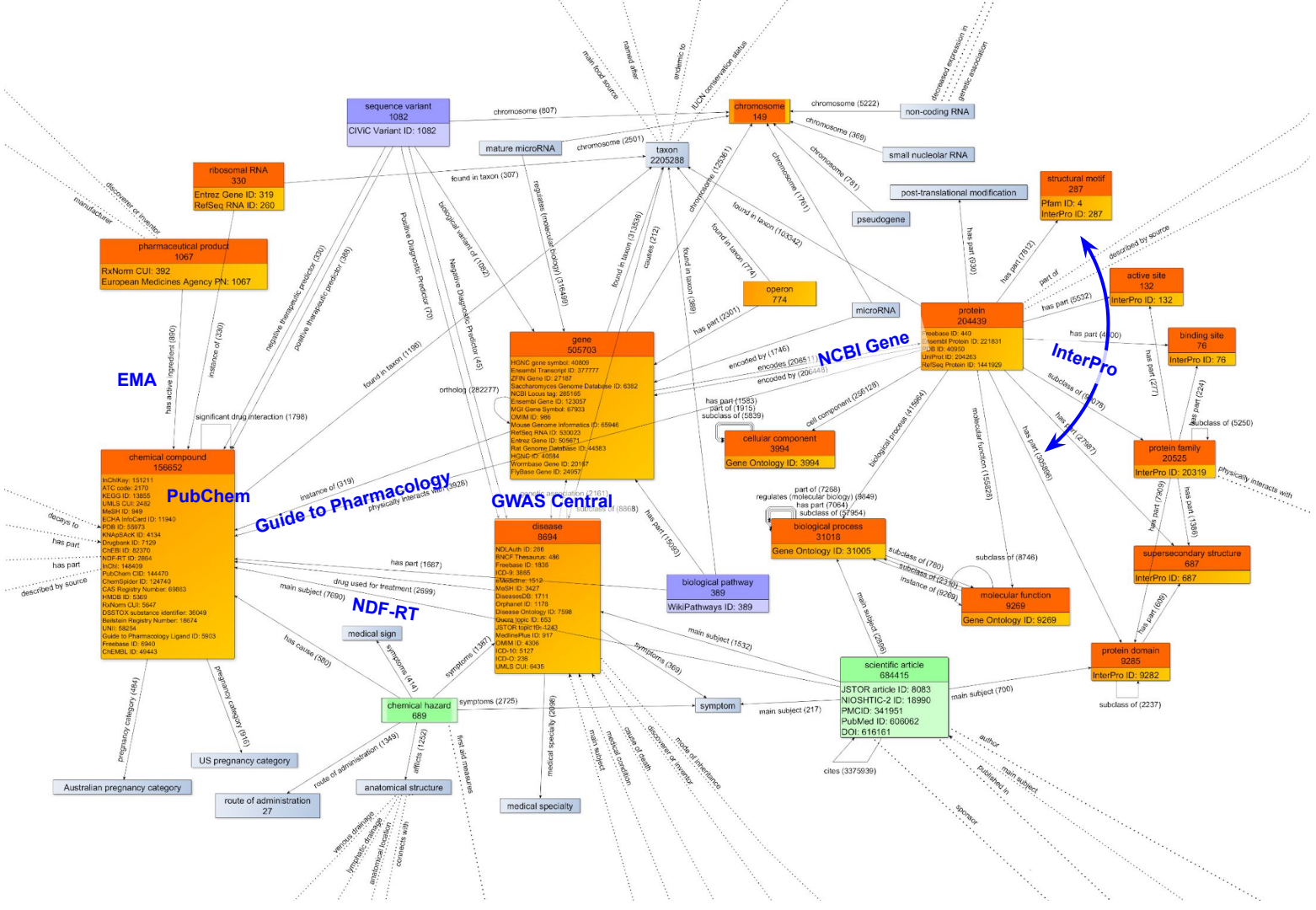
Reelin



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


sequence variant
1082
CIViC Variant ID: 1082
chromosome 149
chromosome (807)
chromosome (5222)
non-coding RNA
chromosome (369)
small nucleolar RNA
mature microRNA
chromosome (2501)
taxon
2205288
ribosomal RNA
330
Entrez Gene ID: 319
RefSeq RNA ID: 260
found in taxon (307)
pseudogene
post-translational modification
structural motif
287
Pfam ID: 4
InterPro ID: 287
pharmaceutical product
1067
RxNorm CUI: 392
European Medicines Agency PN: 1067
operon
774
found in taxon (1033)
found in taxon (774)
microRNA
protein
204439
Freebase ID: 440
Ensembl Protein ID: 221831
SGD ID: 40950
UniProt ID: 294263
RefSeq Protein ID: 1441929
has part (7812)
has part (930)
part of
has part (5532)
active site
132
InterPro ID: 132
binding site
76
InterPro ID: 76
has part (4430)
has part (277)
has part (224)
subclass of (5250)
gene
505703
HGNC gene symbol: 40809
Ensembl Transcript ID: 377777
ZFIN Gene ID: 27187
Saccharomyces Genome Database ID: 6382
NCBI Locus tag: 285165
Ensembl Gene ID: 123057
MGI Gene Symbol: 67933
OMIM ID: 986
Mouse Genome Informatics ID: 65946
RefSeq RNA ID: 830023
Entrez Gene ID: 505671
Rat Genome Database ID: 44583
HGNC ID: 40594
Wormbase Gene ID: 20167
FlyBase Gene ID: 24957
encoded by (1746)
encodes (20651)
encoded by (20645)
has part (1583)
part of (1915)
subclass of (5839)
cellular component
3994
Gene Ontology ID: 3994
cell component (256128)
biological process (159864)
molecular function (153928)
has part (27987)
protein family
20525
InterPro ID: 20319
physically interacts with
chemical compound
156652
InChIKey: 151211
ATC code: 2170
KEGG ID: 13855
UMLS CUI: 2482
MeSH ID: 949
ECHA InfoCard ID: 11940
PDB ID: 55673
KNApSAcK ID: 4134
Drugbank ID: 7129
ChEBI ID: 82370
NDF-RT ID: 2864
InChI: 146409
PubChem CID: 144470
ChemSpider ID: 124740
CAS Registry Number: 69883
HMDB ID: 5369
RxNorm CUI: 5647
DSSTOX substance identifier: 36049
Beilstein Registry Number: 18674
UNII: 58254
Guide to Pharmacology Ligand ID: 5903
Freebase ID: 6940
ChEMBL ID: 49443
PubChem
Guide to Pharmacology
GWAS Central
NCBI Gene
InterPro
EMA
NDF-RT
part of (7268)
regulates (molecular biology) (8649)
has part (7064)
subclass of (57954)
biological process
31018
Gene Ontology ID: 31005
subclass of (780)
subclass of (2310)
instance of (9269)
subclass of (8746)
molecular function
9269
Gene Ontology ID: 9269
supersecondary structure
687
InterPro ID: 687
has part (7409)
has part (1365)
has part (605)
protein domain
9285
InterPro ID: 9282
subclass of (2237)
disease
8694
NDLAuth ID: 286
BNCF Thesaurus: 496
Freebase ID: 1936
ICD-9: 3855
eMedicine: 1512
MeSH ID: 3427
DiseasesDB: 1711
Orphanet ID: 1178
Disease Ontology ID: 7598
Gwoza topic ID: 653
JSTOR topic ID: 1243
MedlinePlus ID: 917
OMIM ID: 4306
ICD-10: 5127
ICD-O: 238
UMLS CUI: 6435
has part (1687)
drug used for treatment (2899)
main subject (7690)
biological pathway
389
WikiPathways ID: 389
has part (1583)
medical sign
chemical hazard
689
symptoms (2725)
symptom
main subject (217)
main subject (1532)
symptoms (369)
symptoms (1387)
symptoms (414)
has cause (580)
scientific article
684415
JSTOR article ID: 8083
NIOSHTIC-2 ID: 18990
PMCID: 341951
PubMed ID: 606062
DOI: 616161
main subject (2898)
main subject (700)
cites (3375939)
author
published in
sponsor
pregnancy category (464)
pregnancy category (916)
US pregnancy category
Australian pregnancy category
route of administration (1349)
route of administration
27
afflicts (1252)
anatomical structure
first aid measures
medical specialty
medical specialty (2098)
has part
decays to
has part
described by source
manufacturer
discoverer of (inventor
has active ingredient (980)
instance of (330)
significant drug interaction (1798)
negative therapeutic predictor (330)
positive therapeutic predictor (386)
Positive Diagnostic Predictor (70)
Negative Diagnostic Predictor (45)
biological variant of (1062)
found in taxon (1196)
regulates (molecular biology) (316499)
found in taxon (313556)
causes (212)
found in taxon (369)
chromosome (12361)
chromosome (1761)
chromosome (781)
has part (2501)
ortholog (282277)
paralog (2161)
subclass of (8868)
has part (1583)
has part (1687)
instance of (319)
physically interacts with (3926)
has cause (580)
symptoms (1387)
symptoms (414)
symptoms (369)
main subject (1532)
main subject (7690)
main subject (217)
main subject (2898)
main subject (700)
pregnancy category (464)
pregnancy category (916)
route of administration (1349)
afflicts (1252)
first aid measures
medical specialty (2098)
medical condition
cause of death
discoverer or inventor
mode of inheritance
main subject
author
published in
sponsor
cites (3375939)
various drainage
lymphatic drainage
anatomical location
connects with
named after
main food source
endemic to
IUCN conservation status
decreased expression in
genetic association
chromosome (5222)
chromosome (369)
chromosome (781)
chromosome (1761)
chromosome (12361)
found in taxon (307)
found in taxon (1033)
found in taxon (774)
found in taxon (313556)
found in taxon (369)
regulates (molecular biology) (316499)
causes (212)
has part (2501)
has part (930)
has part (7812)
part of
has part (5532)
has part (4430)
has part (277)
has part (224)
subclass of (5250)
encoded by (1746)
encodes (20651)
encoded by (20645)
has part (1583)
part of (1915)
subclass of (5839)
cell component (256128)
biological process (159864)
molecular function (153928)
has part (27987)
has part (305696)
part of (7268)
regulates (molecular biology) (8649)
has part (7064)
subclass of (57954)
subclass of (780)
subclass of (2310)
instance of (9269)
subclass of (8746)
has part (7409)
has part (1365)
has part (605)
subclass of (2237)
physically interacts with
ortholog (282277)
paralog (2161)
subclass of (8868)
has part (1583)
has part (1687)
instance of (319)
physically interacts with (3926)
has cause (580)
symptoms (1387)
symptoms (414)
symptoms (369)
main subject (1532)
main subject (7690)
drug used for treatment (2899)
main subject (217)
main subject (2898)
main subject (700)
pregnancy category (464)
pregnancy category (916)
route of administration (1349)
afflicts (1252)
first aid measures
medical specialty (2098)
medical condition
cause of death
discoverer or inventor
mode of inheritance
main subject
author
published in
sponsor
cites (3375939)
various drainage
lymphatic drainage
anatomical location
connects with
has part
decays to
has part
described by source
manufacturer
discoverer of (inventor
has active ingredient (980)
instance of (330)
significant drug interaction (1798)
negative therapeutic predictor (330)
positive therapeutic predictor (386)
Positive Diagnostic Predictor (70)
Negative Diagnostic Predictor (45)
biological variant of (1062)
found in taxon (1196)
named after
main food source
endemic to
IUCN conservation status
decreased expression in
genetic association
described by source

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

Gene Search: choose a gene

Gene | Protein

Gene Name: urease accessory protein UreH HP0067

Entrez ID: 899986

Wikidata ID: Q21629448

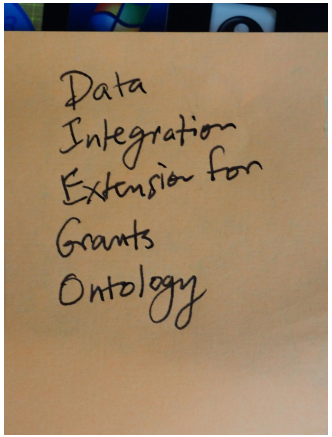
NCBI Locus Tag: HP0067

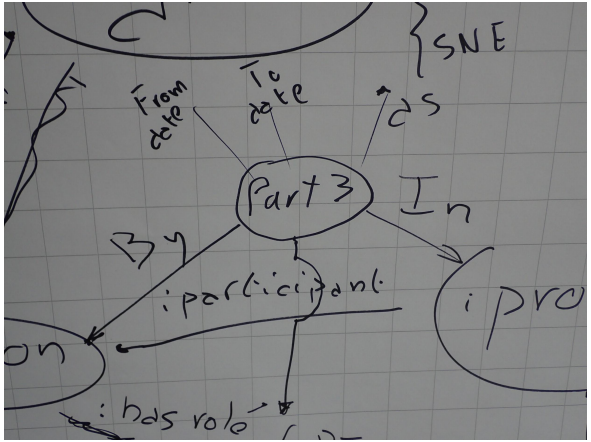
Molecular Function (2) | Biological Process (2) | Cellular Component (1) | InterPro Domains (0) | Enzyme Activity (0) | Operon (0) | Genomic Position

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

FAIRification

Community engagement and model discussion

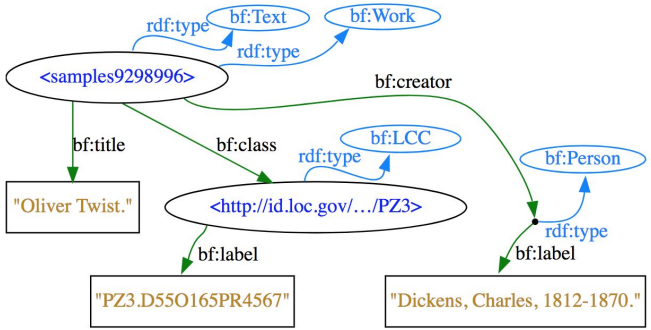




Formally capture and describe model and community consensus

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



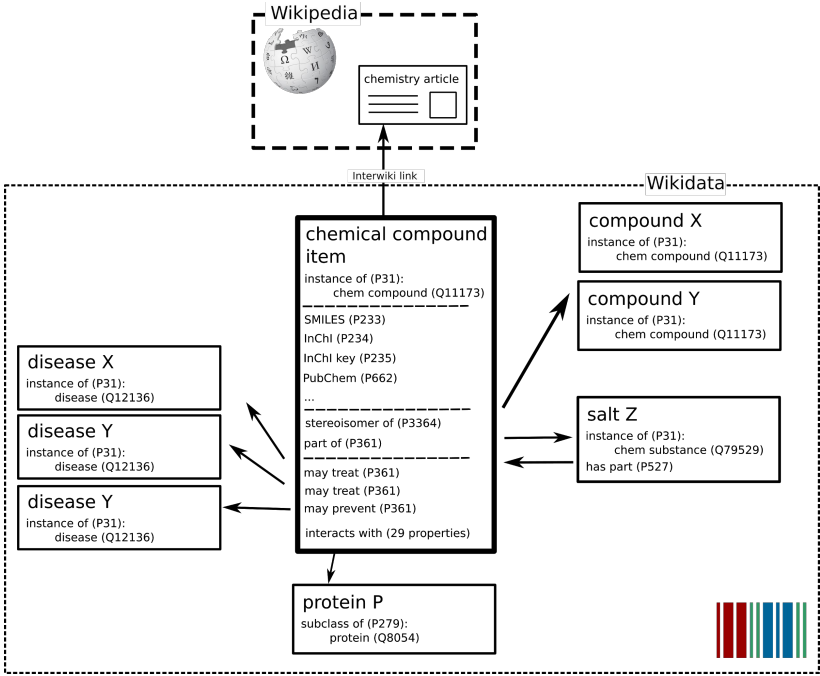
Data (Turtle)

```
<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



Platforms for (semi-) automatic data ingestion

- Quickstatements: https://tools.wmflabs.org/quickstatements/
- Open Refine: https://www.wikidata.org/wiki/Wikidata:Tools/OpenRefine
- Wikidata Integrator: https://github.com/SuLab/WikidataIntegrator

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

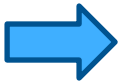
Running Bots

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

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

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

```
1  SELECT ?diseaseGALabel (count (DISTINCT ?gen
2  (group_concat(DISTINCT ?geneLabel; separator
3    ?gene wdt:P2293 ?diseaseGA .          # gene
4    ?diseaseGA wdt:P279* wd:Q3286546 .    # to a type of respiratory system disease
5
6    ?gene wdt:P31 wd:Q7187 ; wdt:P688 ?protein ;    # gene is subclass of "gene" and encodes protein
7          rdfs:label ?geneLabel .
8    FILTER (lang(?geneLabel) = "en")
9    ?protein p:P681 ?s .                            # protein's cell component statement
10     ?s ps:P681 ?cp .                              # get statement value
```

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."

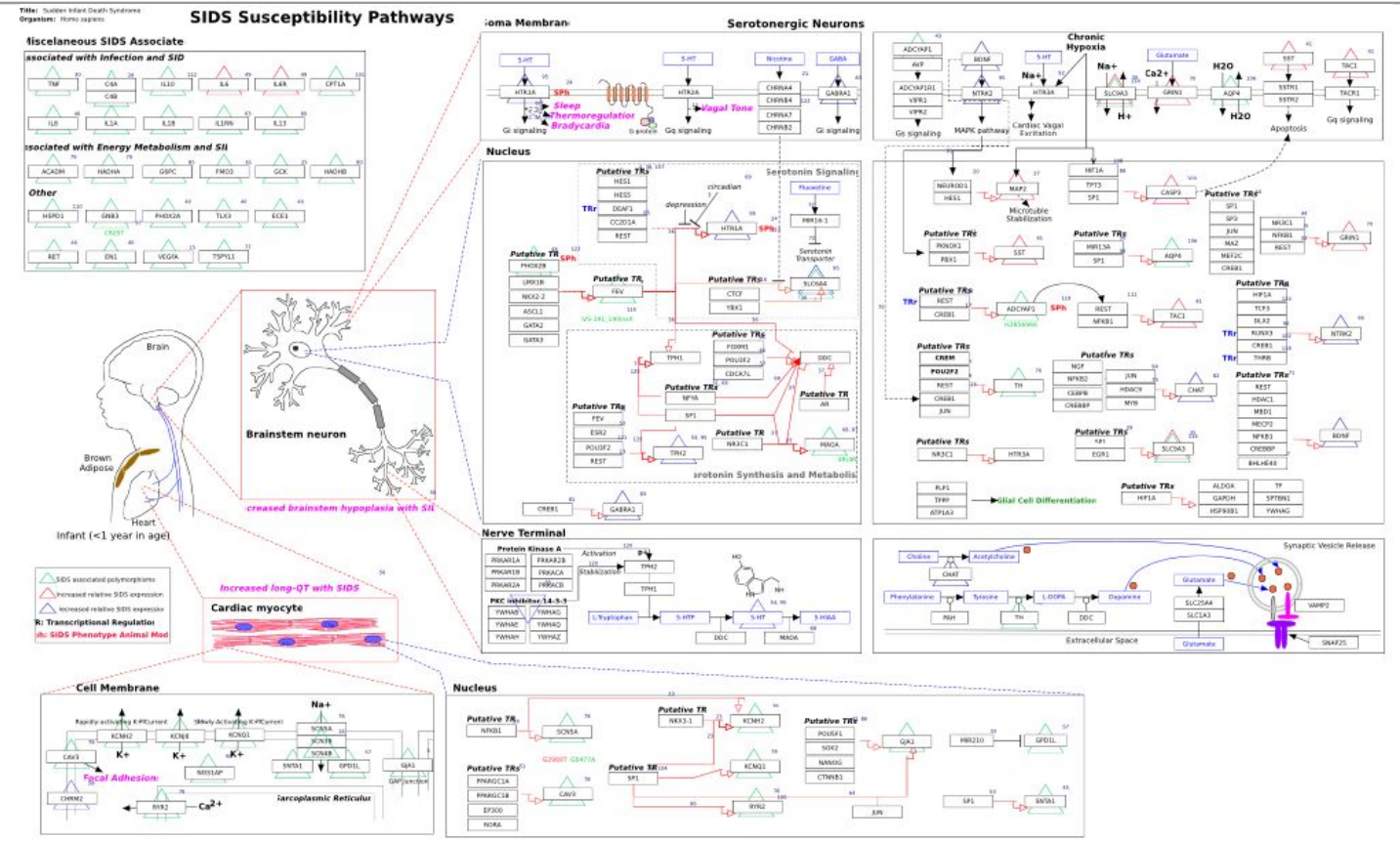
16 genes / 59 pathways

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

EMBL-EBI PIR SIB

Cellosaurus
@Cellosaurus

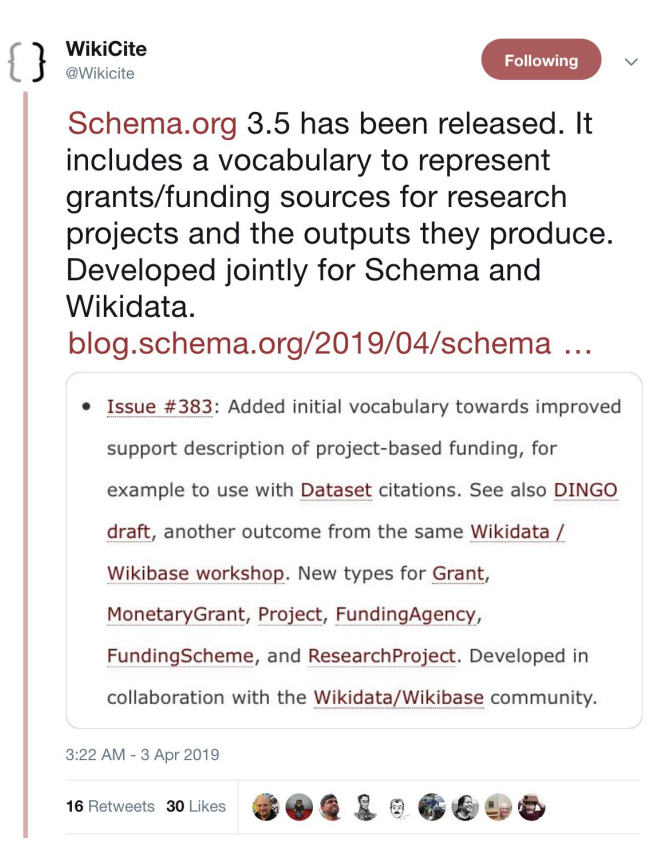
Following

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

Is Wikidata FAIR?

What makes a service FAIR?



Recognition

Rec. 7: Support semantic technologies

Semantic technologies are essential for interoperability and need to be developed, expanded and applied both within and across disciplines.

Wikidata as a cross-disciplinary FAIR data platform

Wikidata (https://www.wikidata.org) is a multilingual collaborative database collecting, reusing and providing structured Open data. The platform hosts information across all areas of knowledge and is tightly integrated with all Wikipedia sites. About 18,000 people contribute in a typical month. The human contributors are aided by hundreds of automated or semi-automated tools that perform similar tasks at scale, based on community-agreed standards.

An identifier-first architecture

Each entity in Wikidata (e.g. an 'item' or a 'lexeme') has a globally unique and persistent identifier that can be used by humans and machines to retrieve information on the topic. Entities can be described using an increasingly rich metadata vocabulary that consists of several thousand uniquely identifiable 'properties'. Some of these express relationships between Wikidata entities, while others can be used to link concepts with concrete values, e.g. the height of a mountain or the pseudonym of a writer.

In contrast to classical Subject Predicate Object triples, Wikidata's data model includes optional qualifiers to make statements more specific, as well as references to highlight the provenance of a specific piece of information. Every entity is linked to multiple different assertions.

The identifier-first architecture has many benefits. It enables Wikidata to support hundreds of languages and allows editors from all over the globe to review, refine, expand, correct or otherwise build on each other's contributions in a FAIR manner.

Wikidata is a FAIR data platform:

- It can be searched and queried in multiple ways, including via SPARQL, the query language of the Semantic Web
- Wikidata is accessible via open, free, and universally implementable protocols, with authentication and authorization where necessary
- Metadata provided by automated tools are usually associated with detailed provenance
- Except for specific circumstances, metadata about deleted data remains available.
- The data and metadata are published under CC0, which allows for reuse without restrictions
- The software for the site and for most of the user-generated tools is openly licensed, which allows an ecosystem of federated FAIR databases to grow around Wikidata.

By acting as an identifier hub, Wikidata helps other resources across and beyond the research landscape – e.g. including the cultural heritage sector – increase their FAIRness.

Image credit: CC-BY Elena Simperl and Alessandro Piscopo, slides 6 & 7 www.slideshare.net/elenasimperl/quality-and-collaboration-in-wikidata

Figure 11: Wikidata case study: a cross-disciplinary FAIR platform

Ticking the criteria

(F)indable

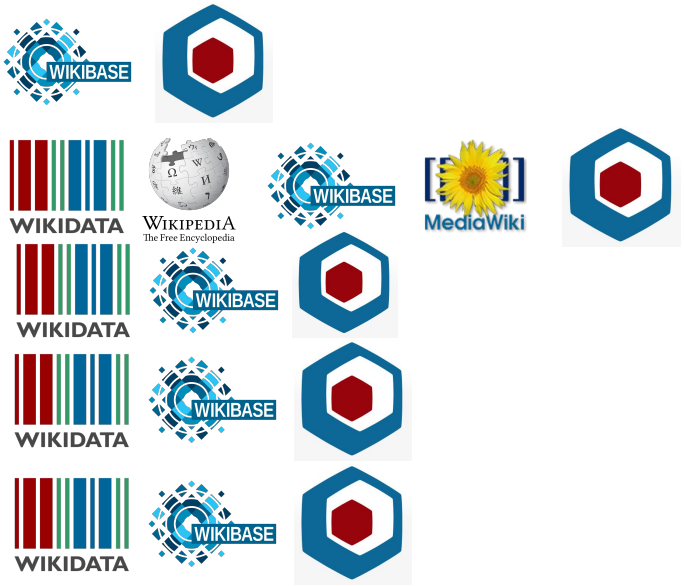
(A)ccesible

(I)nteroperable

(R)eusable

Where is Wikimedia ecosystem located in the Research Data Lifecycle

- Data creation
- Documentation
- Discovery
- Sharing
- Reuse



How do you interoperate/collaborate/engage with other services to help support FAIR data?

Updating Disease information from Disease Ontology in Wikidata

What is the one key change you think needs to happen for us to implement FAIR data and FAIR services?

Ruben Verborgh (UGent-imec) <Ruben.Verborgh@ugent.be>
to Eric, Andra

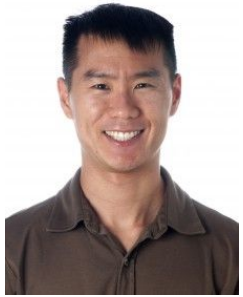
English > Dutch Translate message

I wish I knew such people, but I have trouble finding them myself :-)

Acknowledgements

Núria Queralt Rosinach

Benjamin Good



Andrew Su

Chunlei Wu

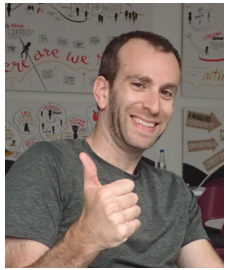
- Elvira Mitraka (Disease Ontology, U Baltimore)
- Gang Fu, Evan Bolton (NIH, PubChem)
- Wikimedia Foundation

Thousands of WikiDatans

Tim Putman

Lynn Schriml

Sebastian Burgstaller



Gregory Stupp

By Helpameout - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=20337311

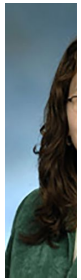
Funding

- National Institutes of General Medical Sciences (R01GM089820)
- NIH Common Fund programs for Big Data to Knowledge (U54GM114833)

Acknowledgements

Núria Queralt Rosi

Lynn Schriml



Funding

- National Inst
  Sciences (R
- NIH Commo
  Data to Kno

Sebastian Burgstaller

Center for Integrative Bioinformatics,
Max Perutz Laboratories,
Vienna Biocenter.

- Elvira Mitraka
  Disease Ontology, U
  Baltimore)
  Gang Fu, Evan Bolton
  NIH, PubChem)
  Wikimedia Foundation

ands of WikiDatans

pameout - Own work, CC BY-SA 3.0,
commons.wikimedia.org/w/index.php?c
0337311

Availability

Wikimedia ecosystem for FAIR data

- www.wikidata.org - Main entry
- query.wikidata.org - Wikidata Query Service
- www.wikibase.org - Infrastructure stack
- github.com/wmde/wikibase-docker - Wikibase docker
- wikidata-shex.wmflabs.org/wiki/Main_Page - Schema

Tools

- Chlambase: chlambase.org
- Scholia: tools.wmflabs.org/scholia/
- Inventaire: inventaire.io
- Open Refine: www.wikidata.org/wiki/Wikidata:Tools/OpenRefine
- …… (many more)

Gene Wiki: github.com/SuLab/GeneWikiCentral

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