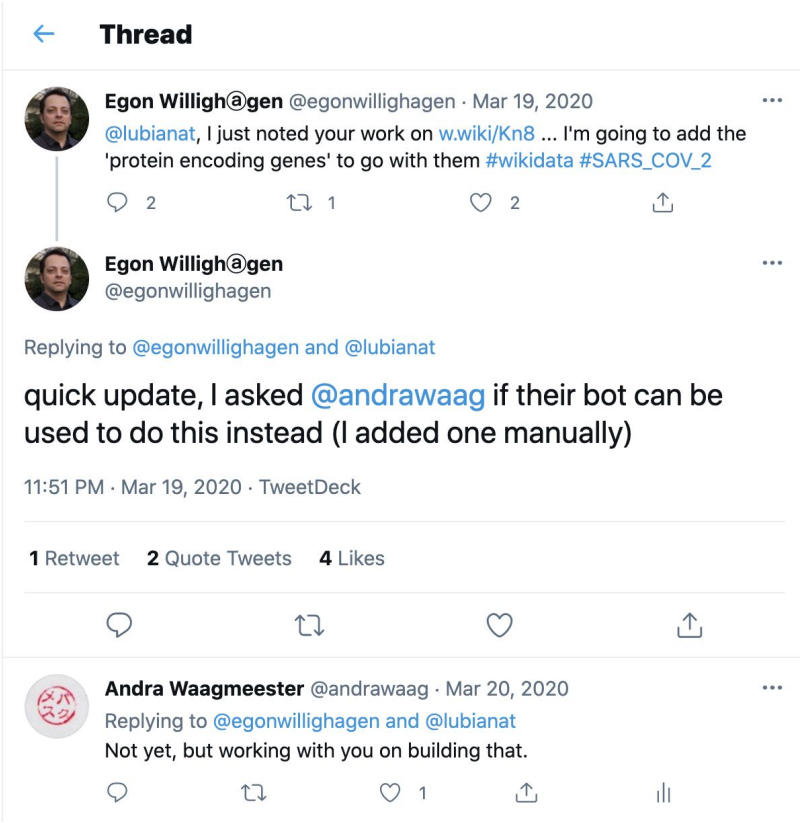
Swift Integration of Knowledge in Wikidata through Linked Data (RDF) and Schemas (ShEx)

Andra Waagmeester

1) Micelio, Antwerp, Belgium | Email: andra@micelio.be, Twitter: @andrawaag

Protocol to add Covid-19 to Wikidata

What we did...



BMC Part of Springer Nature

Search

BMC Biology

Home About Articles Submission Guidelines

Methodology article | Open Access | Published: 22 January 2021

A protocol for adding knowledge to Wikidata: aligning resources on human coronaviruses

Andra Waagmeester, Egon L. Willighagen, Andrew I. Su, Martina Kutmon, Jose Emilio Labra Gayo, Daniel Fernández-Álvarez, Quentin Groom, Peter J. Schaap, Lisa M. Verhagen & Jasper J. Koehorst

BMC Biology **19**, Article number: 12 (2021) | Cite this article

WIKIPEDIA
The Free Encyclopedia

Infrastructure	Resource	Content	Where?
MediaWiki	Wikipedia The Free Encyclopedia	Text	https://<lang>.wikipedia.org https://releases.wikimedia.org/mediawiki/
MediaWiki	Wikimedia Commons	Media	https://commons.wikimedia.org
MediaWiki + Wikibase	Wikidata	Data	https://www.wikidata.org https://www.wikibase.org

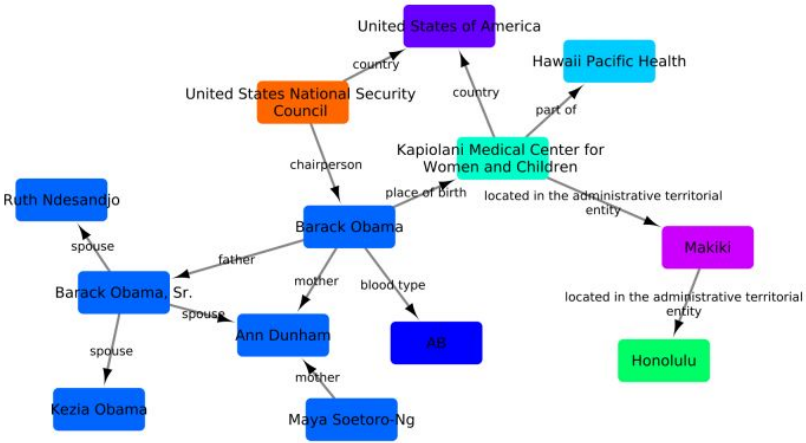
Wikidata is to data as Wikipedia is to text

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)

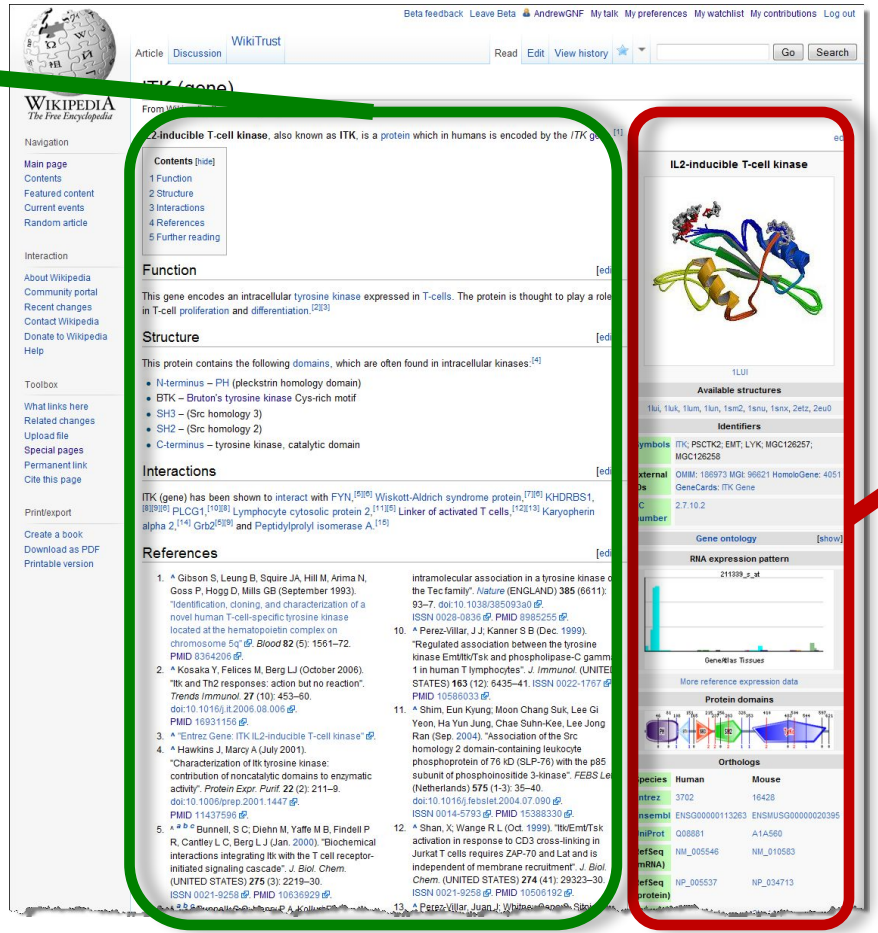
- Stable! Long term support not dictated by funding cycles
- Actively developed
- Already has large number of active users, editors contributors!

A giant graph of knowledge!



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

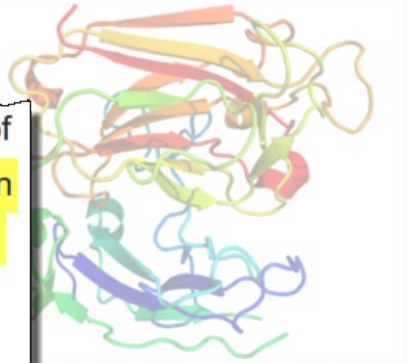
Reelin is a large secreted extracellular matrix glycoprotein that helps regulate processes of neuronal migration and positioning in the developing brain by controlling cell–cell interactions. Besides this important role in early development, reelin continues to work in the adult brain. It modulates synaptic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

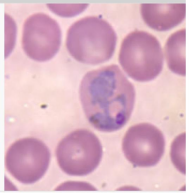
Identifiers	
Symbols	RELN ; LIS2; PRO1598; RL
External	OMIM: 600514 MGI: 103022

Wikipedia: Maintained independently by >300 language communities

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

	Papiamento
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))

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

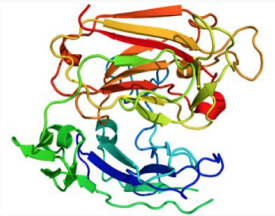
Reelin (RELN)[5] is a large secreted extracellular matrix glycoprotein that helps regulate processes of neuronal migration and positioning in the developing brain by controlling cell-cell interactions. Besides this important role in early development, reelin continues to work in the adult brain. It modulates 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]

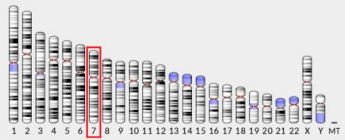
RELN



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

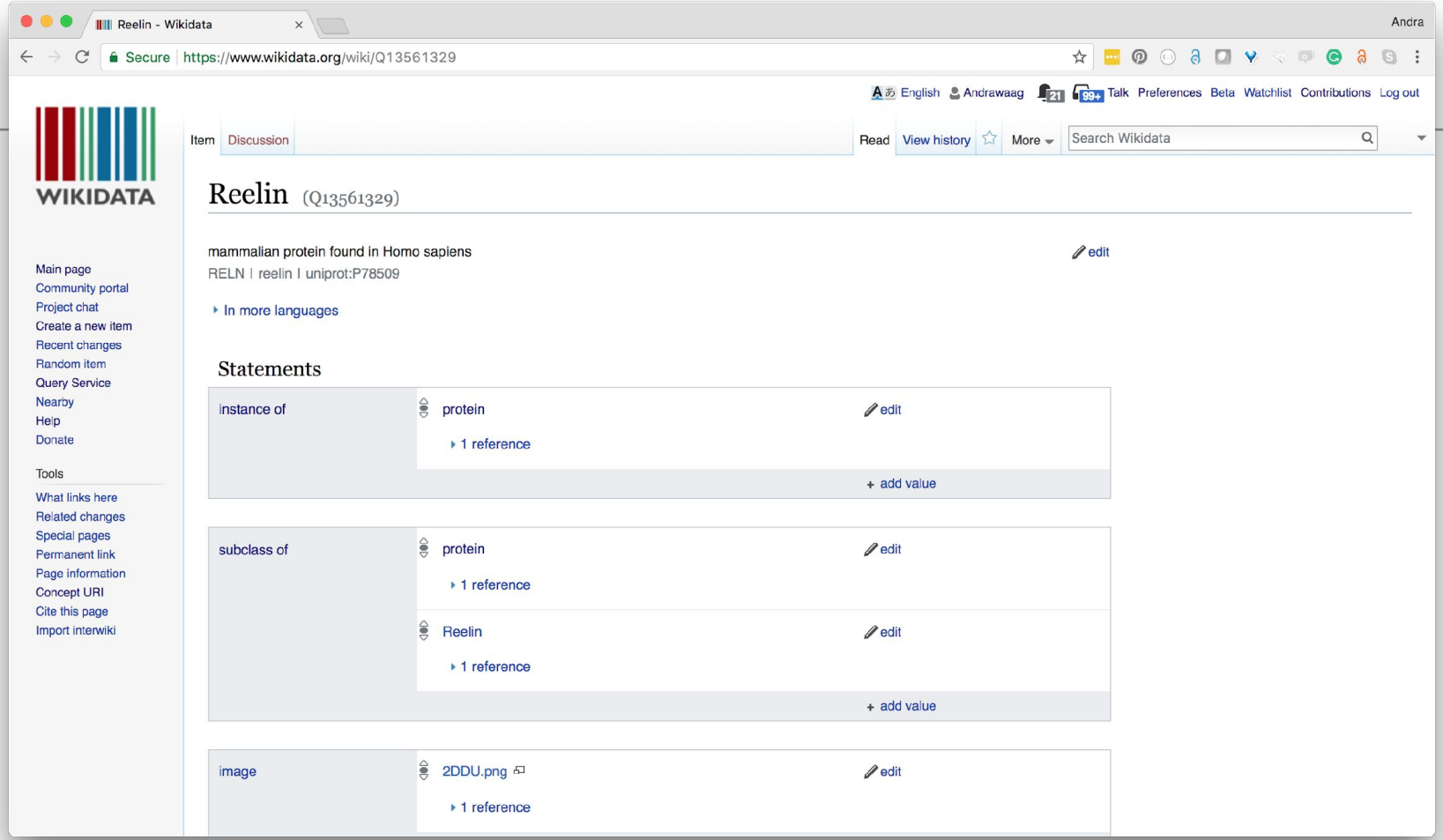
Identifiers	
Aliases	RELN, LIS2, PRO1598, RL, reelin, ETL7
External IDs	OMIM: 600514 MGI: 103022 HomoloGene: 3699 GeneCards: RELN

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

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

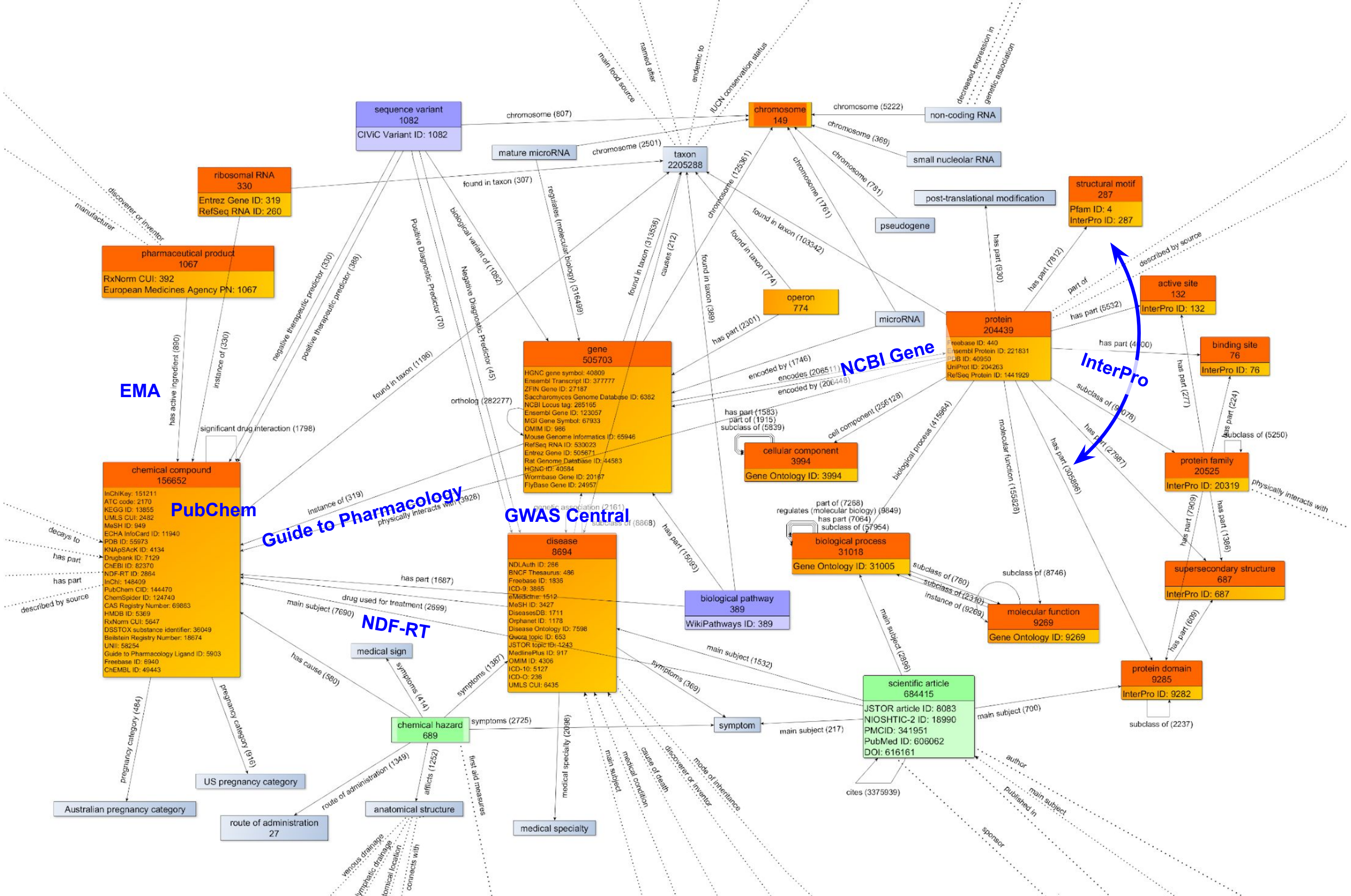
Diff selection: Mark the radio boxes of the revisions to compare and hit enter or the button at the bottom.

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)


sequence variant
1082
CIViC Variant ID: 1082
chromosome (807)
chromosome
149
chromosome (5222)
non-coding RNA
chromosome (369)
small nucleolar RNA
chromosome (781)
pseudogene
mature microRNA
chromosome (2501)
taxon
2205288
ribosomal RNA
330
Entrez Gene ID: 319
RefSeq RNA ID: 260
found in taxon (307)
pharmaceutical product
1067
RxNorm CUI: 392
European Medicines Agency PN: 1067
discoverer or inventor
manufacturer
main food source
named after
endemic to
IUCN conservation status
decreased expression in
genetic association
post-translational modification
structural motif
287
Pfam ID: 4
InterPro ID: 287
described by source
part of
active site
132
InterPro ID: 132
operon
774
microRNA
protein
204439
Freebase ID: 440
Ensembl Protein ID: 221831
PDB ID: 40950
UniProt ID: 204263
RefSeq Protein ID: 1441929
binding site
76
InterPro ID: 76
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: 530023
Entrez Gene ID: 505671
Rat Genome Database ID: 44583
HGNC ID: 40584
Wormbase Gene ID: 20167
FlyBase Gene ID: 24957
EMA
NCBI Gene
InterPro
PubChem
Guide to Pharmacology
GWAS Central
NDF-RT
has active ingredient (890)
instance of (330)
negative therapeutic predictor (330)
positive therapeutic predictor (389)
Positive Diagnostic Predictor (70)
Negative Diagnostic Predictor (45)
biological variant of (1082)
regulates (molecular biology) (316499)
found in taxon (313536)
causes (212)
chromosome (125361)
found in taxon (389)
found in taxon (774)
found in taxon (103342)
chromosome (1761)
has part (2301)
has part (930)
has part (7812)
has part (5532)
has part (4100)
encoded by (1746)
encodes (206511)
encoded by (206448)
found in taxon (1196)
ortholog (282277)
significant drug interaction (1798)
has part (1583)
part of (1915)
subclass of (5839)
cell component (256128)
cellular component
3994
Gene Ontology ID: 3994
biological process (415964)
molecular function (158828)
has part (305896)
has part (27987)
subclass of (91078)
has part (277)
has part (224)
subclass of (5250)
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: 55973
KNApSAcK ID: 4134
Drugbank ID: 7129
ChEBI ID: 82370
NDF-RT ID: 2864
InChI: 148409
PubChem CID: 144470
ChemSpider ID: 124740
CAS Registry Number: 69863
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
decays to
has part
has part
described by source
instance of (319)
physically interacts with (3928)
genetic association (2161)
subclass of (8868)
has part (15093)
part of (7268)
regulates (molecular biology) (9849)
has part (7064)
subclass of (57954)
biological process
31018
Gene Ontology ID: 31005
has part (7909)
has part (1386)
supersecondary structure
687
InterPro ID: 687
disease
8694
NDLAuth ID: 286
BNCF Thesaurus: 486
Freebase ID: 1836
ICD-9: 3865
eMedicine: 1512
MeSH ID: 3427
DiseasesDB: 1711
Disease Ontology ID: 7598
Orphanet ID: 1178
Quora topic ID: 653
JSTOR topic ID: 1243
MedlinePlus ID: 917
OMIM ID: 4306
ICD-10: 5127
ICD-O: 236
UMLS CUI: 6435
has part (1687)
drug used for treatment (2699)
main subject (7690)
biological pathway
389
WikiPathways ID: 389
subclass of (780)
subclass of (2310)
instance of (9269)
subclass of (8746)
molecular function
9269
Gene Ontology ID: 9269
has part (609)
medical sign
has cause (580)
symptoms (414)
symptoms (1387)
main subject (1532)
symptoms (369)
main subject (2896)
scientific article
684415
JSTOR article ID: 8083
NIOSHTIC-2 ID: 18990
PMCID: 341951
PubMed ID: 606062
DOI: 616161
main subject (700)
protein domain
9285
InterPro ID: 9282
subclass of (2237)
pregnancy category (484)
pregnancy category (916)
chemical hazard
689
symptoms (2725)
symptom
main subject (217)
US pregnancy category
Australian pregnancy category
route of administration (1349)
afflicts (1252)
first aid measures
medical specialty (2098)
main subject
medical condition
cause of death
discoverer or inventor
mode of inheritance
cites (3375939)
author
main subject
published in
sponsor
route of administration
27
anatomical structure
medical specialty
venous drainage
lymphatic drainage
anatomical location
connects with

Getting data in..

License: CC0

- All structured data from the main, Property, Lexeme, and EntitySchema namespaces is available under the Creative Commons CC0 License

License	Add to Wikidata	Add to Commons	Add to Wikipedia
CC0	+	+	+
CC-BY	+	+	+
CC-BY-NC	-	-	-
CC-BY-SA	-	+	+
CC-BY-ND	-	-	-
CC-BY-NC-SA	-	-	-
CC-BY-NC-ND	-	-	-

Eligibility for inclusion

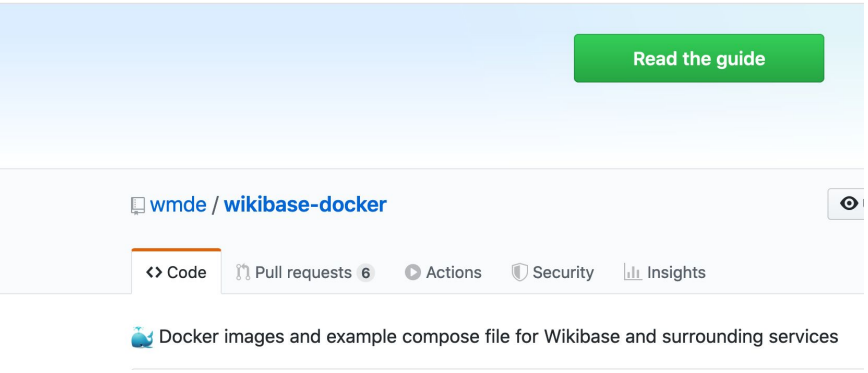
1. It contains at least one valid sitelink to a page on Wikipedia, Wikivoyage, Wikisource, Wikiquote, Wikinews, Wikibooks, Wikidata, Wikispecies, Wikiversity, or Wikimedia Commons.
2. It refers to an instance of a clearly identifiable conceptual or material entity. The entity must be notable, in the sense that it can be described using serious and publicly available references.
3. It fulfills some structural need, for example: it is needed to make statements made in other items more useful.

https://www.wikidata.org/wiki/Wikidata:Notability

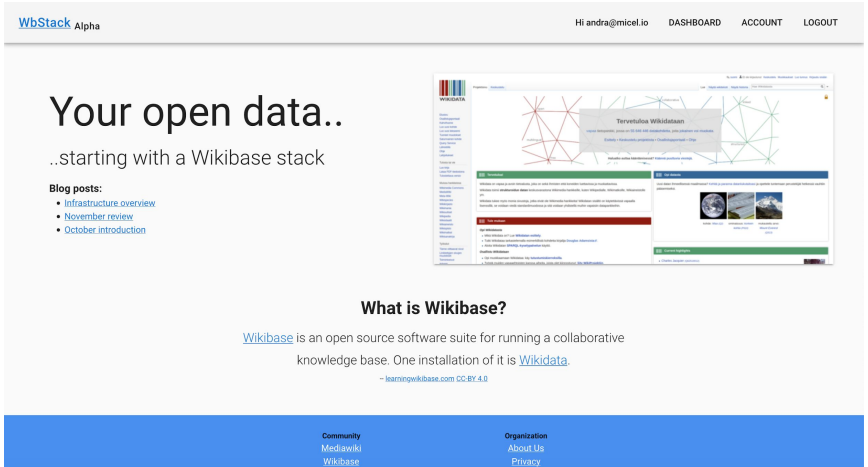
Infrastructure	Resource	Content	Where?
MediaWiki	Wikipedia The Free Encyclopedia	Text	https://<lang>.wikipedia.org https://releases.wikimedia.org/mediawiki/
MediaWiki	Wikimedia Commons	Media	https://commons.wikimedia.org
MediaWiki + Wikibase	Wikidata	Data	https://www.wikidata.org https://www.wikibase.org

Wikibase and WBStack

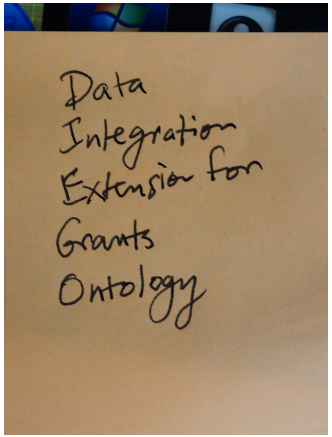
github.com/wmde/wikibase-docker

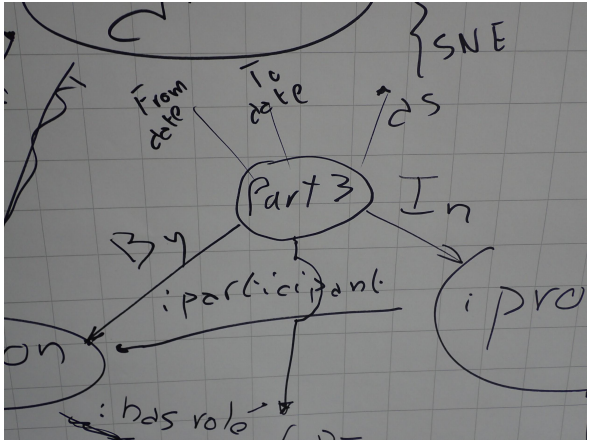


wbstack.com



Community engagement and model discussion

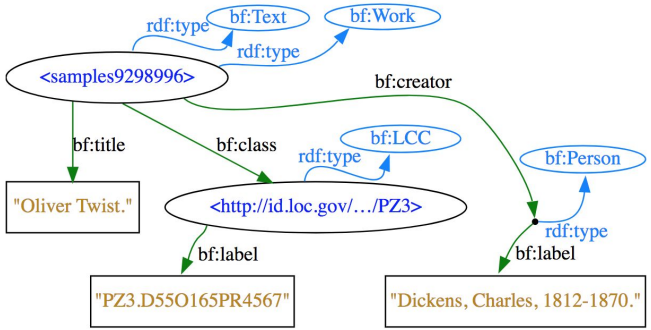




Formally capture and describe model and community consensus

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" .
```

https://www.wikidata.org/wiki/EntitySchema:E37

Donate

Print/export

Download as PDF

Tools

What links here

Related changes

Special pages

pt	gene humano			edit

```
# E108: genome_assembly
IMPORT <https://www.wikidata.org/wiki/Special:EntitySchemaText/E108>
PREFIX E108: <https://www.wikidata.org/wiki/Special:EntitySchemaText/E108#>

# E109: human chromosome
IMPORT <https://www.wikidata.org/wiki/Special:EntitySchemaText/E109>
```

check entities against this Schema | edit |

Enter an entity to check e.g.Q42 Check

```
p:P31 @<#P31_instance_of_gene> ;
```

```
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX p: <http://www.wikidata.org/prop/>
```

```
<#P31_instance_of_gene> {
   ps:P31 @<#gene_types> ;      # Instance of [P31] gene types
   prov:wasDerivedFrom @<#ncbi-gene-reference> OR @<#ensembl-gene-reference>
}
```

```
start = @<#wikidata-human-gene>
```

```
(
        p:P644 @<#P644_genomic_start> ; # Its genomic start location
        p:P645 @<#P645_genomic_end> ; # Its genomic end location
)* ; # Zero or more start and end locations.
```

```
# Value statements contain either actual values, or pointers to other Wikidata items.
Identifier statements capture
# external identifiers, erroneous statements are those that are errors.
```

ShEx2 — Simple Online Validator

```
# Shape Expression for Human genes in Wikidata
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX p: <http://www.wikidata.org/prop/>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX pq: <http://www.wikidata.org/prop/qualifier/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX prv: <http://www.wikidata.org/prop/reference/value/>
PREFIX pr:  <http://www.wikidata.org/prop/reference/>
PREFIX ps: <http://www.wikidata.org/prop/statement/>

BASE <http://www.wikidata.org/entity/>

start = @<#wikidata-human-gene>

# Query with results
# SELECT * WHERE {?item wdt:P31 wd:Q7187 ; wdt:P703 wd:Q15978631 .} LIMIT 10

# Indicates which shape to use to start iterating over the graph if none is
provided.

# wikidata-human gene is the main shape for a human gene data model in Wikidata.
Each line between the brackets
# represents the structure than can be enforced to validate human gene annotations
in Wikidata
```

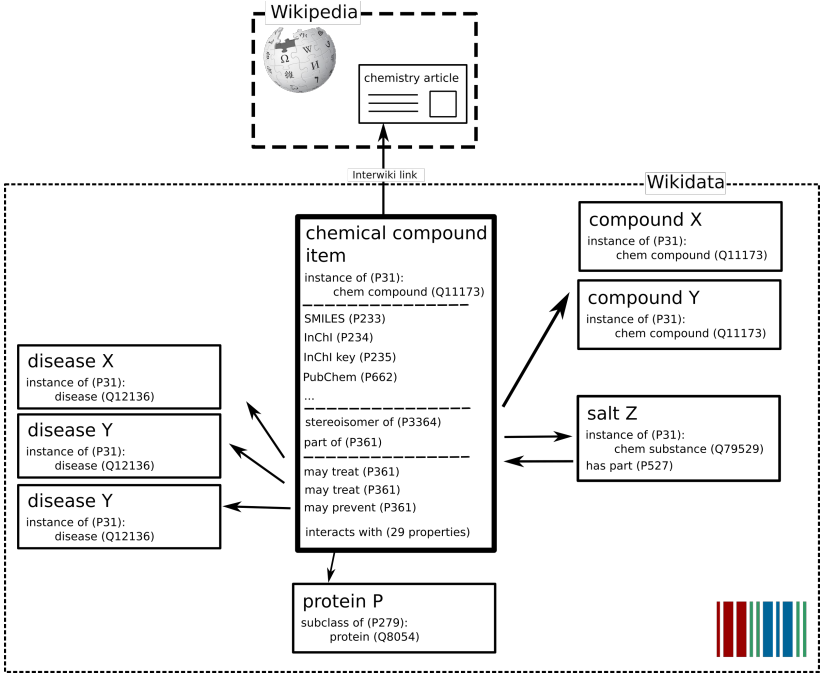
abort (ctl-enter)

Query | Entities to check

<http://www.wikidata.org/entity/Q414043>	@	START	- ✓
<http://www.wikidata.org/entity/Q415594>	@	START	- ✓
<http://www.wikidata.org/entity/Q416426>	@	START	- ✓
<http://www.wikidata.org/entity/Q417169>	@	START	- ✓
<http://www.wikidata.org/entity/Q417743>	@	START	- ✓
<http://www.wikidata.org/entity/Q418553>	@	START	- ✓

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

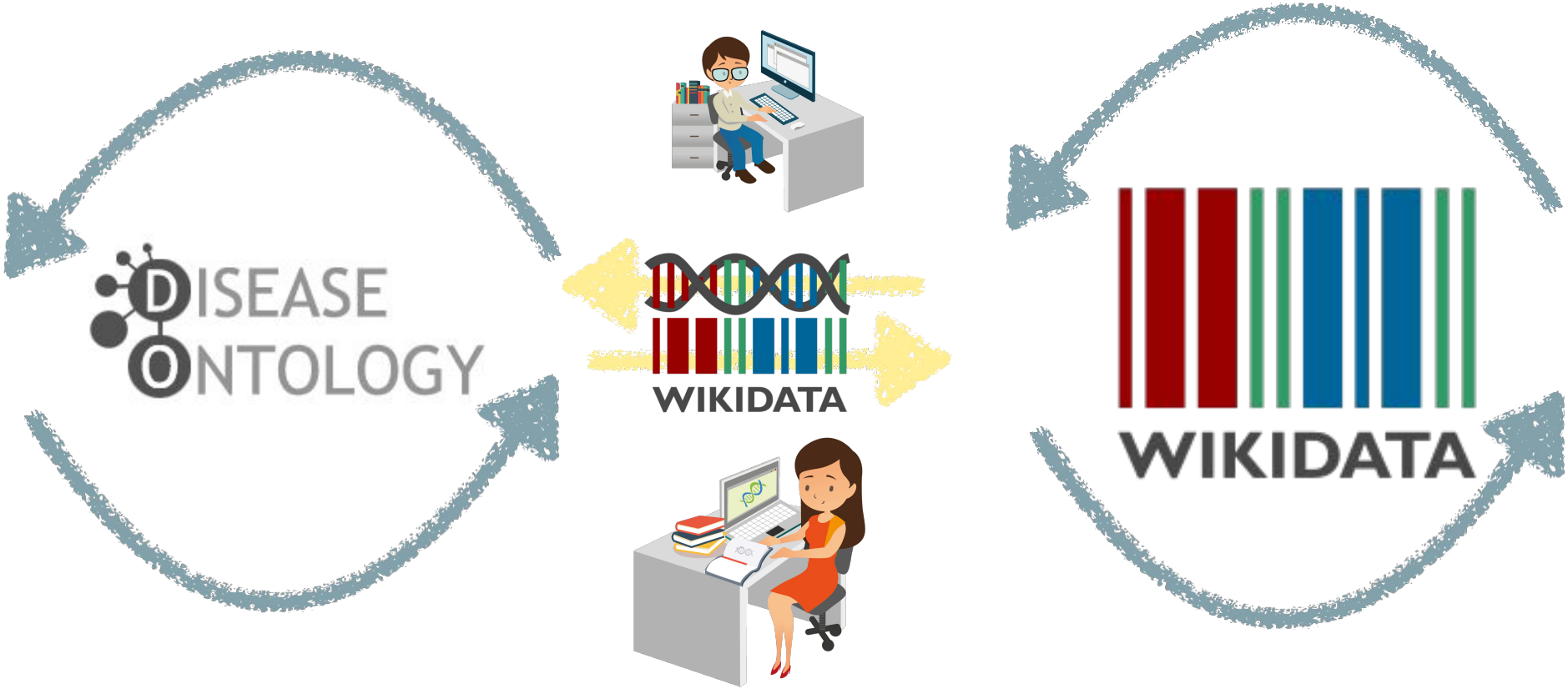
Build Queue

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

Feedback loop



Wikidata reconciliation

- Schema reconciliation
  - Schema extraction
    - Using similar items as templates
    - Wikiprojects
    - sheXer
  - Property proposals
- Data reconciliation
  - On labels
  - On shared identifiers
  - IRI mapping

```
start = @<#wikidata-virus-gene>

<#wikidata-virus-gene> {
  p:P31                                  @<#P31_instance_of_gene> ;
  p:P279                                 @<#P279_subclass_of_gene>? ;
  p:P688                                 @<#P688_encodes>? ; # Zero or one
geneproducts.
  p:P703                                 @<#P703_found_in_taxon_virus> ; # In
which taxonomy and where in that taxonomy this gene is found

  # Identifiers
  p:P351                                 @E266:P351_ncbi_gene_id ; # Exactly
one ncbi gene identifier
  p:P594                                 @E266:P594_ensembl_gene_id* ; # Zero
or more Ensembl gene identifier

  p:P2393                                @E266:P2393_ncbi_locus_tag?; # NCBI
Locus tag
}

## Statement details
<#P31_instance_of_gene> {
  ps:P31                                 [wd:Q7187] ;    # Instance of [P31]
gene
  prov:wasDerivedFrom   @E265:ncbi-gene-reference OR @E265:ensembl-gene-
reference ;
}

<#P279_subclass_of_gene> {
  ps:P279                                @<#gene_types> ; # Subclass of [P279]
gene types <gene_types>
  prov:wasDerivedFrom   @E265:ncbi-gene-reference OR @E265:ensembl-gene-
reference ;
}

<#P644_genomic_start> {
  ps:P644                                xsd:string ;  # genomic start [P644]
value
  pq:P659                        @E108:sequence_assembly+ ;  # Qualifier
indicating the applicable genomic assembly versions.
  prov:wasDerivedFrom   @E265:ensembl-gene-reference ;
}
```

Entity schemas and Shape Expressions at Wikidata

RDF and knowledge graphs, the good parts

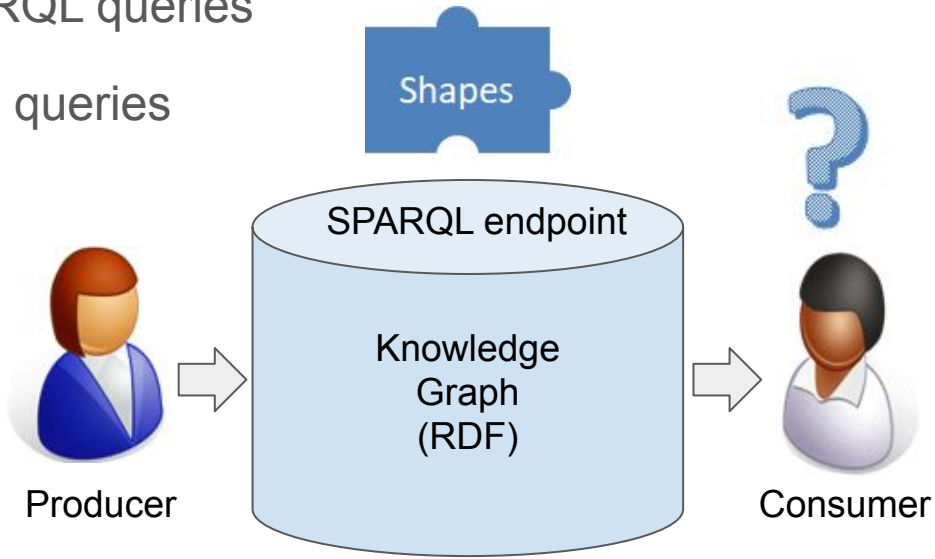
- Data integration
  - Merging RDF graphs automatically
  - RDF as a basis for knowledge representation
- Flexibility
  - Data that can be adapted to multiple environments
  - Reusable data by default
- Tools
  - Data stores and SPARQL endpoints
  - Multiple serializations: Turtle, JSON-LD, RDF/XML,...
  - Embeddable in HTML (Microdata/RDFa)

RDF and knowledge graphs, other parts...

Consuming & producing data from RDF

SPARQL endpoints are usually not well documented

- Typical documentation = set of SPARQL queries
- Difficult to know where to start doing queries



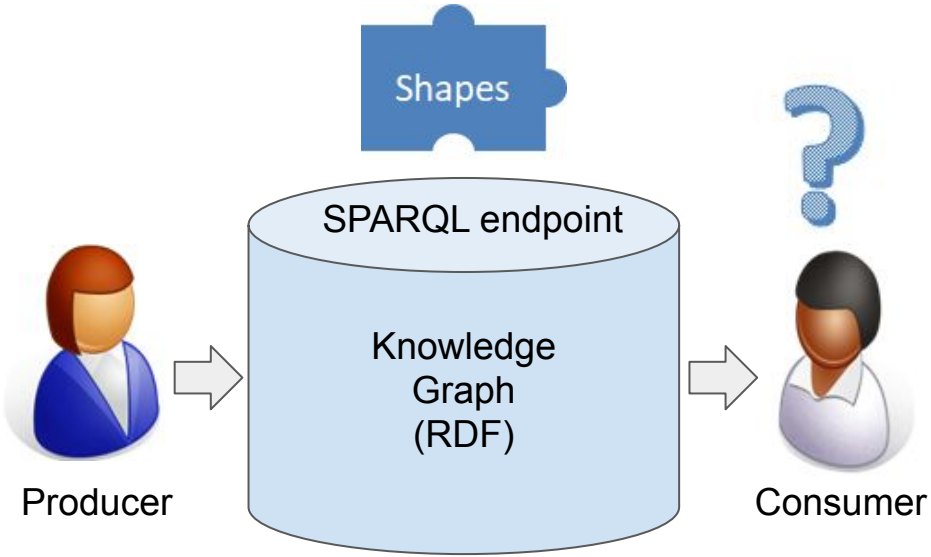
Why Shape Expressions?

For producers

- Understand the contents they will produce
- Ensure contents have the expected structure
- Advertise and document the structure
- Generate user interfaces

For consumers

- Understand content's structure
- Validate before processing
- Query generation & optimization



Shape Expressions

Language to describe and validate RDF data

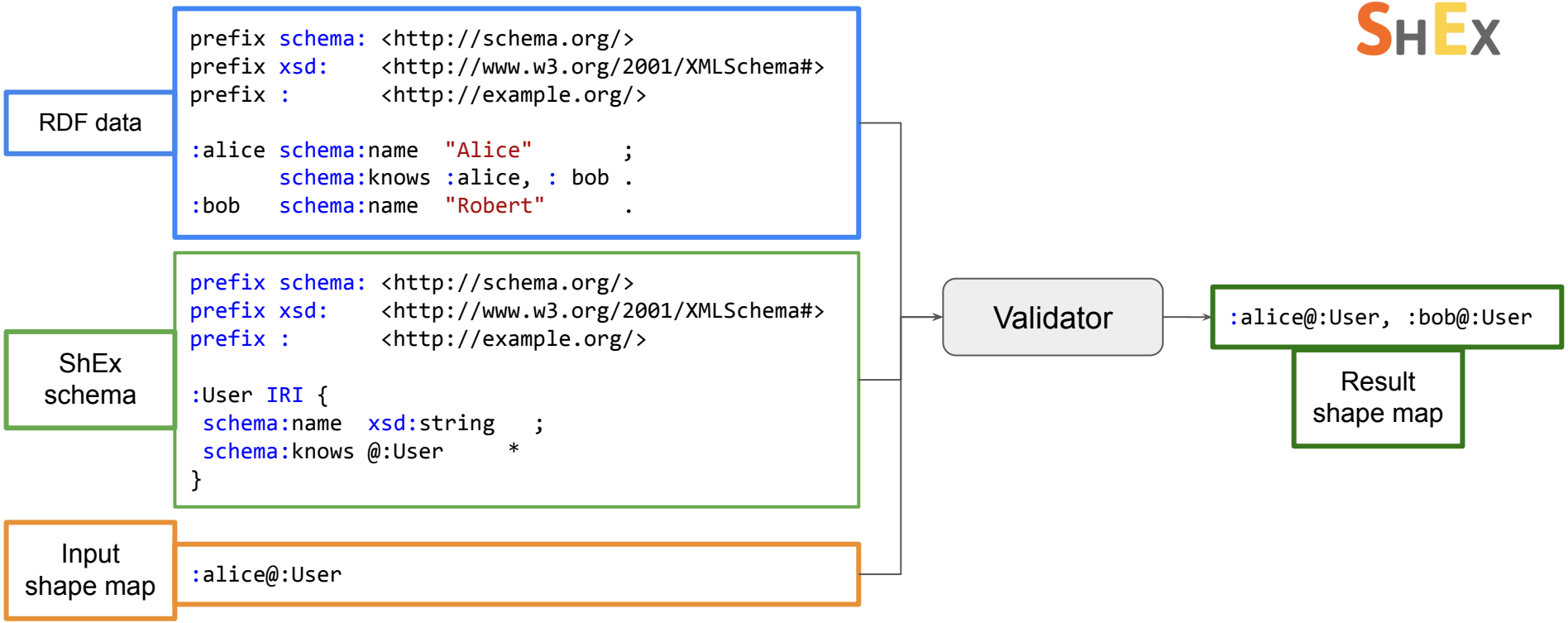
- Human readable
  - Intended audience: domain experts
  - Syntax inspired by Turtle and SPARQL
- Machine processable
  - Formal semantics
  - Several syntaxes (Compact, RDF, JSON-LD)
  - Open source implementations:
    - Javascript, Scala, Java, Python, …
  - Online demos: RDFShape, ShEx-simple

ShEx schema

```
prefix schema: <http://schema.org/>
prefix xsd:    <http://www.w3.org/2001/XMLSchema#>
prefix :       <http://example.org/>

:User IRI {
  schema:name  xsd:string   ;
  schema:knows @:User     *
}
```

Shape Expressions example



[Try it with RDFShape]

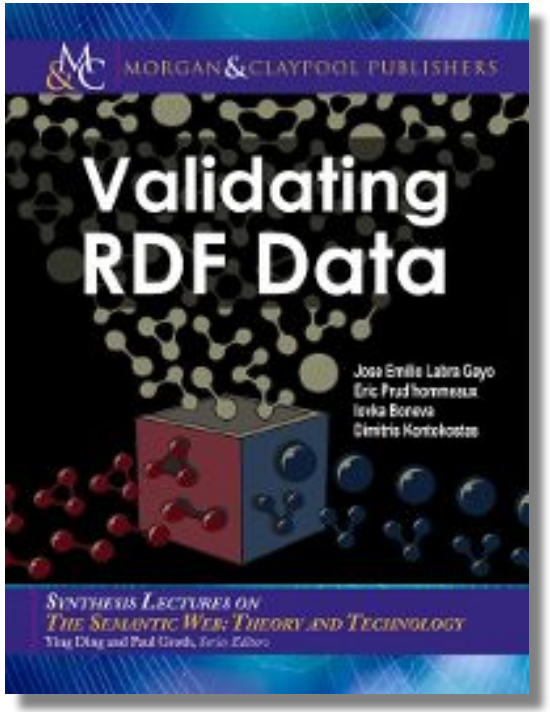
More info about Shape Expressions

Validating RDF Data book

http://book.validatingrdf.com/

Shapes applications and tools tutorial at ISWC'20

http://www.validatingrdf.com/tutorial/iswc2020/



Shape Expressions at Wikidata

Introduced in May, 2019 as a new namespace: Entity Schemas

- Enn = Entity schemas, e.g. Author: E42

Other namespaces:

- Qnn = Entities, e.g. Douglas Adams: Q42
- Pnn = Properties, e.g. country of citizenship: P27
- Lnn = Lexemes (lexicographical data), e.g.: ”answer” L42
- Snn = Senses
- ..

Some use cases and tools

Describing and validating entities with entity schemas

Authoring entity schemas

Extracting schemas from existing data: sheXer

Generating User interfaces from entity schemas: shapeForms

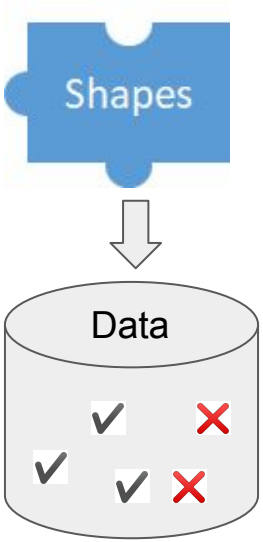
Entity schemas ecosystem

Describing and validating entities with entity schemas

Describe expected shape of entities

Check if entities conform to that shape

Filter entities according to conformance



Example: Author (E42)

shex-simple.toolforge.org/wikidata/packages/shex-webapp/doc/shex-simple.html?data=Endpoint:%20http...

ShEx2 — Simple Online Validator

```
PREFIX p: <http://www.wikidata.org/prop/>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX ps: <http://www.wikidata.org/prop/statement/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX psn: <http://www.wikidata.org/prop/statement/value-normalized/>
prefix ui:   <http://www.w3.org/ns/ui#>

# This schema is used as an example for some Entity schemas tutorials
# Example queries:
# 5 authors:
#  SELECT * WHERE { ?person wdt:P106 wd:Q36180 } LIMIT 5
# One specific author (Q42):
#  SELECT DISTINCT ?author WHERE {   VALUES ?author {  wd:Q42  }  }

start = @<AuthorShape>

<AuthorShape> EXTRA wdt:P31 wdt:P106 {

      wdt:P31 [ wd:Q5 ] // ui:label "instance of"@en ;

      wdt:P735 @<GivenName> * // ui:label "Given name"@en ;

      wdt:P21 IRI ? // ui:label "sex or gender"@en ;
```

validate (ctl-enter)

Query | Entities to check

<http://www.wikidata.org/entity/Q42>	START	✓
<http://www.wikidata.org/entity/Q272>	START	✓
<http://www.wikidata.org/entity/Q360>	START	✗
<http://www.wikidata.org/entity/Q377>	START	✓
<http://www.wikidata.org/entity/Q392>	START	✗

✓wd:Q42@START

✓wd:Q272@START

✗wd:Q360@!START

```
  validating http://www.wikidata.org/entity/Q360 as //www.wikidata.org/wiki/Special:EntitySchemaText/AuthorShape:
    validating "1971-07-03T00:00:00Z"^^http://www.w3.org/2001/XMLSchema#dateTime: exceeds cardinality
  OR
  validating "Julian Assange": exceeds cardinality
  OR
```

Authoring and visualizing entity schemas

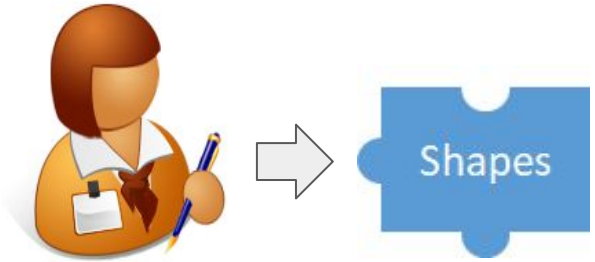
YaSHE: Editor with syntax highlighting, error detection, auto-complete, etc.

ShEx-Author: Visual editor

RDFShape: RDF playground

Conversion, Querying, visualization, Validation…)

Can be used to visualize Shape Expressions



Wikishape

Similar to RDFShape, but specialized for Wikidata/Wikibase

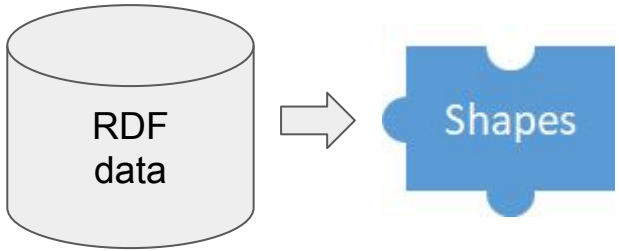
Extracting schemas from existing data

sheXer extracts ShEx schemas from RDF data

Identify common structure of a given set of items

Items can be selected by Class, SPARQL query, shapeMap, etc.

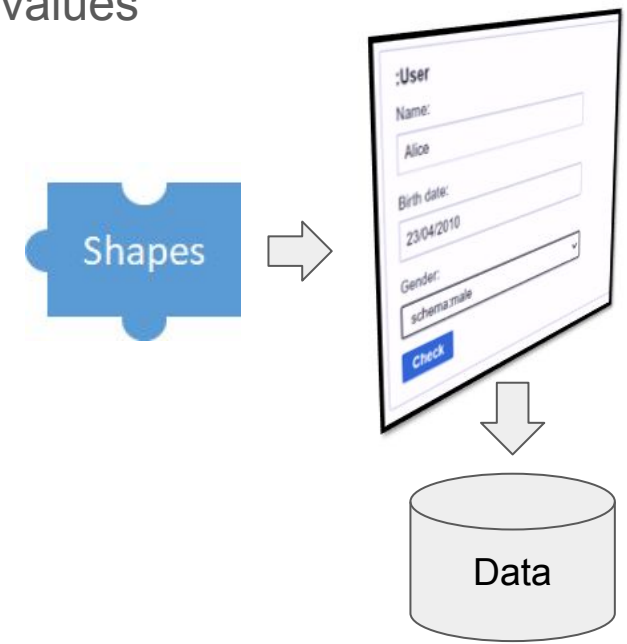
Integrated in RDFShape and WikiShape



Creating UIs from entity schemas

Generate Forms from entity schemas

- Suggest/check properties and values
- Prototypes
  - CRADLE
  - ShapeForms



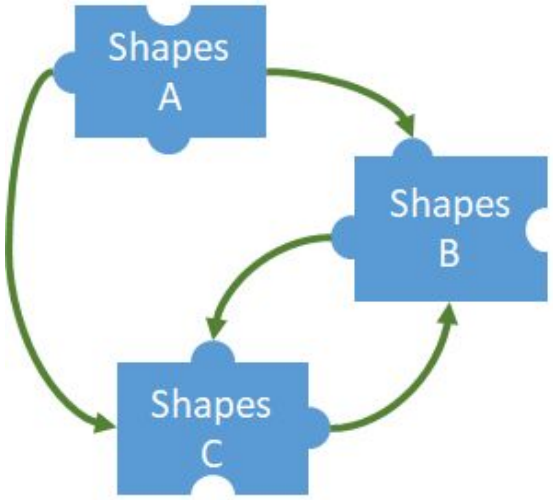
Entity schemas ecosystem

Increasing adoption of entity schemas at Wikidata

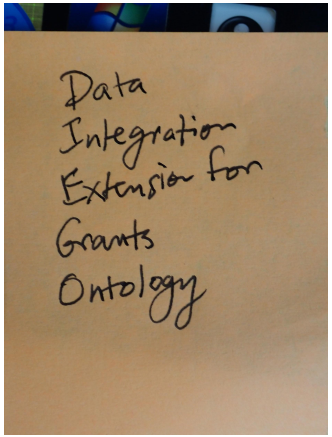
- Schemas project: https://www.wikidata.org/wiki/Wikidata:Schemas
- Directory of entity schemas

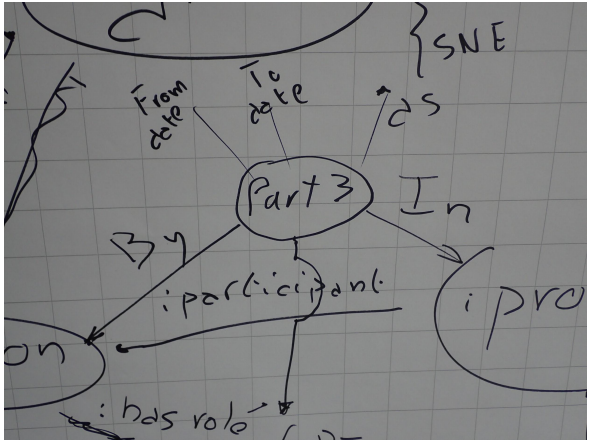
New challenges

- Collaborative work and different views
- Trade-off: quality of data vs freedom
- Domain experts from different disciplines



Community engagement and model discussion

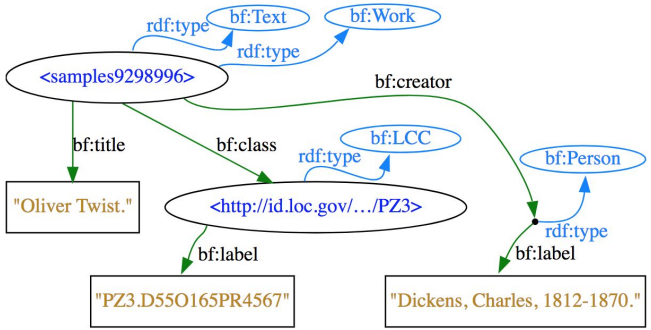




Formally capture and describe model and community consensus

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



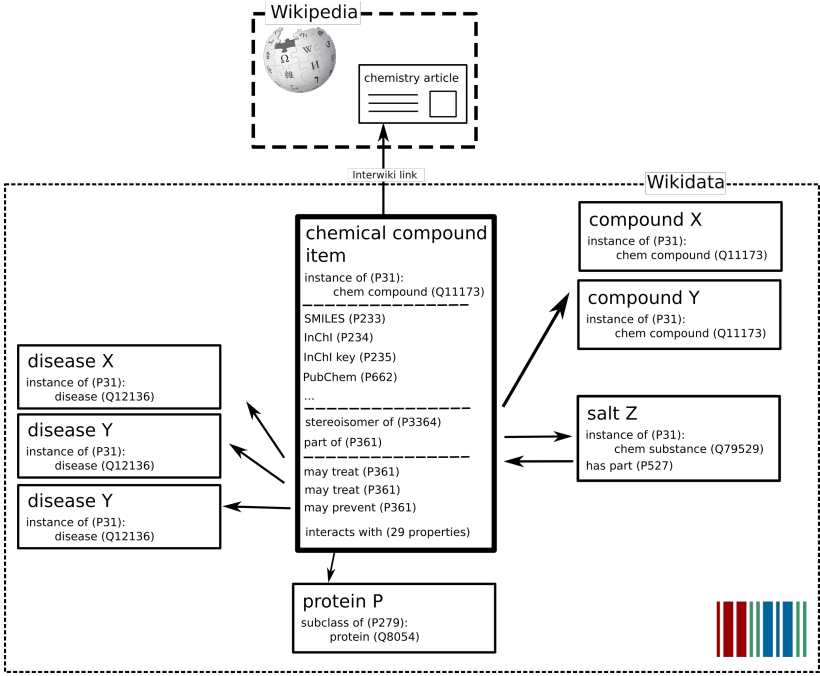
Data (Turtle)

```
<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

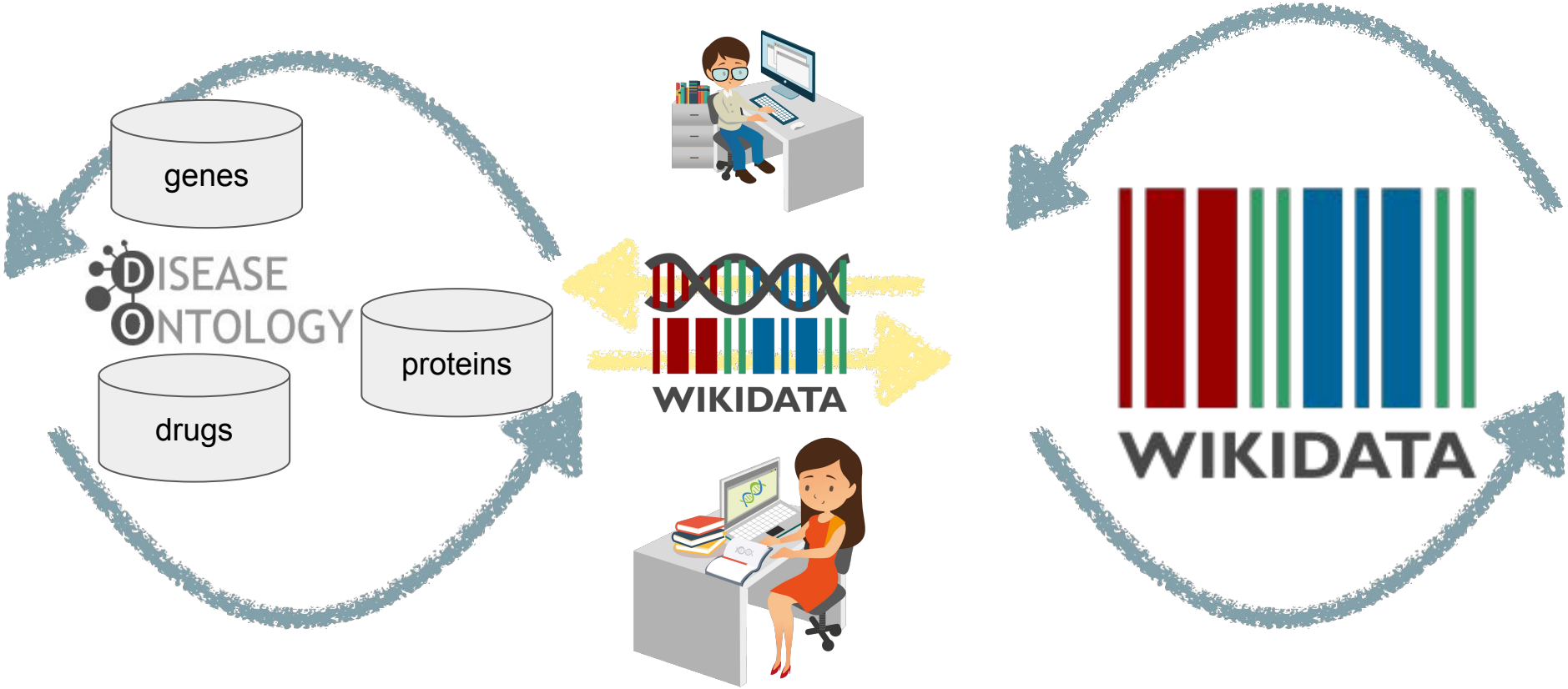
Build Queue

Running Bots

All | **Running** | +

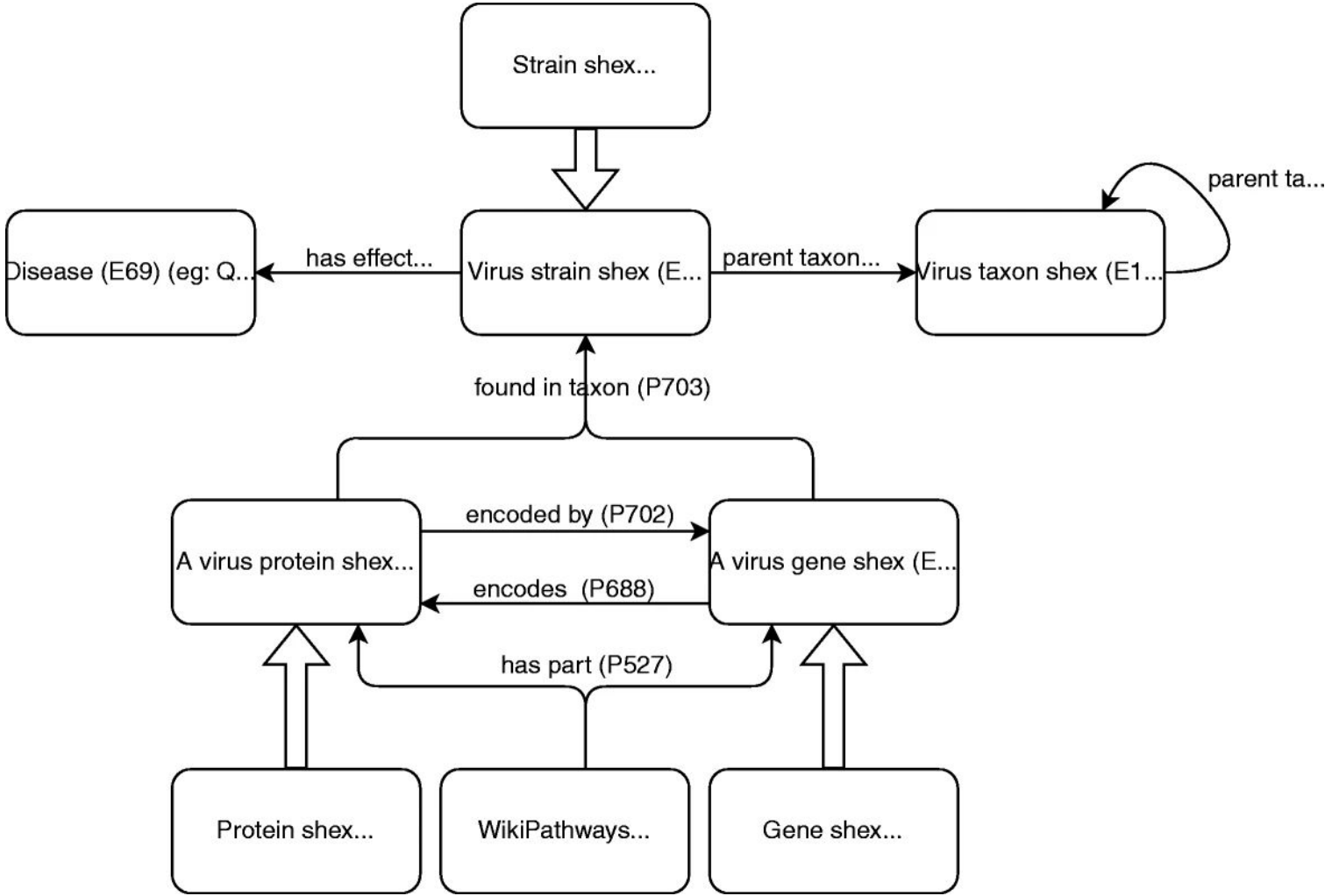
S	Name	Last Success ↑	Last Failure
	ProteinBot_homo_sapiens	1 day 21 hr - #12	N/A
	GOBot_bigmem	2 days 15 hr - #15	9 days 15 hr - #14
	GeneBot_Homo_sapiens	2 days 19 hr - #25	2 days 20 hr - #24
	Disease_Ontology	2 days 23 hr - #11	4 days 13 hr - #8
	GeneDiseaseBot	2 days 23 hr - #9	1 mo 6 days - #2

Feedback loop



Corona viruses added to Wikidata

Virus strain	NCBI Taxon ID	Wikidata Qid	# Genes	# Proteins
SARS virus	694009	Q278567	14	11
Middle East respiratory syndrome coronavirus	1335626	Q4902157	11	9
Human coronavirus NL63	277944	Q8351095	7	6
Human coronavirus 229E	11137	Q16983356	8	8
Human coronavirus HKU1	290028	Q16983360	9	9
Human coronavirus OC43	31631	Q16991954	9	8
SARS-CoV-2	2697049	Q82069695	11	27


Strain shex...
parent ta...
Disease (E69) (eg: Q...
has effect...
Virus strain shex (E...
parent taxon...
Virus taxon shex (E1...
found in taxon (P703)
A virus protein shex...
encoded by (P702)
encodes (P688)
A virus gene shex (E...
has part (P527)
Protein shex...
WikiPathways...
Gene shex...

Conclusions & future work

Towards an entity schemas ecosystem

Wikidata as a hub for collaborative work

Further work

- Apply the protocol to other domains in wikidata
- Wikibase related projects
- Wikidata subsetting work (see supplementary slides)

Acknowledgments

Maastricht University (NL)
- Egon L. Willighagen
- Martina Kutmon

The Scripps Research Institute (USA)
- Andrew I. Su,

University of Oviedo (ES)
- Daniel Fernández-Álvarez,

Meise Botanic Garden (BE)
- Quentin Groom,

Wageningen University (NL)
- Peter J. Schaap
- Jasper J. Koehorst

Intervacc (NL)
- Lisa M. Verhagen

Gene Wiki

ShEx CG

Wikidata Community

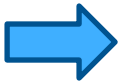
Funders

- National Institute of General Medical Sciences (R01 GM089820)
- Alfred P. Sloan Foundation (grant number G-2019-11458)
- Spanish Ministry of Economy and Competitiveness (Society challenges: TIN2017-88877-R)
- Netherlands Organisation for Scientific Research funded UNLOCK project (NRGWI.obrug.2018.005)
- SYNTHESYS+ a Research and Innovation action funded under H2020-EU.1.4.1.2. Grant agreement ID: 823827.
- ZonMw (grant number: 10430012010015)

Use..

Simple data retrieval

"Retrieve genes with GWAS association with asthma"



39 genes

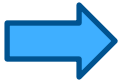
gene	geneLabel		gene	geneLabel		gene	geneLabel		gene	geneLabel
Q5013317	COL22A1		Q18027370	IGSF3		Q18053559	CDHR3		Q14903974	SMAD3
Q14912759	SLC22A5		Q18045382	HPSE2		Q18045669	ATG3		Q18033889	IL1RL1
Q14914243	PSAP		Q18048437	IL33		Q18035037	RAD50		Q17917202	ERBB4
Q14907990	SLC30A8		Q18051900	PYHIN1		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

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."

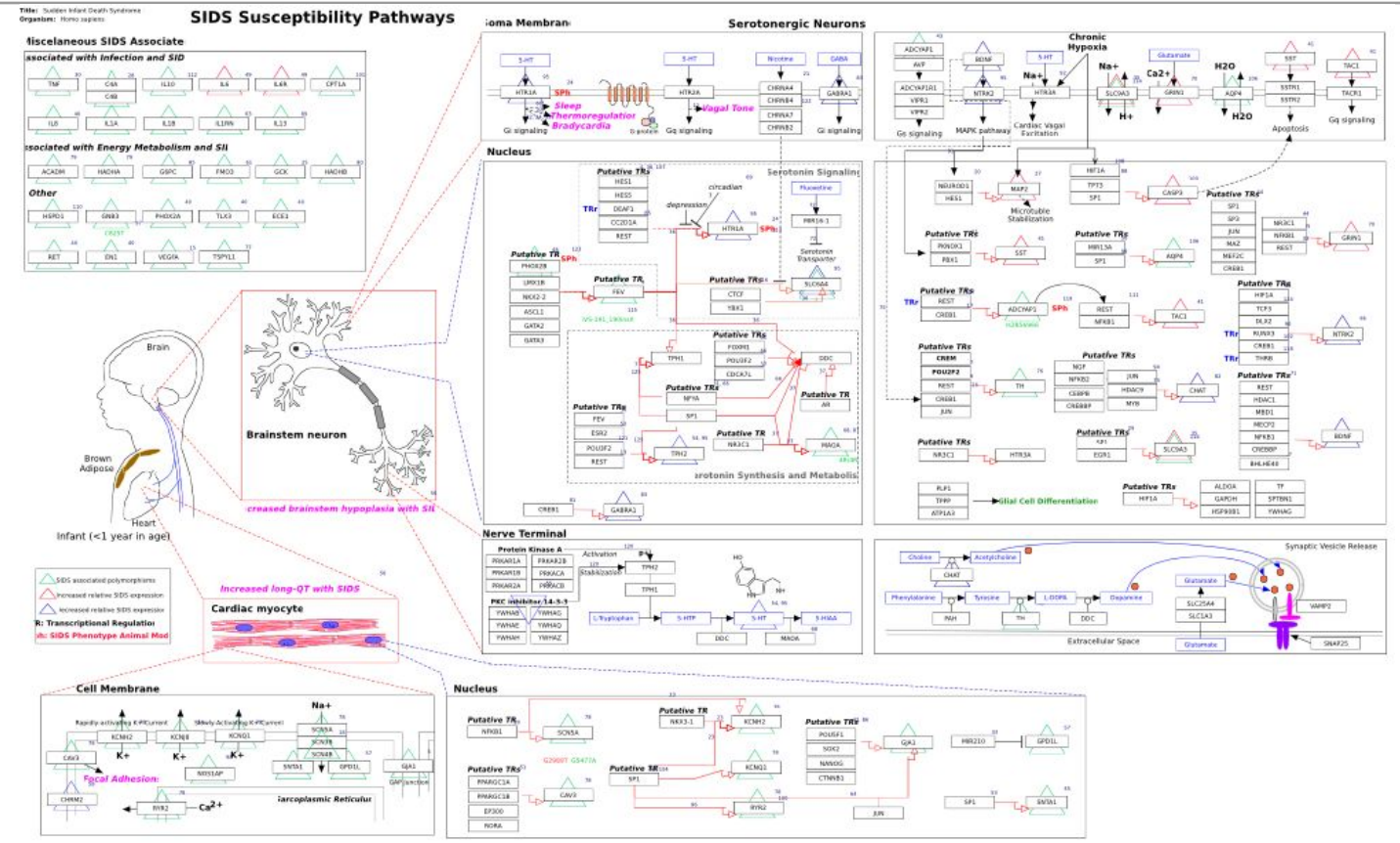
16 genes / 59 pathways

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". }
}
```

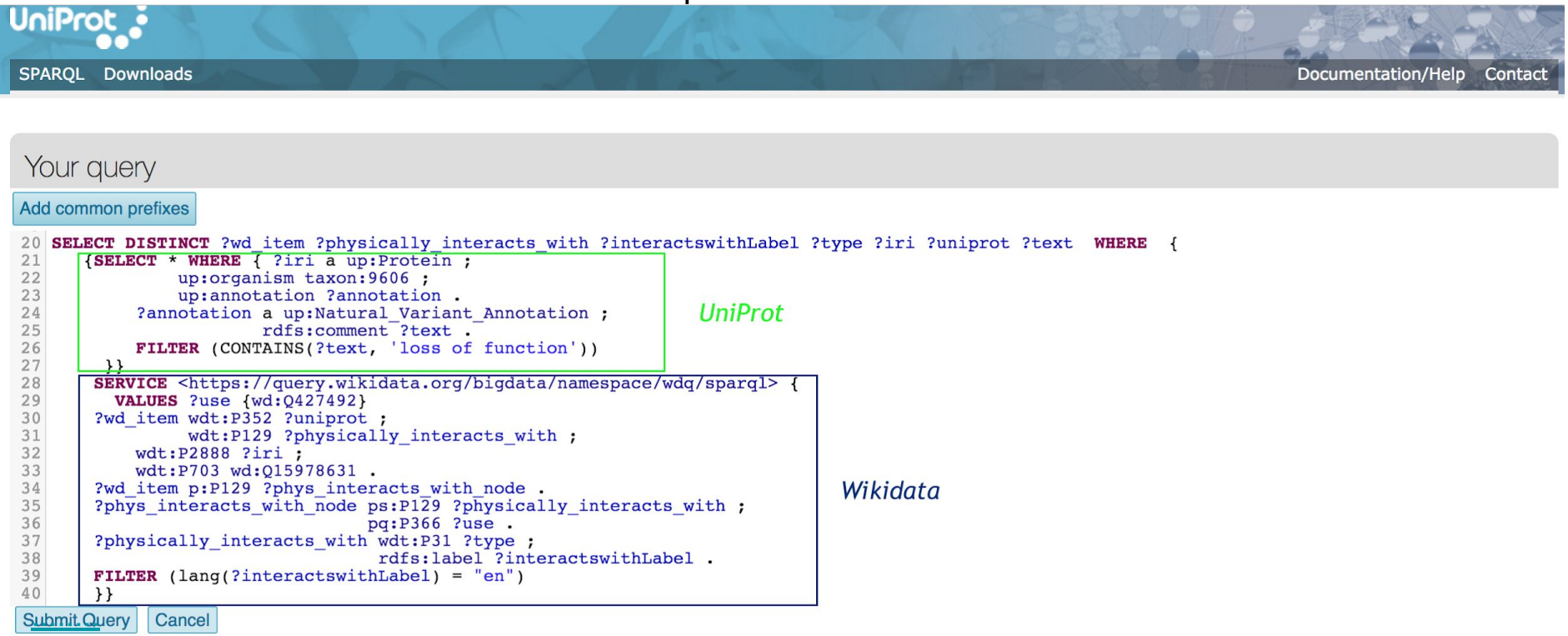
Wikidata

Wikipathways

Wikidata

Try me....

From a remote SPARQL endpoint to Wikidata



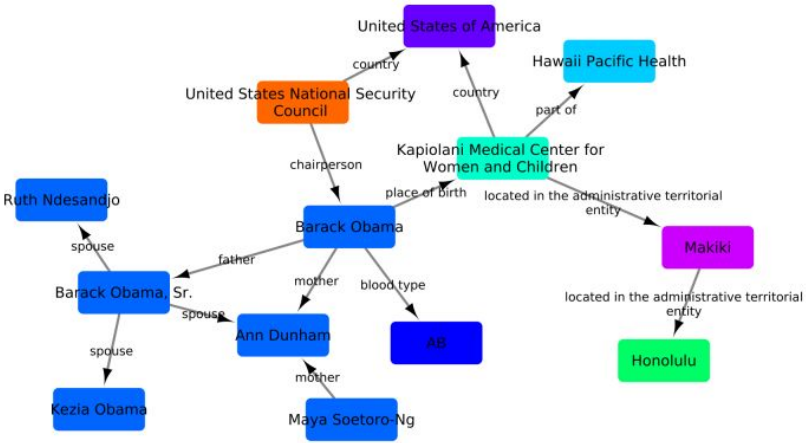
Wikidata is to data as Wikipedia is to text

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

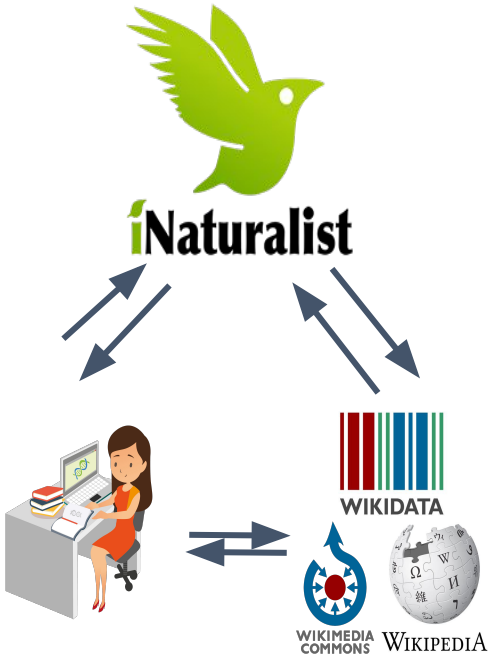
- Completely free, even for commercial usage (CC0)
- Anybody can contribute
- Covers all domains of knowledge
- Extensive item history, talk pages, projects, users
- Integration with the semantic web
- High performance query engine (SPARQL)

- 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!



Introducing Wikiproject iNaturalist



Acknowledgements

Wikidata as a FAIR knowledge graph for the life sciences

Andra Waagmeester, Gregory Stupp, Sebastian Burgstaller-Muehlbacher, Benjamin M Good, Malachi Griffith, Obi Griffith, Kristina Hanspers, Henning Hermjakob, Kevin Hybiske, Sarah M. Keating, Magnus Manske, Michael Mayers, Elvira Mitraka, Alexander R. Pico, Timothy Putman, Anders Riutta, Núria Queralt-Rosinach, Lynn M. Schriml, Denise Slenter, Ginger Tsueng, Roger Tu, Egon Willighagen, Chunlei Wu, Andrew I Su

doi: https://doi.org/10.1101/799684

Thousands of Wikidatans

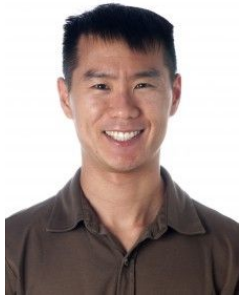
By Helpameout - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=20337311

Funding

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

Acknowledgements

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)