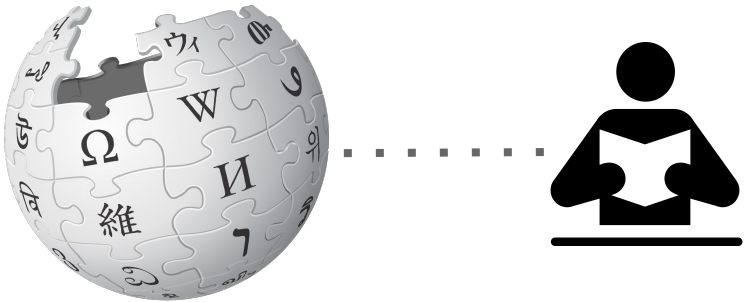
[*citations needed*]

for the sum of all human knowledge

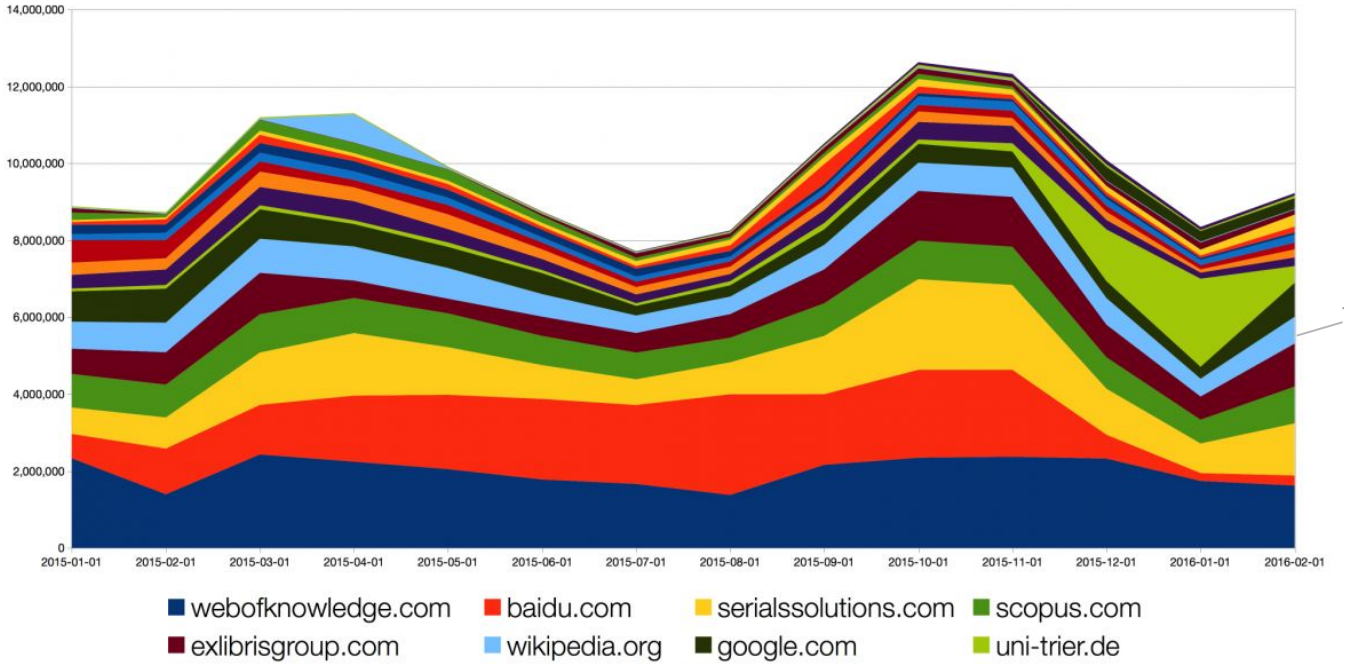
COASP 2016 • September 21, 2016

Dario Taraborelli
@readermeter

1. a major entry point into the scholarly literature

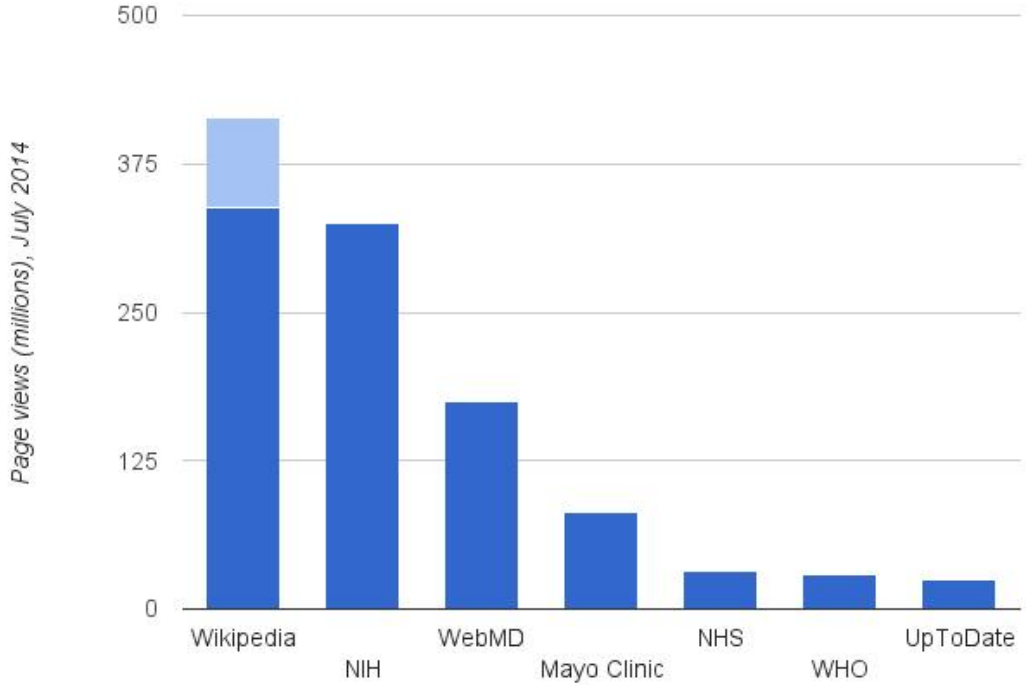


top sources of DOI lookups



http://crosstech.crossref.org/2014/02/many-metrics-such-data-wow.html
http://blog.crossref.org/2016/05/https-and-wikipedia.html

world’s most accessed online medical resource



Heilman and West (2015) doi.org/10.2196/jmir.4069

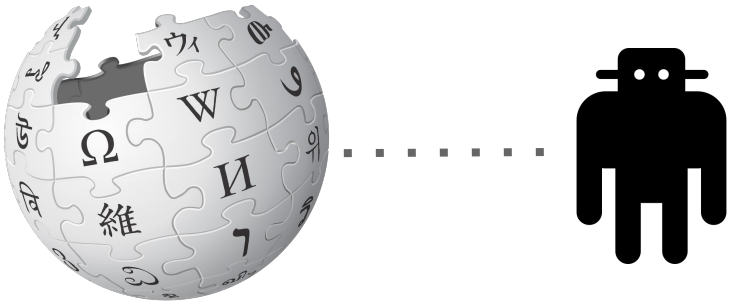
most visited resource on Ebola in West Africa

Most used internet site in Liberia, Sierra Leone and Guinea for Ebola during 2014 outbreak

Greater than CNN, CDC and WHO

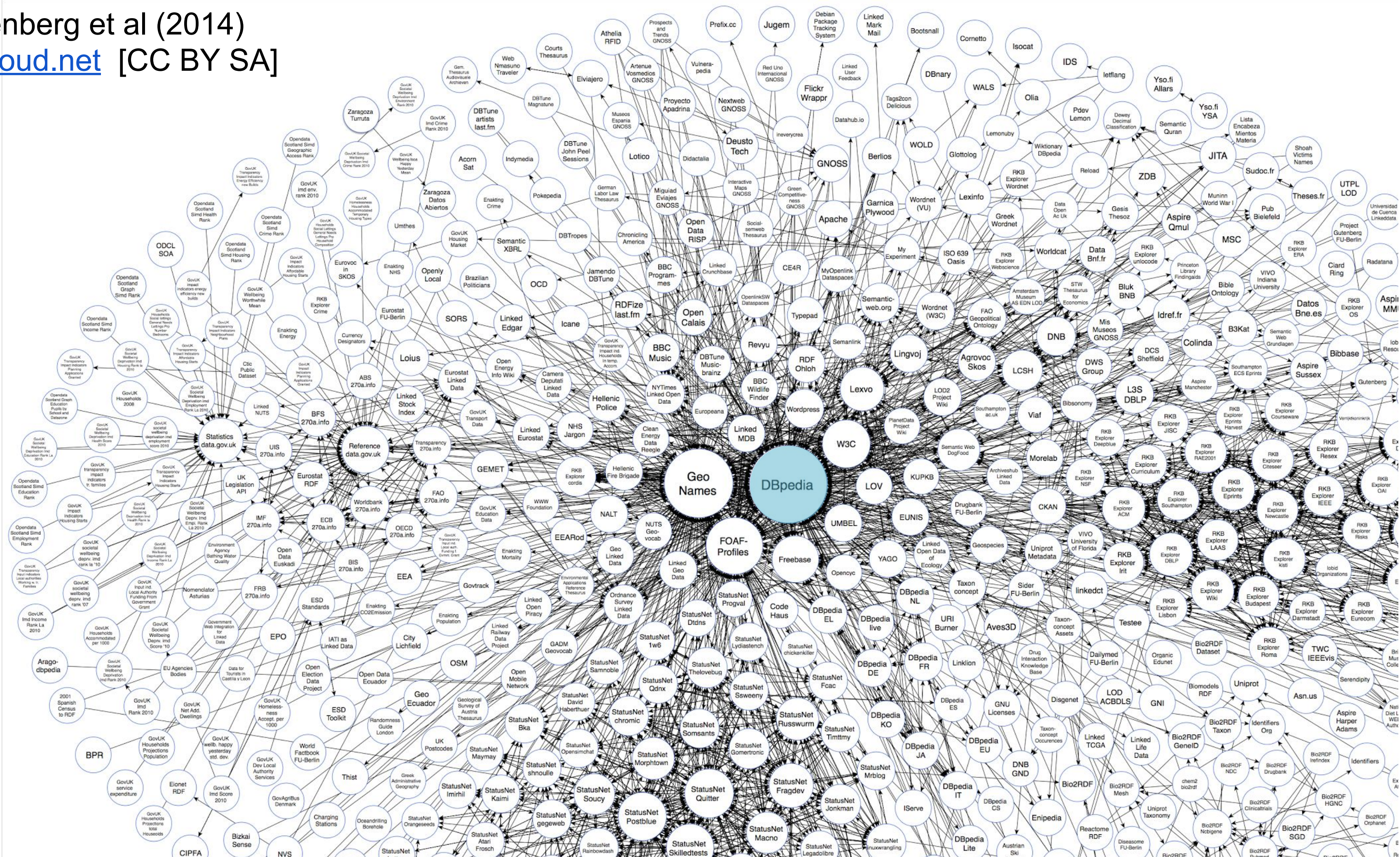
Heilman (2016) http://tinyurl.com/jfuyduv

2. central hub in the linked open data ecosystem



Schmachtenberg et al (2014)

http://lod-cloud.net [CC BY SA]



Challenges 🤔

Machine-readable knowledge base
Editable by anyone
Supporting human + algorithmic curation
Comprehensive coverage
Transparently verifiable


WIKIDATA

Free knowledge base *that anyone can edit*

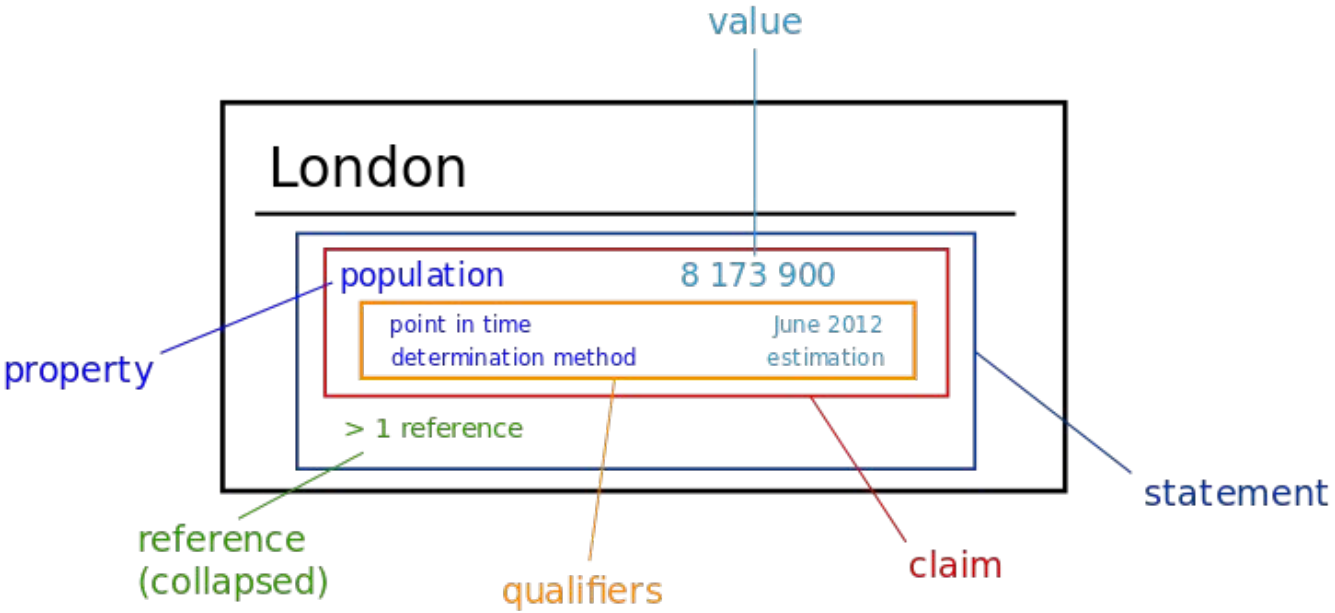
Launched in 2012

Integrated with Wikipedia and other sister projects

Statistics (September 2016)

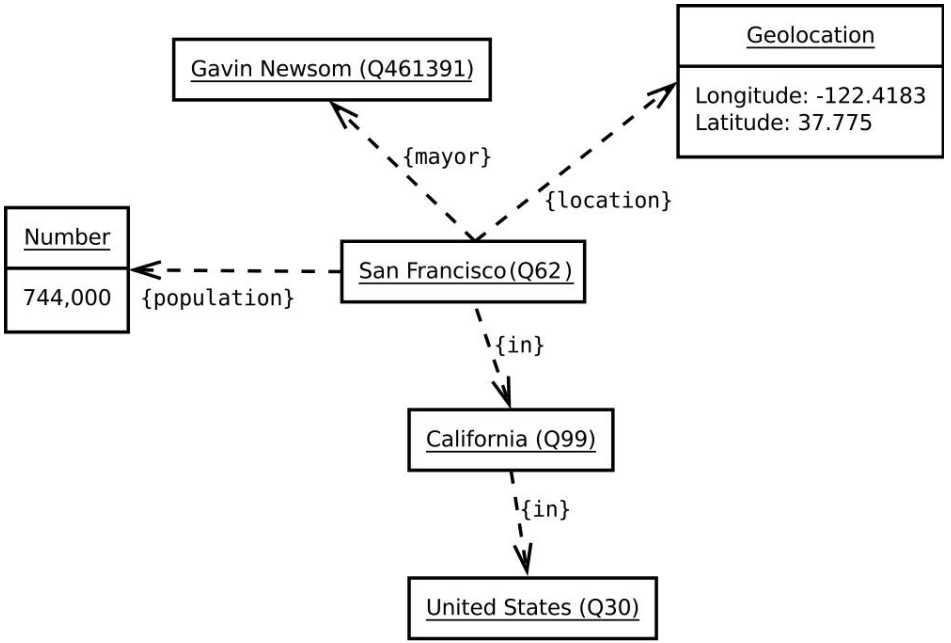
- Over 20M items
- Over 100M statements
- Fastest growing active editor population among largest Wikimedia projects

Wikidata’s anatomy



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

Wikidata’s anatomy

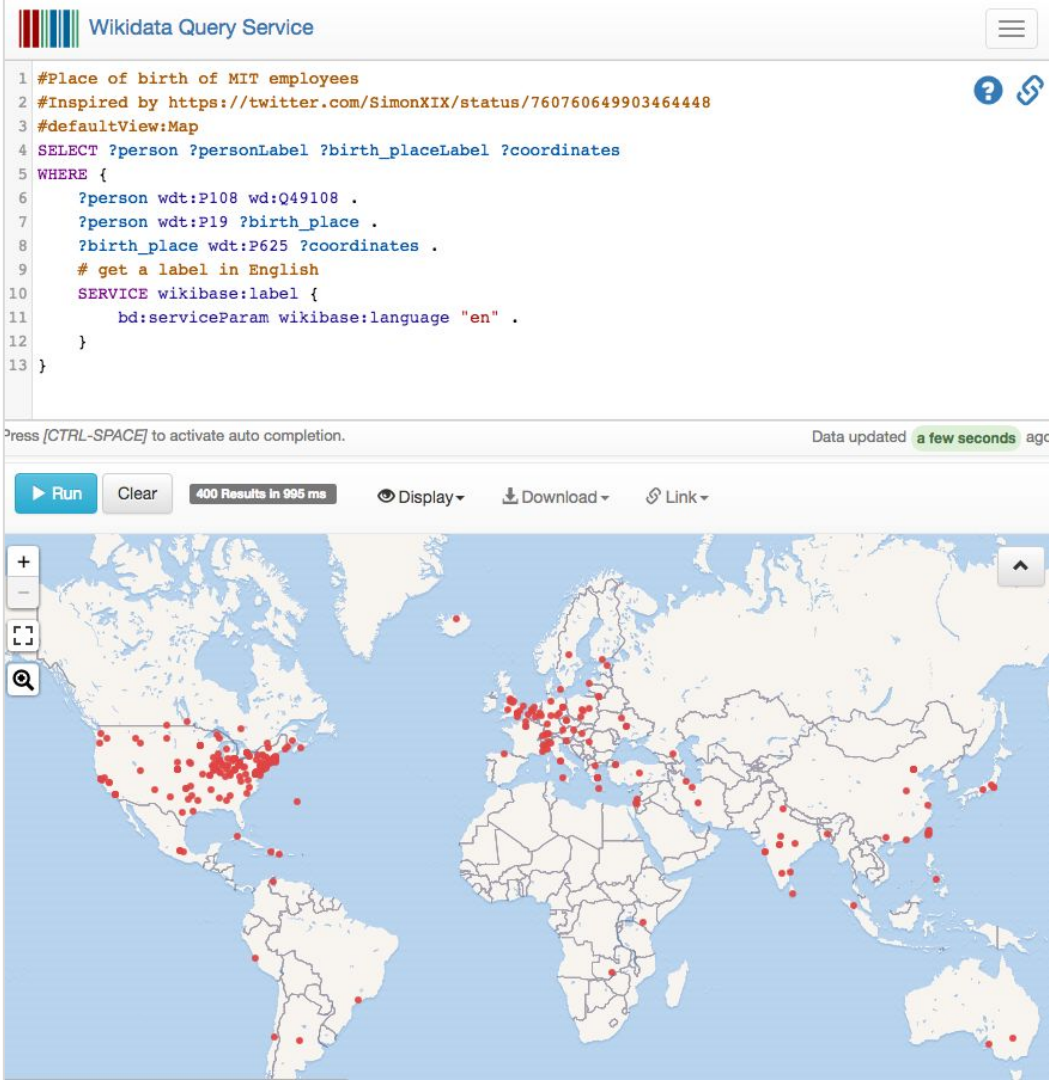


Linked data, San Francisco, Jeblad
https://commons.wikimedia.org/wiki/File:Linked_Data_-_San_Francisco.svg [CC BY SA]

Wikidata queries

Birth place of people employed by MIT

SPARQL:
https://t.co/cDR4Lt7V6P



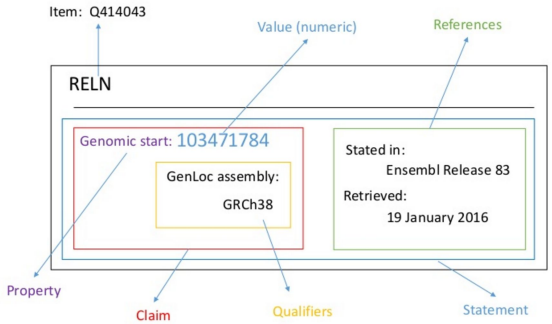
Wikidata queries

Authors with a known location and ORCID

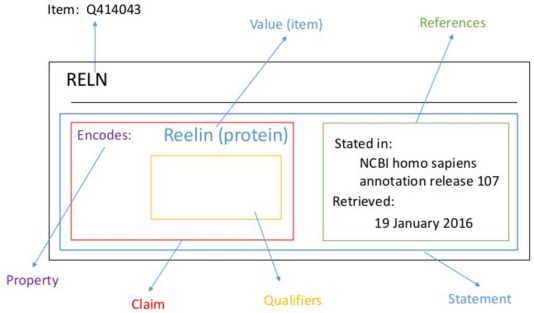
SPARQL:
http://tinyurl.com/h2lqv9y

Expert curation of scientific open data

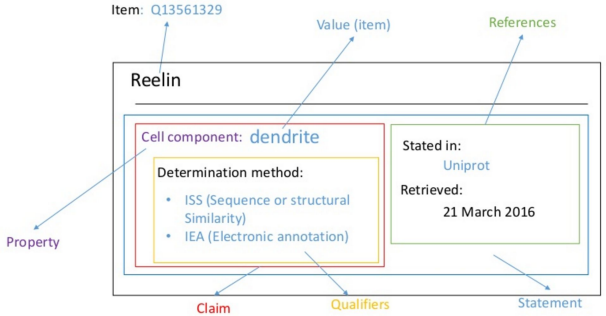
Genomic position for Reelin gene



Linking the Reelin gene to a protein it encodes



Gene ontology annotation for Reelin protein with evidence codes modeled as qualifiers



Benjamin Good (2016) *Opportunities and challenges presented by Wikidata in the context of biocuration*
http://tinyurl.com/hk9qrmz

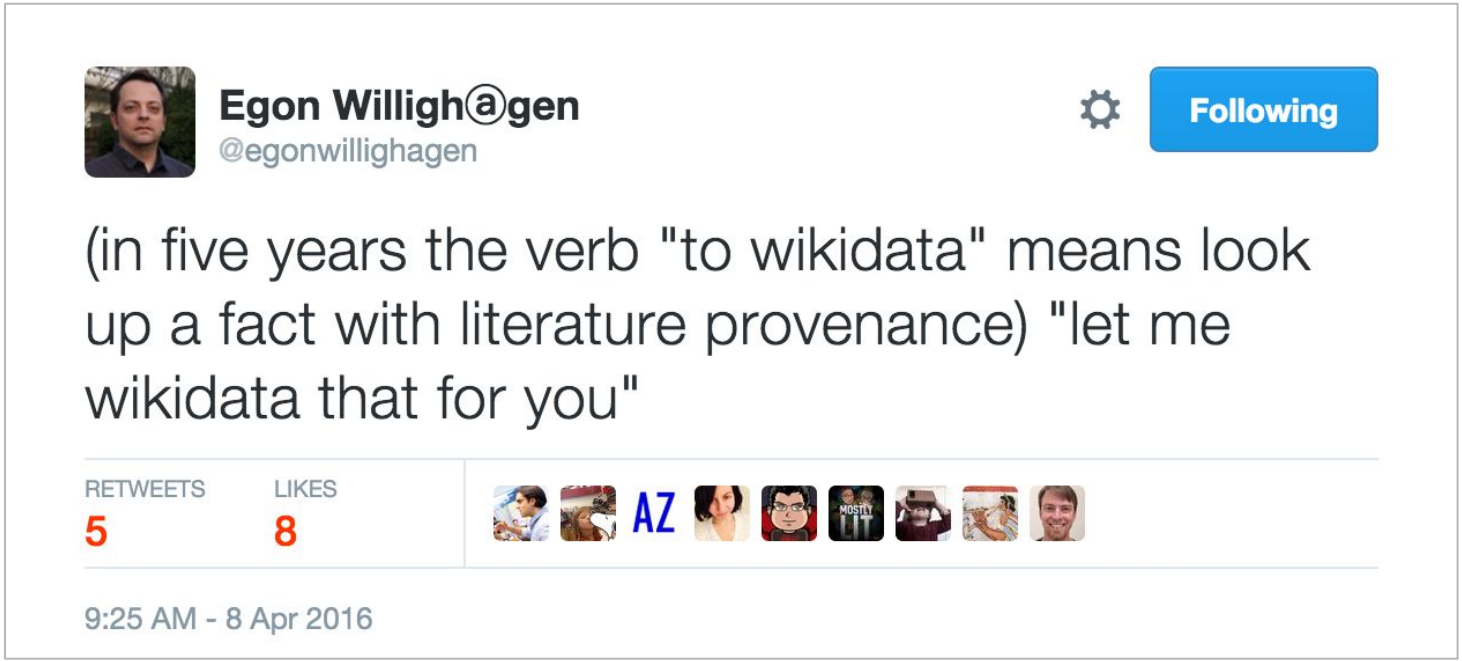
Expert curation of scientific open data

Get all known *drug-drug interactions* for Methadone via its CHEMBL id

Get a list of all diseases *known to be treated* by Metformin

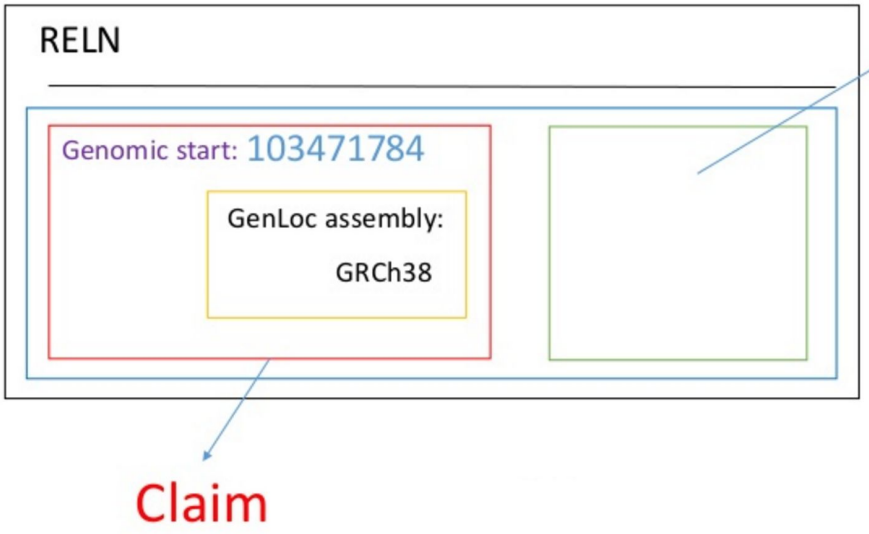
Get a list of all diseases that *might be treated* by Metformin

Gene Wiki: WIkidata SPARQL examples
https://bitbucket.org/sulab/wikidatasparqlexamples/overview



https://twitter.com/egonwillighagen/status/718474906858582016

Computable trust



Add References

1. Add references
2. Check that references concur with the claim or not
3. Estimate 'truthiness' of claim
4. Provide humans with sources to follow up.

Benjamin Good (2016) *Opportunities and challenges presented by Wikidata in the context of biocuration*
http://tinyurl.com/hk9qrmz

WIKIMEDIA FOUNDATION

WIKIMEDIA DEUTSCHLAND

Alfred P. Sloan FOUNDATION

GORDON AND BETTY MOORE FOUNDATION

Crossref

WikiCite: goals

Build a repository of all Wikimedia citations
and bibliographic metadata

Design data models and technology to improve the coverage, quality, standards-compliance and machine-readability of citations and bibliographic metadata in Wikimedia projects

@wikicite • meta.wikimedia.org/wiki/WikiCite

All biomedical OA review articles of the last 5 years

cites (P2860)

bibliographic citation | citation

citation from one creative work to another

Type: WikibaseItem
subproperty of: cites is not a subproperty of any other property
instance of: cites is a(n) Wikidata property for items about works

Property Usage

Entities	77635 PD 0332991, a selective cyclin D kinase 4/6 inhibitor, preferentially inhibits proliferation of luminal estrogen receptor-positive human breast cancer cell lines in vitro , Animal cell differentiation patterns suppress somatic evolution , Tetrahydrohyperforin and octahydrohyperforin are two new potent inhibitors of angiogenesis ,
Values	How Wikipedia Works , The Assayer , The Wealth of Networks , Planned Parenthood v. Casey , Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid , Sociobiology: The New Synthesis , Sidereus Nuncius , The Extended Mind , The Double Helix , Imperialism, the Highest Stage of Capitalism , The Descent of Man, and Selection in Relation to Sex , Griswold v. Connecticut , The Structure of Scientific Revolutions , What Is to Be Done? , Epperson v. Arkansas , As We May Think , View of Delft , The Third Chimpanzee , Roe v. Wade , On the origin of species , ... further results
Typical Properties	PMCID , PubMed ID , page , volume , issue , DOI , short author name , published in , publication date , title , original language of work , main subject , external data available at , article ID , ZooBank publication ID , Chinese Library Classification , NIOSHTIC-2 ID , sponsor , JSTOR article ID
Statements	501929 (6.47 per entity)
Qualifiers	has quality 242 , URL 2
Uses as qualifier	7
Uses in references	5

https://tools.wmflabs.org/sqid/#/view?id=P2860

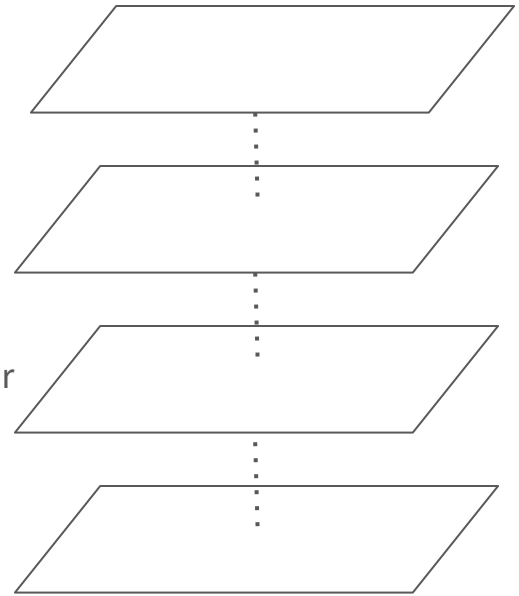
The Zika corpus

Encyclopedic layer

Expert annotation layer

Bibliographic metadata layer

Open citation graph layer



The Zika corpus

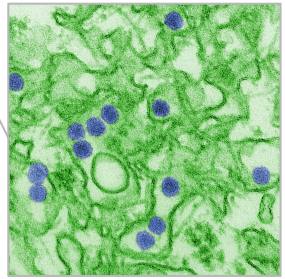
Encyclopedic layer

Zika virus

From Wikipedia, the free encyclopedia

This article is about the virus. For the disease, see Zika fever. For the current outbreak, see 2015–16 Zika virus epidemic.

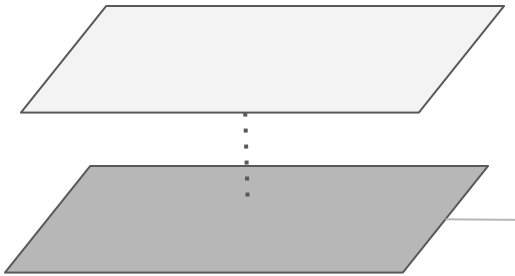
Zika virus (**ZIKV**) is a member of the virus family *Flaviviridae* and the genus *Flavivirus*.[3] It is spread by daytime-active *Aedes* mosquitoes, such as *A. aegypti* and *A. albopictus*.[3] Its name comes from the Zika Forest of Uganda, where the virus was first isolated in 1947.[4] Zika virus is related to the dengue, yellow fever, Japanese encephalitis, and West Nile viruses.[4] Since the 1950s, it has been known to occur within a narrow equatorial belt from Africa to Asia. From 2007 to 2016, the virus spread eastward, across the Pacific Ocean to the Americas, leading to the 2015–16 Zika virus epidemic.



The Zika corpus

Encyclopedic layer

Expert annotation layer



pathogen transmission process

mosquito borne transmission (type of insect borne pathogen tranmission)

Reference

stated in: Concurrent outbreaks of dengue, chikungunya and Zika virus infections – an unprecedented epidemic wave of mosquito-borne viruses in the Pacific 2012–2014 (scientific article)

contact transmission (type of direct pathogen transmission)

placental transmission (type of congenital pathogen transmission process)

Reference

stated in: Zika virus damages the human placental barrier and presents marked fetal neurotropism (scientific article)

sexual intercourse (insertion of a male's penis into a female's vagina for the purposes of sexual pleasure, reproduction, or both)

Reference

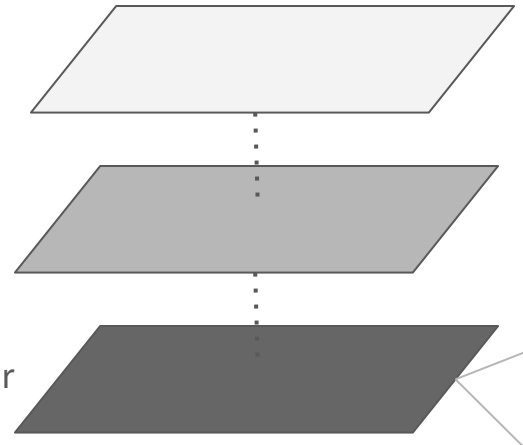
stated in: Potential sexual transmission of Zika virus (scientific article)

The Zika corpus

Encyclopedic layer

Expert annotation layer

Bibliographic metadata layer



Potential sexual transmission of Zika virus (Q22330722)

scientific article

instance of: Potential sexual transmission of Zika virus is a(n) scientific article

Statements

Own statements	From related entities
page	359-61
volume	21
issue	2
cites	Zika virus. I. Isolations and serological specificity (1952 scientific article)
9 statements	Rapid spread of emerging Zika virus in the Pacific area (scientific article)
	Zika Virus, French Polynesia, South Pacific, 2013
published in	Emerging Infectious Diseases (medical journal)
title	Potential sexual transmission of Zika virus [en]
publication date	2015-02
original language of work	English (West Germanic language originating in England)
main subject	Zika virus (species of virus)
	Zika fever (infectious arboviral disease)
author	Didier Musso (researcher) series ordinal : 1
6 statements	Van-Mai Cao-Lormeau (virologist) series ordinal : 6
	Anita Teissier (epidemiologist) series ordinal : 5

Identifiers

PMCID	4313657
PubMed ID	25625872
DOI	10.320...141363

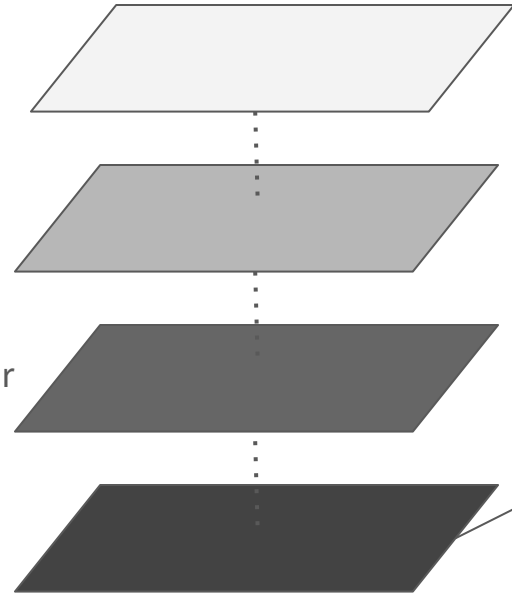
The Zika corpus

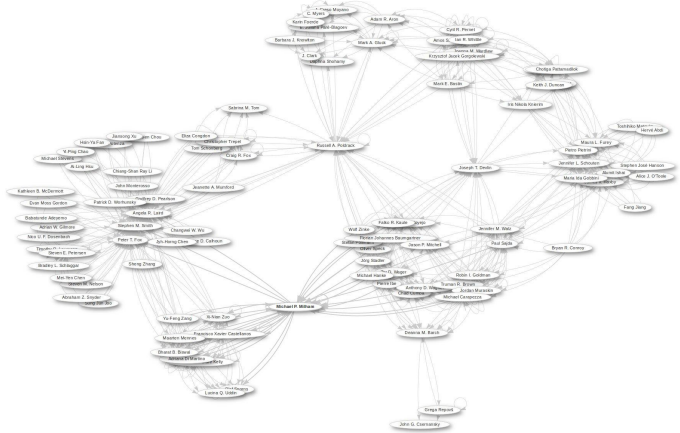
Encyclopedic layer

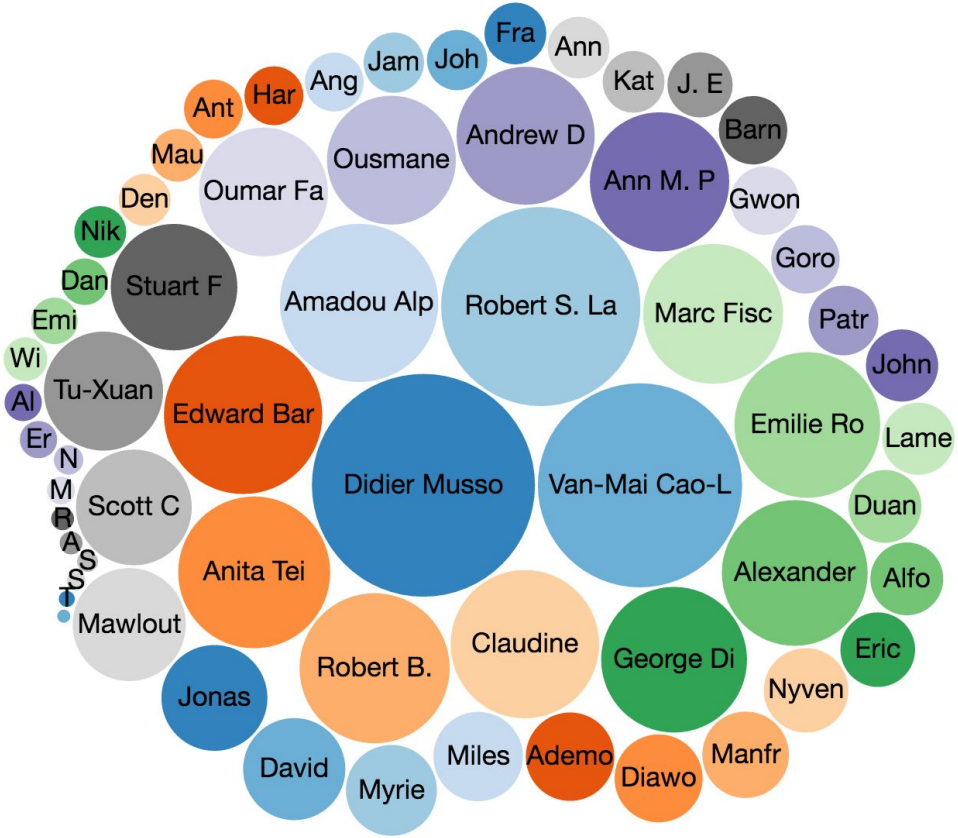
Expert annotation layer

Bibliographic metadata layer

Open citation graph layer







Most cited authors in the Zika research corpus (filtered by journal, OA status or type of statement)

SPARQL: http://tinyurl.com/jb8da68

Wikidata	Sentence	Extracted Statement	Reference
`<Germany, participant of, Miracle of Cordoba>`	*"(...) The Miracle of Cordoba, when they eliminated Germany from the 1978 World Cup"*	`<Germany, eliminated in, Miracle of Cordoba>`	The Telegraph
`<Germany, team manager, Franz Beckenbauer>`	*"In 1984 Beckenbauer was appointed manager of the West German team"*	`<West German team, manager, Beckenbauer>`	Encyclopædia Britannica
`<Germany, inception, 1908>`	*"The story of the DFB's national team began (...) on April 5th 1908"*	`<DFB's national team, start, 1908>`	DFB
`<Germany, captain, Michael Ballack>`	*"Michael Ballack, the captain of the German national football team"*	`<German national football team, captain, Michael Ballack>`	Spiegel

Semi-automated recommendation of entities, missing statements, references for unsourced statements

https://www.wikidata.org/wiki/Wikidata:Primary_sources_tool
https://meta.wikimedia.org/wiki/Grants:IEG/StrepHit:_Wikidata_Statements_Validation_via_References
https://meta.wikimedia.org/wiki/Grants:Project/WikiFactMine

all statements citing a *New York Times article*

most popular journals cited by statements of any item that is a subclass of *economics*

all statements citing the works of *Joseph Stiglitz*

all statements citing journal articles by *physicists at Oxford University in the 1970s*

all statements citing a journal article that was *retracted*

all statements citing *a source that cites a journal article that was retracted*

https://meta.wikimedia.org/wiki/WikiCite_2016/Report/Group_5

Asks 

1. release open citation data

Distributing references via Crossref: blog.crossref.org/2016/06/distributing-references-via-crossref.html

1. release open citation data
2. use licenses supporting content mining

The Right to Read Is the Right to Mine: blog.okfn.org/2012/06/01/the-right-to-read-is-the-right-to-mine
Crossref Text and Data Mining Services: tdmsupport.crossref.org/

Accelerate the *discoverability*, *reusability*, and *societal impact* of open access

Thank you

Acknowledgments

Daniel Mietchen, Jonathan Dugan, Lydia Pintscher, Cameron Neylon, James Hare, James Heilman, Magnus Manske, Egon Willighagen, the Gene Wiki team (especially Andra Waagmeester, Tim Putman, Benjamin Good), the ContentMine team, the University of Chicago Knowledge Lab, all WikiCite 2016 participants and Wikidata Source Metadata project contributors.

Additional image credits

Library, National Park Service Collection thenounproject.com/term/library/191/ [CC0]
Robot, Creative Stall thenounproject.com/term/robot/132360/ [CC BY]
Open Access logo commons.wikimedia.org/wiki/File:Open_Access_logo_PLoS_transparent.svg [CC0]

dario@wikimedia.org • @readermeter • @Wikidata • @WikiCite • @WikiResearch