

Open curation of knowledge on biodiversity: An introduction to Wikidata



Q43624

Smithsonian Tropical Research Institute • 27 February 2019

Dario Taraborelli
@readermeter

Outline

1. My journey: research and advocacy on open knowledge
2. Wikipedia and scientific knowledge dissemination
3. An introduction to Wikidata
4. Wikidata for science
5. Best practices to support open curation

My journey with open knowledge

Team



Dario Taraborelli
Director, Head of Research



Leila Zia
Senior Research Scientist, Lead



Jonathan Morgan
Senior Design Researcher



Isaac Johnson
Research Scientist



Bahodir Mansurov
Software Engineer



Miriam Redi
Research Scientist



Diego Sáez-Trumper
Research Scientist



Erik Zachte
Senior Data Analyst

External team members



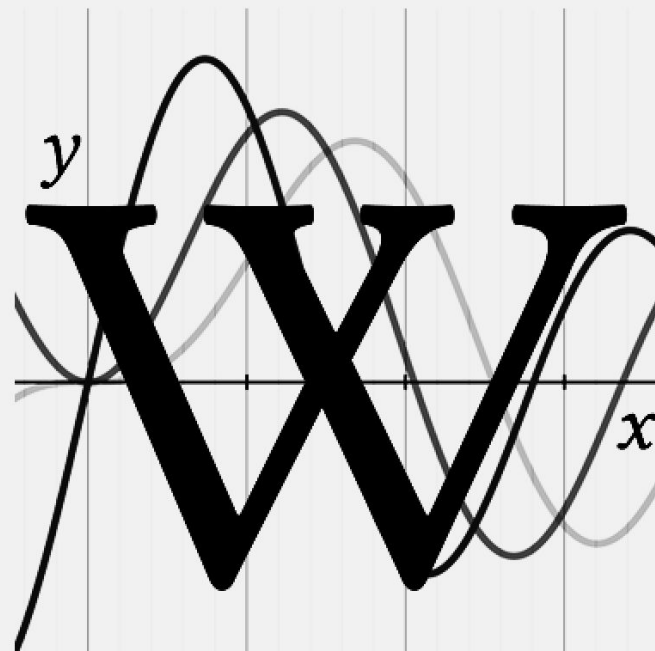
Aaron Halfaker
*Principal Research Scientist,
Scoring Platform*



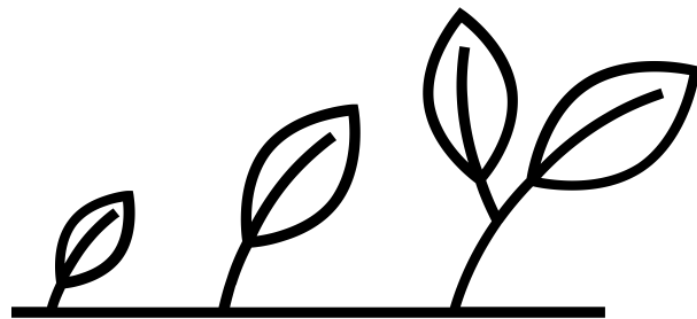
Morten Warncke-Wang
Research Fellow



Robert West
*Research Fellow; Assistant
Professor, EPFL*



Wikimedia Research



[About](#)[Team](#)[Collaborators](#)[Projects](#)[Publications](#)[News](#)[Events](#)[Contact](#)

Wikimedia Research

Wikimedia Research is a team of [scientists and engineers](#) using data to understand and empower millions of readers and contributors who interact with Wikipedia and its sister projects on a daily basis.

We turn research questions into [publicly shared knowledge](#). We design and test new technology, produce empirical insights to support new products and programs, and publish research informing the organization's and the movement's strategy.

We are strongly committed to principles of transparency, privacy, and open collaboration. We collaborate with [a network of researchers](#) in the industry and academia, and publish [all our output](#) in the open.

Featured Projects

Knowledge Gaps

Developing systems to identify content gaps across Wikipedia languages and recommend contributions based on editors' interests.

Scoring Platform

Enhancing the impact of machine classifiers for quality control while minimizing their detrimental effects.

Structured Citations

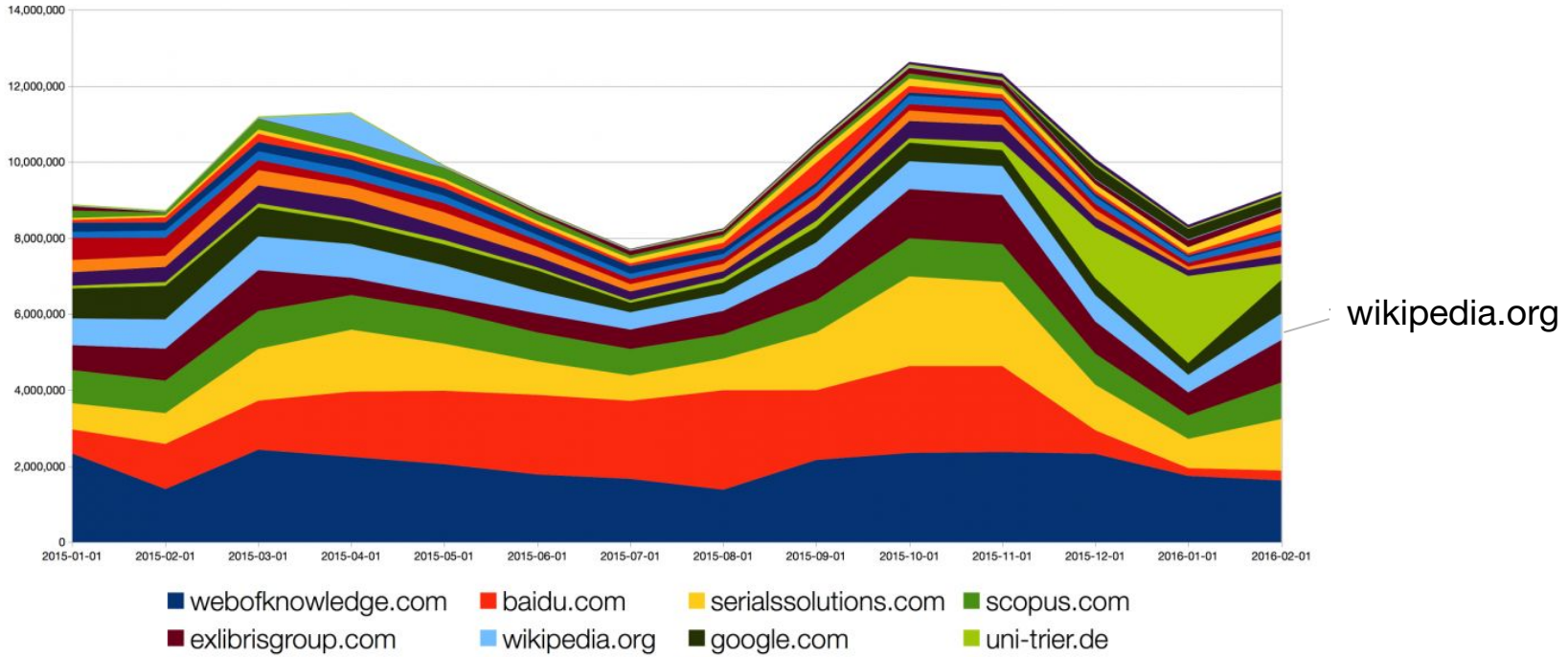
Building a structured repository of citable sources that can serve Wikipedia and other open knowledge projects.



A major entry point into the scholarly literature



One of the top sources of readership to the scholarly literature



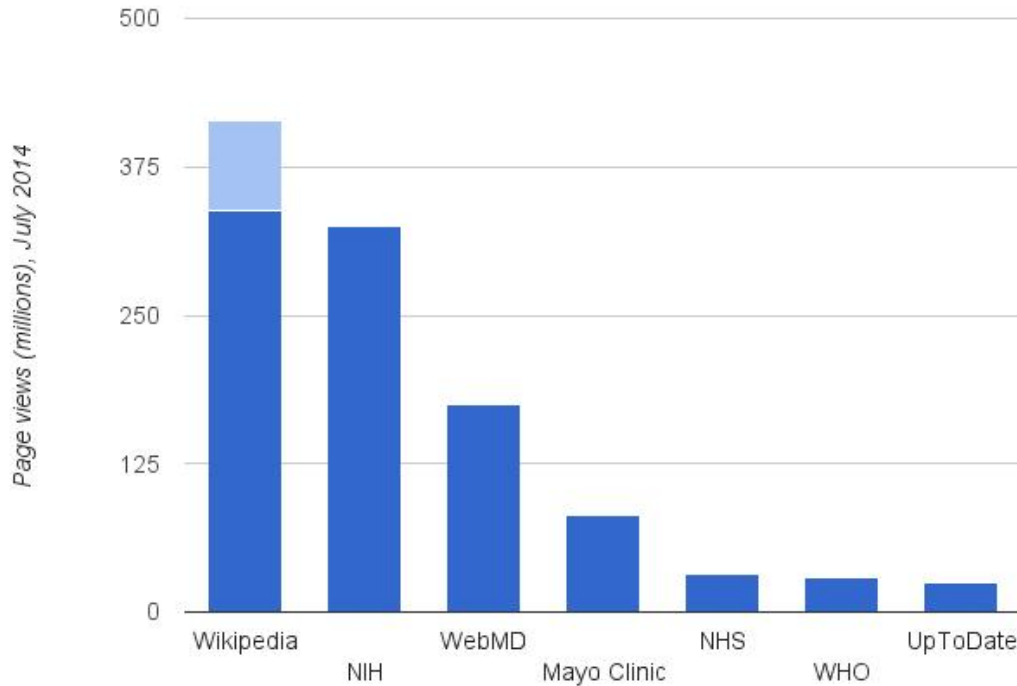
crosstech.crossref.org/2014/02/many-metrics-such-data-wow.html
blog.crossref.org/2016/05/https-and-wikipedia.html

A preferred resource for (pre-)research

Results: The participants valued lecture notes (73.7%) and Wikipedia (74%) as their most important online sources for knowledge acquisition. Missing integration of e-learning into teaching was seen as the major pitfall (58.7%). The learning management system was mostly used for study information (68.3%), preparation of exams (63.3%) and lessons (54.5%). Clarity (98.3%), teaching-related contexts (92.5%) and easy use of e-learning offers (92.5%) were rated highest. Interactivity was most important in free-text comments (n = 123).

Back et al. (2016) *Int J Med Educ.* 7:267-273; doi.org/10.5116/ijme.57a5.f0f5

The world's most visited online medical resource



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

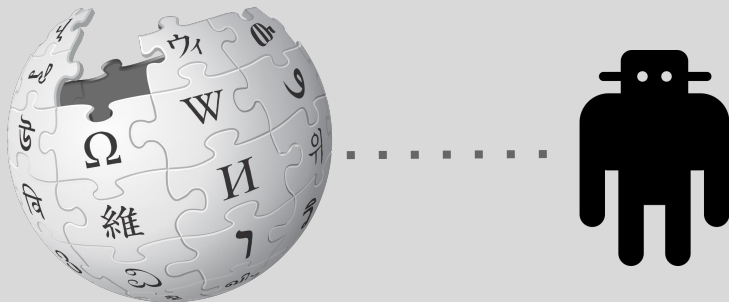
The most accessed resource on Ebola in West Africa

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

Greater than CNN, CDC and WHO



A primary data source for platforms and algorithms

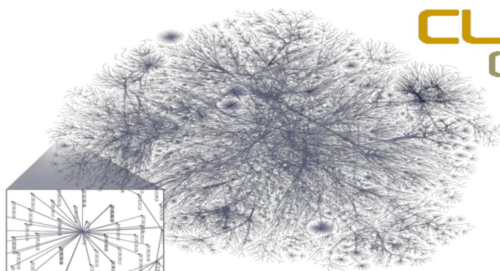


SQuAD

The Stanford Question Answering Dataset

What is SQuAD?

Stanford Question Answering Dataset (SQuAD) is a new reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage. With 100,000+ question-answer pairs on 500+ articles, SQuAD is significantly larger than previous reading comprehension datasets.



CLOUD Graph

Rich Data Graphs for MapReduce

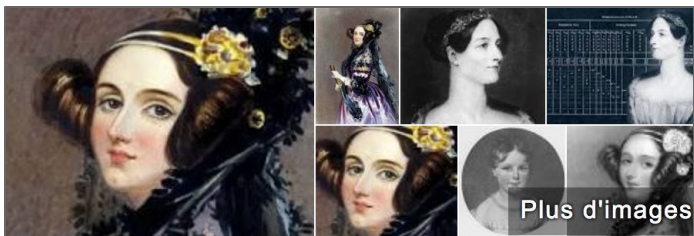
Consuming and producing complex HBase structures in MapReduce™ with on-line NLP enhanced 1.5 Billion Word Wikipedia Text Corpus Example

Scott Cinnamond – TerraMeta Software Inc.
<http://cloudgraph.org>

The Wikipedia Corpus



SEARCH			FREQUENCY	CONTEXT
79	Andes	A B C		endemic to the region, surpassing the diversity of any other hotspot. The small tree "
80	Ambracia	A B C	1372	Amber is fossilized tree resin (not sap), which has been appreciated for its color
81	Amber	A B C		the basis of their chemical constituents. Because it originates as a soft, sticky tree res
82	Amber	A B C		by the Sun, Pliny cites opinions that are well aware of its origin in tree resin, citing the
83	Amber	A B C		Our forefathers, too, were of opinion that it is the juice of a tree , and for this reason g
84	Amber	A B C		of 'succinum' and one great proof that it is the produce of a tree of the pine genus, is
85	Amber	A B C		. It is thought that, in addition to exuding onto the surface of the tree , amber resin als
86	Amber	A B C		created using combinations of labdanum, benzoin resin, copal (itself a type of tree res
87	Alphorn	A B C		generally spruce but sometimes pine. In former times the alphorn maker would find a
88	Alder	A B C		or brown, which is also a root for the English words elk and another tree : elm, a tree c
89	Alder	A B C		also a root for the English words elk and another tree: elm, a tree , distantly related to



Plus d'images

Ada Lovelace

Programmeuse informatique

Ada Lovelace, de son nom complet Augusta Ada King-Lovelace, née Ada Byron le 10 décembre 1815 à Lymington, dans le Hampshire, est une mathématicienne et une scientifique britannique. Elle est considérée comme la première programmeuse d'ordinateur.

Date et lieu de naissance : 10 décembre 1815, Lymington, Hampshire, Royaume-Uni

Date et lieu de décès : 27 novembre 1852, Marylebone, Londres, Royaume-Uni

Nom complet : Augusta Ada King-Noel, Countess of Lovelace

Époux : William King-Noel, 1st Earl of Lovelace (m. 1835)

Enseignement : Université de Londres

Enfants : Anne Blunt, Byron King-Noel, Viscount of Milbanke, 2nd Earl of Lovelace

Recherches associées

Voir d'autres



Charles Babbage



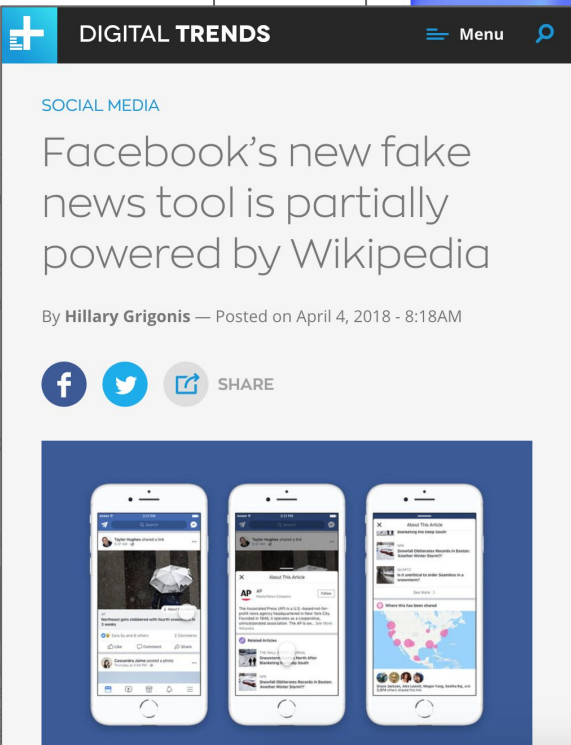
Lord Byron
Père



Grace Hopper

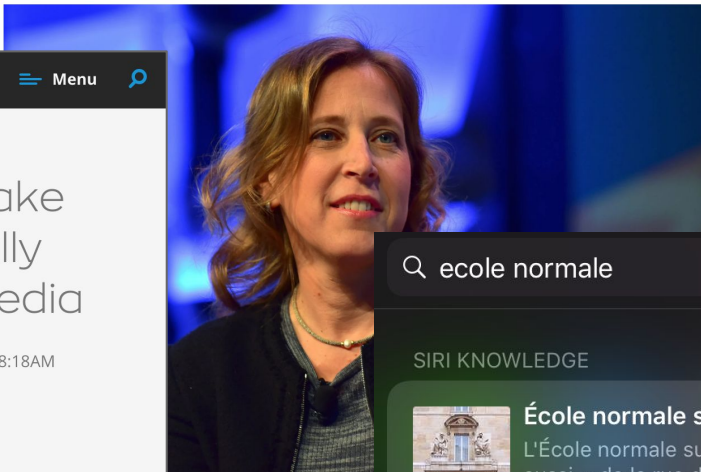


Anne Isabella Milbanke
Mère

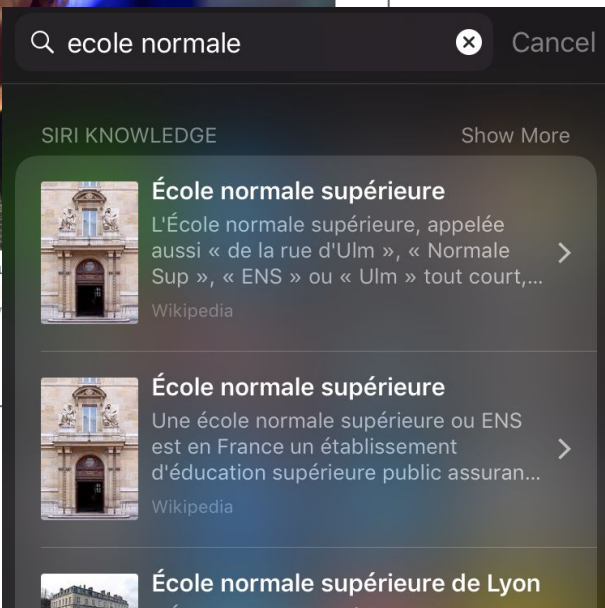


The New York Times

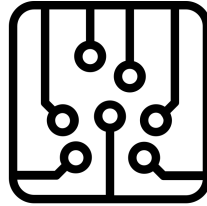
You Tube May Add to the Burdens of Humble Wikipedia



Last week, YouTube's chief executive, Susan Wojcicki, announced that the company was using Wikipedia's help to deal with the platform. Jason Bollenbacher/Getty Images



The epistemic backbone of the internet



Imagine a world in which
everyone can freely share in
the sum of all knowledge.

Imagine a world in which
~~every~~~~one~~~~thing~~ can freely share in
the sum of all **structured** knowledge.

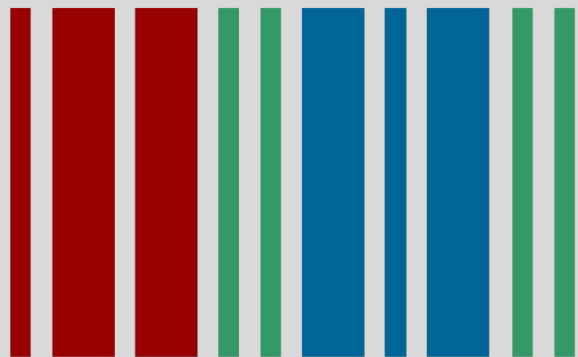
Machine-readable knowledge

Editable by anyone

Support open human + algorithmic curation

Comprehensive coverage

Transparently verifiable



WIKIDATA

Infoboxes

Marie Skłodowska Curie



c. 1920

Born	Maria Salomea Skłodowska 7 November 1867 Warsaw, Kingdom of Poland , then part of Russian Empire ^[1]
Died	4 July 1934 (aged 66) Passy, Haute-Savoie, France
Cause of death	Aplastic anemia from exposure to radiation
Residence	Poland, France
Citizenship	Poland (by birth) France (by marriage)
Alma mater	University of Paris ESPCI ^[2]

Infoboxes

Marie Skłodowska Curie



c. 1920

Born	Maria Salomea Skłodowska 7 November 1867 Warsaw, Kingdom of Poland , then part of Russian Empire ^[1]
Died	4 July 1934 (aged 66) Passy, Haute-Savoie , France
Cause of death	Aplastic anemia from exposure to radiation
Residence	Poland, France
Citizenship	Poland (by birth) France (by marriage)
Alma mater	University of Paris ESPCI ^[2]

Born	Maria Salomea Skłodowska 7 November 1867 Warsaw, Kingdom of Poland , then part of Russian Empire ^[1]
Died	4 July 1934 (aged 66) Passy, Haute-Savoie , France
Cause of death	Aplastic anemia from exposure to radiation
Residence	Poland, France
Citizenship	Poland (by birth) France (by marriage)
Alma mater	University of Paris ESPCI ^[2]

Infoboxes

	居里照，约1920年摄
出生	玛丽亚·萨洛美娅·斯克沃多夫斯卡 Maria Salomea Skłodowska 1867年11月7日 俄罗斯帝国波兰王国华沙 ^[1]
逝世	1934年7月4日（66歲） 法国帕西
死因	再生不良性贫血
公民权	波兰（出生） 法国（成婚）
母校	巴黎大学 巴黎高等物理化工学院
知名于	放射性 钋 镭
配偶	皮埃尔·居里（1895年成婚）
儿女	伊雷娜·约里奥-居里 兹兹·居里

معلومات شخصية	
الاسم عند الولادة	Maria Salomea Skłodowska
الميلاد	7 نوفمبر 1867 وارسو، فستولا لاند، روسيا (في بولندا حاليًا)
الوفاة	4 يوليو 1934 (66 سنة) باسي، باريس، فرنسا
سبب الوفاة	فقر الدم اللاتنسجي ^[1]
مكان الدفن	مقبرة العظام
الإقامة	باريس
مواطنة	روسية (بموجب تبعية بولندا) للالإمبراطورية الروسية)
الجنسية	بولندية، ثم فرنسية فيما بعد
العرق	بولنديون
الديانة	ملحدة
عضوة في	الأكاديمية الملكية السويدية للعلوم، والأكاديمية الروسية للعلوم، والأكاديمية الوطنية للطب
الزوج	بيار كوري (1895-1906)
أبناء	إيرين جوليو-كوري، وإيف كوري

জন্ম	নভেম্বর ৭ ১৮৬৭ ওয়ারশ, কংগ্রেস পোল্যান্ড
মৃত্যু	জুলাই ৪, ১৯৩৪ (৬৬ বছর) Sancellemoz, ফ্রান্স
জাতীয়তা	পোলীয়, ফরাসি
কর্মক্ষেত্র	পদার্থবিজ্ঞান এবং রসায়ন
প্রতিষ্ঠান	সরবোন
প্রাক্তন ছাত্র	সরবোন এবং ইএসপিসিআই
পিএইচডি উপদেষ্টা	অঁরি বেকেরেল
পিএইচডি ছাত্রেরা	অঁদ্রে-লুই দ্যবিয়েঁ Marguerite Catherine Perey
পরিচিতির কারণ	তেজস্ক্রিয়তা
উল্লেখযোগ্য পুরস্কার	 পদার্থবিজ্ঞানে নোবেল পুরস্কার (১৯০৩) রসায়নে নোবেল পুরস্কার (১৯১১)



- 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

Item Discussion

Read

View history

Search Wikidata

Marie Curie (Q7186)

French-Polish physicist and chemist

Maria Salomea Skłodowska | Maria Skłodowska-Curie | M. Curie

In more languages

Statements

instance of



human

edit

1 reference

+ add value

part of



Pierre and Marie Curie

edit

1 reference

+ add value

Wikidata is to data what Wikipedia is to text

- All data CC0
- Anybody can contribute
- Covers all domains of knowledge
- Fully version controlled and collaborative
- Integrated with the semantic web via RDF / open APIs
- High performance query engine (SPARQL)
- Stable. Not tied to short-term funding cycles
- Actively developed, full stack is open source
- Active community. Fastest growing Wikimedia project

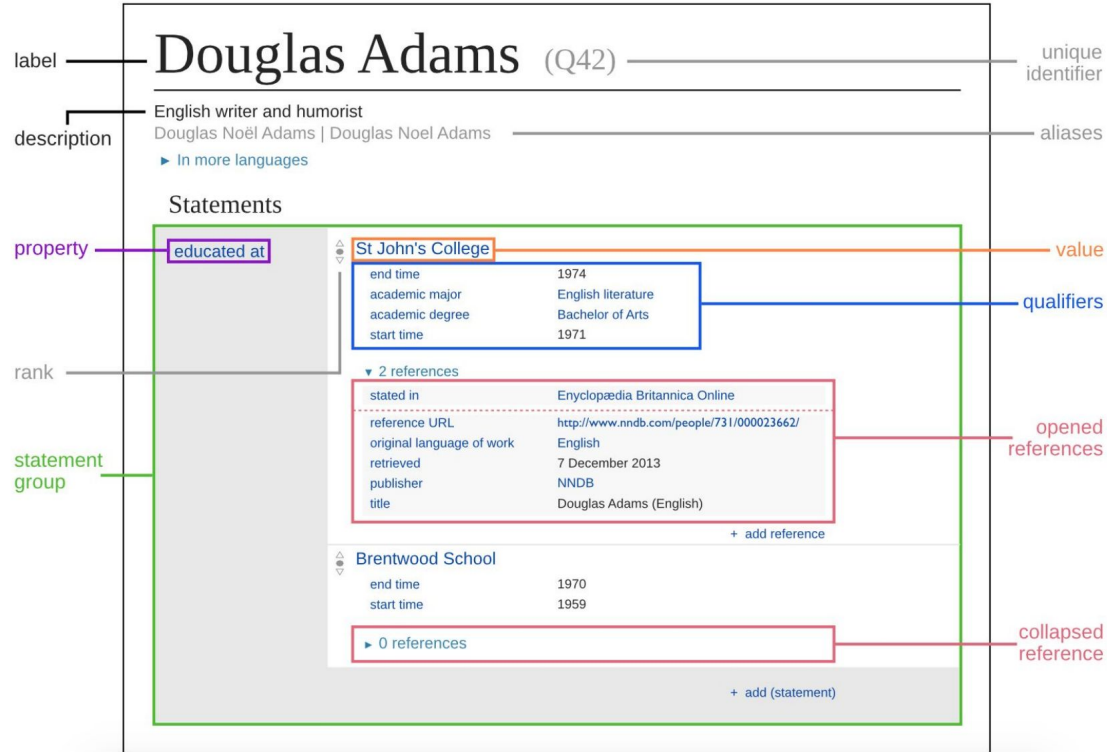


55 million
entities

680M statements • 840M edits

[as of February 2019]

Wikidata's anatomy



Multilingual labels and descriptions

Language	Label	Description	Also known as 
Spanish	Marie Curie	química y física polaca naturalizada francesa	Maria Skłodowska-Curie Maria Salomea Skłodowska Marie Skłodowska-Curie Madame Curie Mme Curie Marie Curie-Skłodowska Madame Pierre Curie M. Curie
Traditional Chinese	瑪麗·居里	物理學家，化學家	居里夫人
Chinese	玛丽·居里	物理学家，化学家	居里夫人
Afrikaans	Marie Curie	Fraans-Poolse fisikus en chemikus	
Amharic	ማሪ ኩሪ	No description defined	
Aragonese	Marie Curie	No description defined	
Arabic		ماري كوري	عالمة كيمياء بولندية
Egyptian Arabic		مارى كورى	No description defined
Assamese	মেডাম কুৰী	No description defined	
Asturian	Marie Curie	química y física polaco-francesa	Maria Salomea Skłodowska Marie Skłodowska-Curie

Statements and references

Statements

instance of

 human  edit

▼ 1 reference

stated in

data.bnf.fr

retrieved

10 October 2015

reference URL

<http://data.bnf.fr/ark:/12148/cb121447141>

+ add reference

+ add value

Identifiers

Identifiers

Mathematics Genealogy

Project ID



177944

▼ 0 references

VIAF ID



76353174

▼ 0 references

Types of content in Wikidata

- People
 - Places
 - Taxa
 - Buildings
 - Organizations
 - Artworks
 - Events
 - Astronomical Bodies
 - ...
- Chemicals
 - Processes
 - Theorems
 - Concepts
 - Creative works
 - Journals
 - Publishers
 - Funders
 - ...

Wikidata for Research



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

Biomedical data



- All human, mouse genes and proteins (swissprot)
- All Gene Ontology terms
- All Human Disease Ontology terms
- All FDA approved drugs
- 109 reference microbial genomes

[Mitraka et al \(2015\) *Semantic Web Applications for the Life Sciences*](#)
[Burgstaller-Muelbacher et al \(2016\) *Database*](#)
[Putman et al \(2016\) *Database*](#)

Biomedical 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

repurpose drugs as potential new treatments for a disease

Gene Wiki: Wikidata SPARQL examples

<https://github.com/SuLab/wikidatasparqlexamples>

Data about biodiversity

White-necked Jacobin (Q1084570)

species of bird

Florisuga mellivora

 [edit](#)

▼ In more languages

Language	Label	Description	Also known as
English	White-necked Jacobin	species of bird	Florisuga mellivora
German	Weißnackenkolibri	Art der Gattung Florisuga	Florisuga mellivora Jakobinerkolibri
Spanish	Florisuga mellivora	especie de ave	Colibrí nuquiblanco Colibri nuquiblanco
French	Colibri jacobin	espèce d'oiseaux	Florisuga mellivora
Italian	Florisuga mellivora	specie di uccello	
Latin	Florisuga mellivora	No description defined	
Dutch	Witnekkolibrie	soort uit het geslacht Florisuga	

[All entered languages](#)

White-necked Jacobin [[Q1084570](#)]

Data about biodiversity

taxon name	⌵ ⌶ Florisuga mellivora edit taxon author Carl Linnaeus year of taxon name 1758 publication instance of recombination ! ▶ 12 references
+ add value	
taxon rank	⌵ ⌶ species edit ▶ 12 references
+ add value	
parent taxon	⌵ ⌶ Jacobin edit ▶ 1 reference
+ add value	

White-necked Jacobin [[Q1084570](#)]

Data about biodiversity

image



White-necked jacobin (*Florisuga mellivora mellivora*) male Tr.jpg
4,174 × 2,783; 1.74 MB

 sex or gender male organism

0 references



White-necked Jacobin 4.jpg
2,702 × 1,930; 3.38 MB

White-necked Jacobin [[Q1084570](#)]

Data about biodiversity

taxon common name	 jacobin cyffredin (Welsh)  edit
	▶ 1 reference
	 White-necked Jacobin (English)  edit
	▶ 3 references
	 Weißnackenkolibri (German)  edit
	▶ 2 references
	 Colibrí nuquiblanco (Spanish)  edit
	▶ 2 references
	 colibrí nucablanc (Catalan)  edit
	▶ 2 references
	 kolibřík bělokrký (Czech)  edit
	▶ 2 references
	 Jacobino cuello blanco (Spanish)  edit
	▶ 1 reference

White-necked Jacobin [[Q1084570](#)]

Data about biodiversity

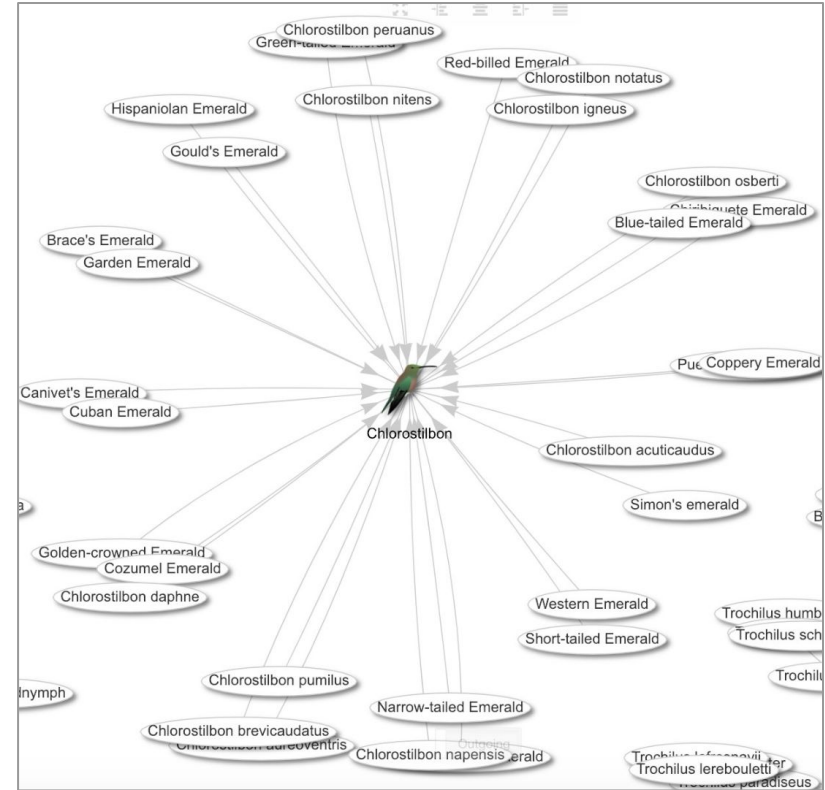
Identifiers

ITIS TSN	555070 ► 1 reference + add value	edit
NCBI Taxonomy ID	304640 ► 1 reference + add value	edit
Freebase ID	/m/05h68s ► 1 reference + add value	edit
Encyclopedia of Life ID	1048777 ► 1 reference + add value	edit

Wikipedia (20 entries) edit	
ast	Florisuga mellivora
bg	Florisuga mellivora
ca	Colibrí de clatell blanc
ceb	Florisuga mellivora
cy	Jacobin cyffredin
de	Weißnackenkolibri
en	White-necked jacobin
es	Florisuga mellivora
eu	Florisuga mellivora
fa	گل‌آشام سفیدگردن
fi	Mesikolibri
fr	Colibri jacobin
hu	Örvös jakobinuskolibri
nl	Witnekkolibrie
pl	Nektareczek błękitny
pnb	چٹے تھلے والی پدی چڑی
ru	Колибри-якобин
sv	Halsbandsjakobin
vi	Florisuga mellivora
war	Florisuga mellivora

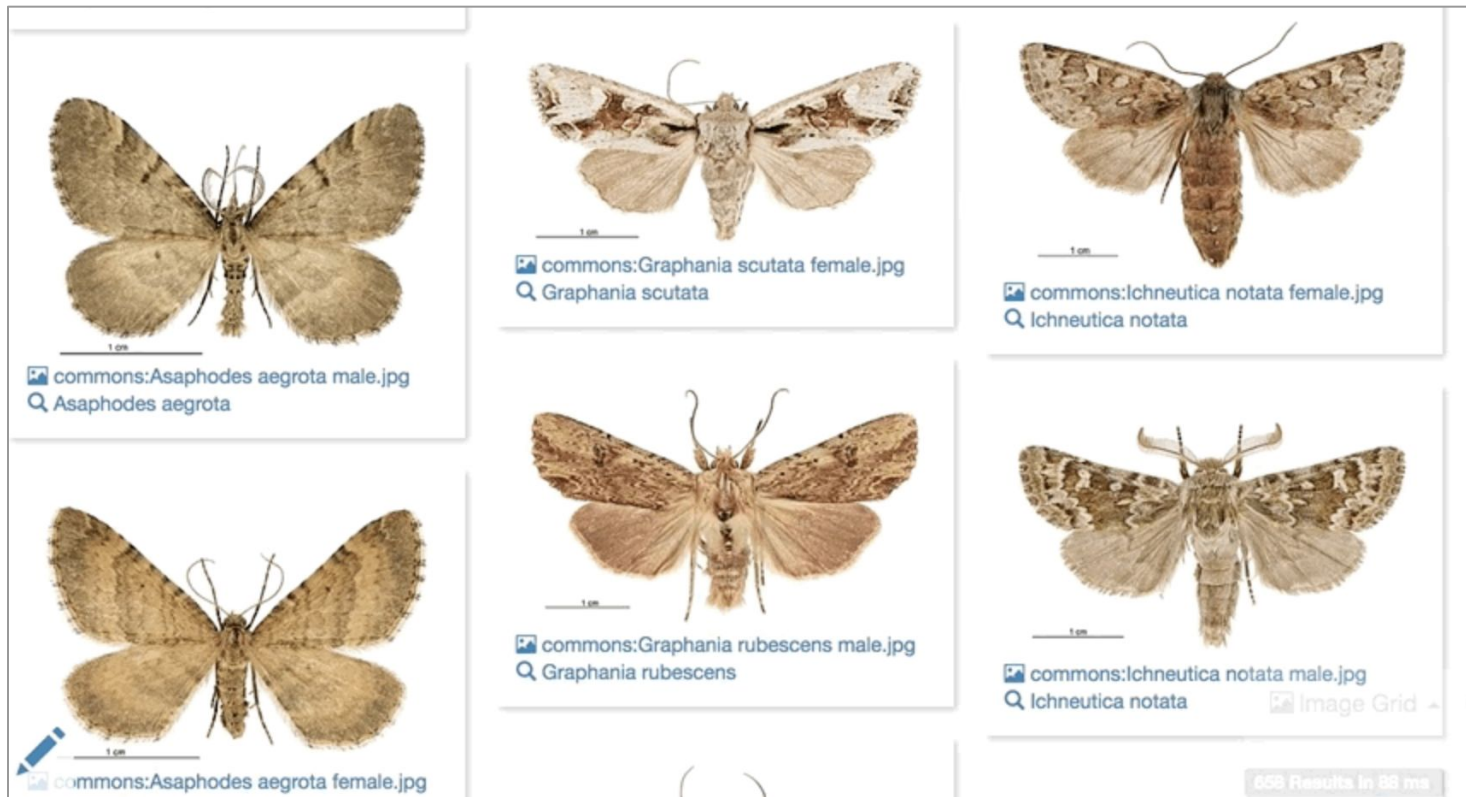
White-necked Jacobin [[Q1084570](#)]

Data about biodiversity



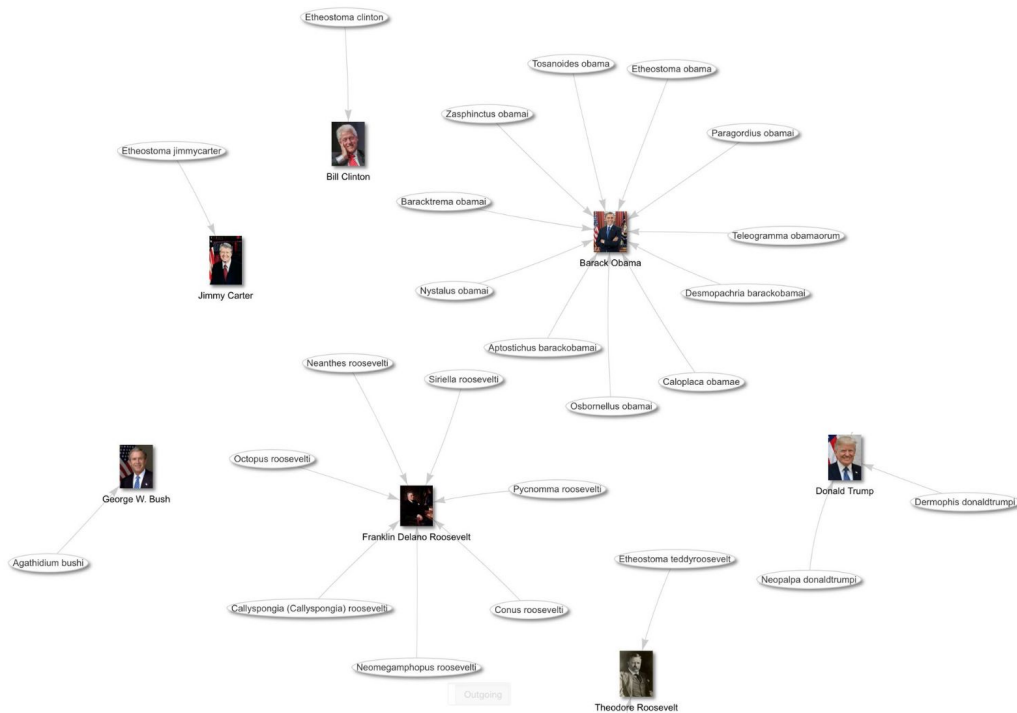
Species of hummingbirds: <http://tinyurl.com/ya543y55>

Data about biodiversity



[[SPARQL query](#)]

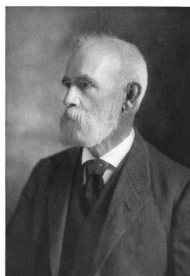
Data about biodiversity



1. Barack Obama: **11**
2. Franklin Delano Roosevelt: **7**
3. Donald Trump, Theodore Roosevelt: **2**
4. George W. Bush, Jimmy Carter, Bill Clinton: **1**

Species named after US presidents: <http://tinyurl.com/y8q3bw7a>

Ornithologists with American citizenship

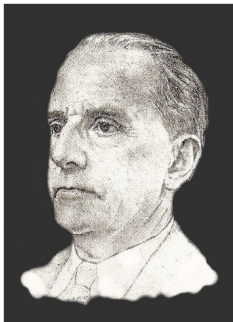


William Brewster
Zimansky, 1916

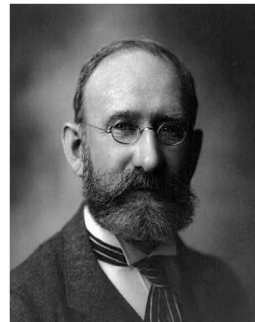
 commons:William Brewster o...
William Brewster



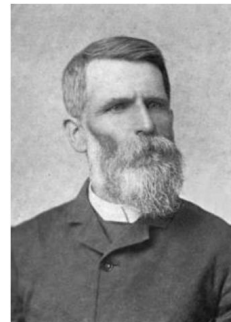
 commons:Herbert Girton Deignan.jpg
Herbert Girton Deignan



 commons:WILLIAM HENRI PHELP...
William H. Phelps, Sr.



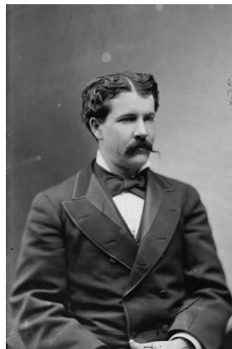
 commons:Henry Wetherbee Henshaw ...
Henry Wetherbee Henshaw



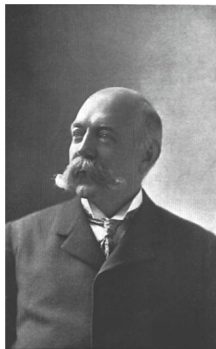
 commons:Joseph Beal Steere.png
Joseph Beal Steere



 commons:W B Barrows.jpg
Walter Bradford Barrows



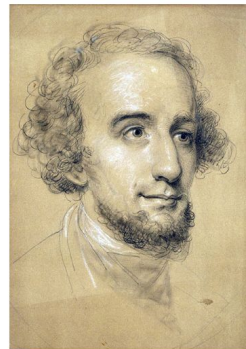
 commons:Joseph H. Acklen.jpg
Joseph H. Acklen



 commons:Sennett George Burrit...
George Burritt Sennett



 commons:George Ord, in Rhoads repr...
George Ord



 commons:Titian Ramsay Peale 1799...
Titian Peale

[[SPARQL query](#)]

STRI on Wikidata

Smithsonian Tropical Research Institute (Q1673980)

research institute based in Panama

 [edit](#)

STRI

▼ [In more languages](#)

Language	Label	Description	Also known as
English	Smithsonian Tropical Research Institute	research institute based in Panama	STRI
German	Smithsonian Tropical Research Institute	No description defined	
Spanish	Instituto Smithsonian de Investigaciones Tropicales	centro de investigaciones en Panamá	STRI Smithsonian Tropical Researc...
French	Smithsonian Tropical Research Institute	institut de recherche au Panama	
Italian	Smithsonian Tropical Research Institute	centro di ricerca nel Panamá	
Latin	Smithsonian Tropical Research Institute	No description defined	
Dutch	Smithsonian Tropical Research Institute	No description defined	

[All entered languages](#)

Scientists affiliated with STRI

Employees and affiliated

Past and present employees, affiliated, and members

Show entries

Search:

Works	Researcher	Description	Orcid
49	Helene Muller-Landau	American ecologist	0000-0002-3526-9021
42	Carlos Jaramillo	paleobotanist and paleontologist	
42	Benjamin L Turner	researcher	0000-0002-6585-0722
38	Karen R. Lips	US herpetologist	0000-0002-2719-1551
16	Rachel Collin	evolutionary biologist	0000-0001-5103-4460
15	Matthew C. Larsen	director, Smithsonian Tropical Research Institute	0000-0001-8157-9621
15	Aaron O'Dea	researcher	0000-0001-5495-4764
10	Blanca Figuerola	researcher	0000-0003-4731-9337
10	Kaoru Kitajima	researcher	0000-0001-6822-8536
8	Michiel van Breugel	researcher	0000-0003-2778-7803

[\[SPARQL query\]](#)

Individual scientists

Helene Muller-Landau (Q57161571)

American ecologist

 [edit](#)

Helene C Muller-Landau | Helene C. Muller-Landau

▼ In more languages

Language	Label	Description	Also known as
English	Helene Muller-Landau	American ecologist	Helene C Muller-Landau Helene C. Muller-Landau
German	Helene Muller-Landau	No description defined	
Spanish	Helene Muller-Landau	ecóloga estadounidense	
French	Helene Muller-Landau	No description defined	
Italian	Helene Muller-Landau	ecologa americana	
Latin	Helene Muller-Landau	No description defined	
Dutch	Helene Muller-Landau	No description defined	

Helene Muller-Landau - <https://tools.wmflabs.org/scholia/author/Q57161571>

Individual scientists

Statements

instance of



human

▼ 0 references

Individual scientists

List of publications

Show entries

Search:

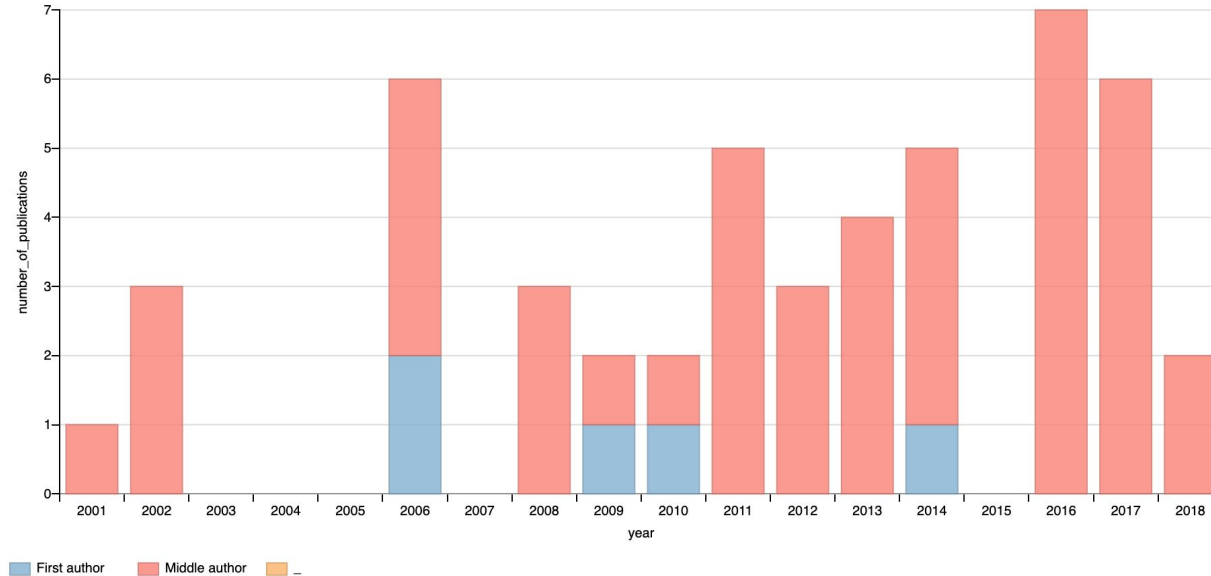
Date	Work	Type	Pages	Venue	Authors
2018-03-02	Resource acquisition and reproductive strategies of tropical forest in response to the El Niño-Southern Oscillation.	scholarly article		Nature Communications	Helene Muller-Landau
2018-02-16	Drivers and mechanisms of tree mortality in moist tropical forests.	scholarly article		New Phytologist	Thomas L Powell, Henrik Hartmann, Chonggang Xu, Helene Muller-Landau, Maurizio Mencuccini, Lara M Kueppers, Paulo M Brando, Charlotte Grossiord
2017-09-18	Vegetation demographics in Earth System Models: A review of progress and priorities.	scholarly article		Global Change Biology	Thomas L Powell, Anna T Trugman, Ashley M. Matheny, Tao Zhang, Chonggang Xu, Marcos Longo, Helene Muller-Landau, Shawn P Serbin
2017-07-18	Drought-induced mortality patterns and rapid biomass recovery in a terra firme forest in the Colombian	scholarly article		Ecology	Helene Muller-Landau

Helene Muller-Landau - <https://tools.wmflabs.org/scholia/author/Q57161571>

Individual scientists

Missing publications? [Search for unlinked works on Wikidata](#) or add new ones with [sourcemd](#).

Number of publications per year

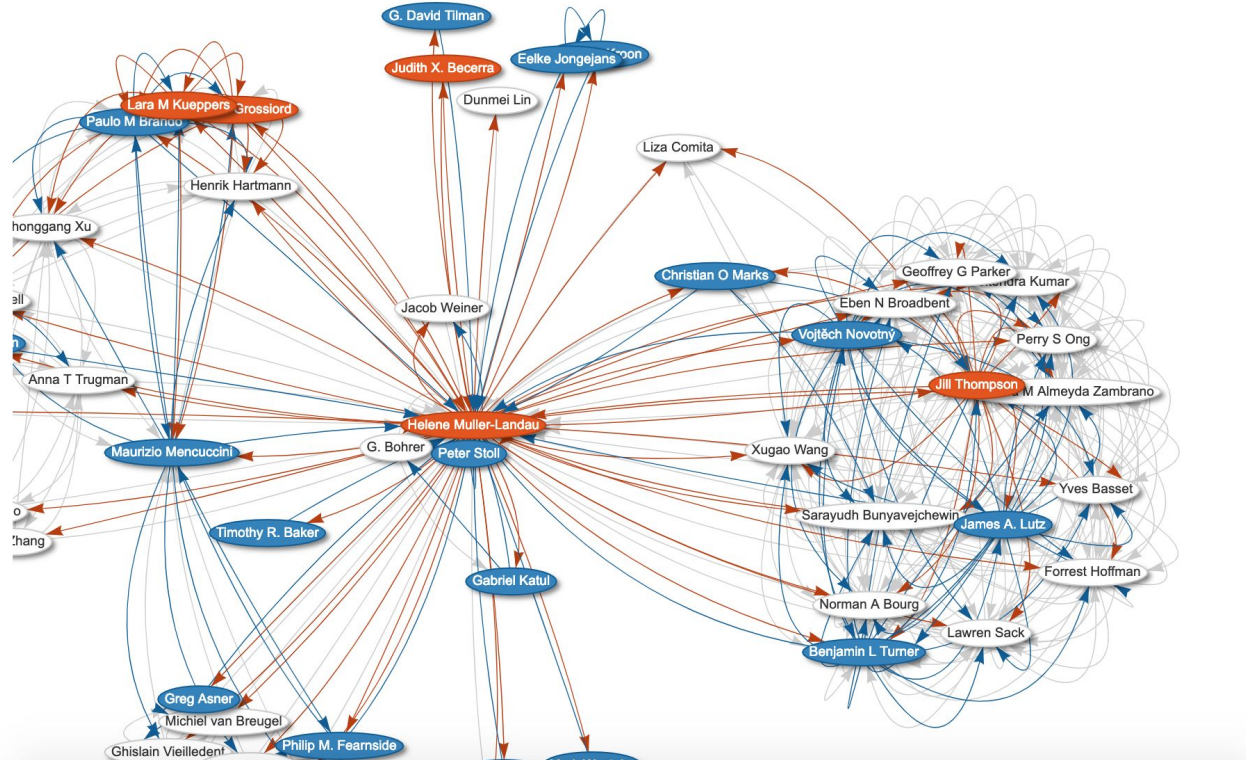


Helene Muller-Landau - <https://tools.wmflabs.org/scholia/author/Q57161571>

Individual scientists



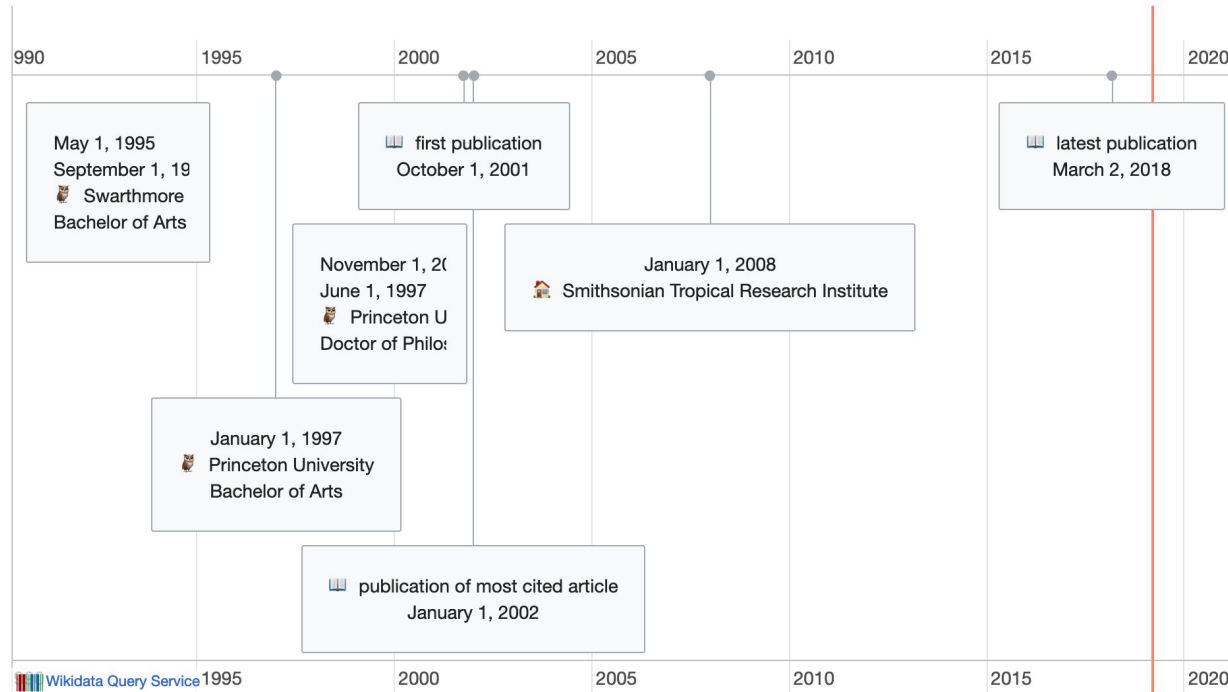
Individual scientists



Helene Muller-Landau - <https://tools.wmflabs.org/scholia/author/Q57161571>

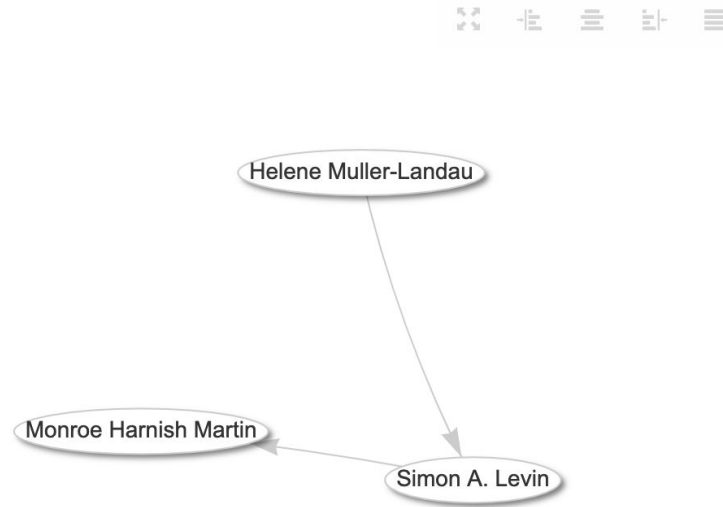
Individual scientists

Timeline



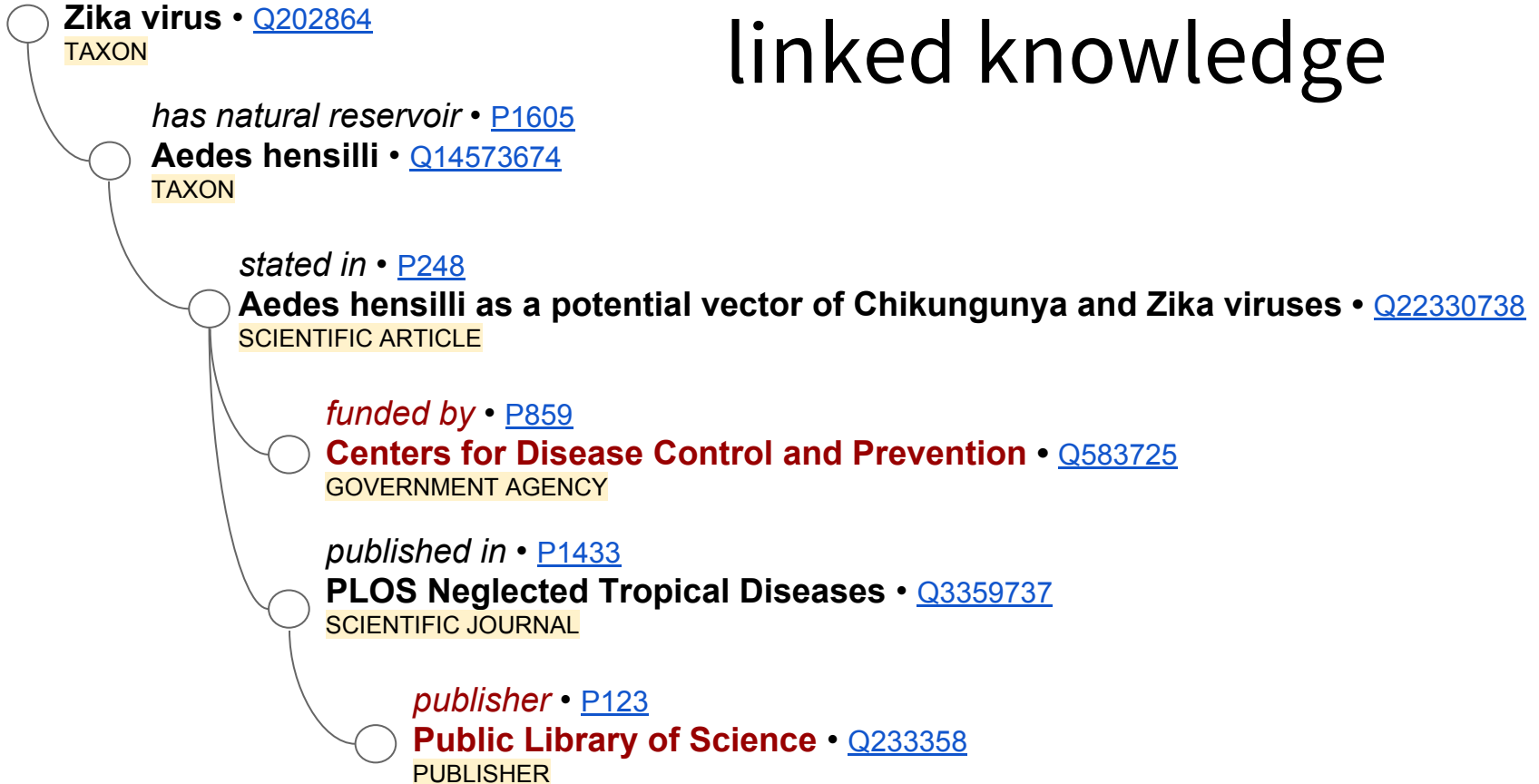
Individual scientists

Academic tree



Helene Muller-Landau - <https://tools.wmflabs.org/scholia/author/Q57161571>

linked knowledge



open linked knowledge
that anyone can edit
and anyone can access



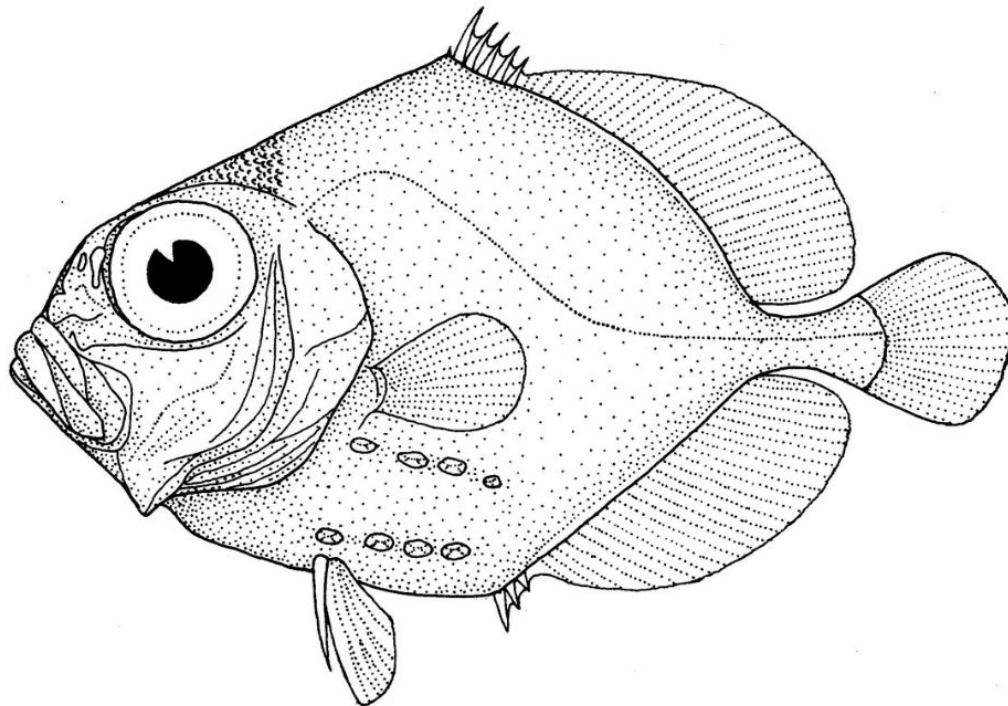


Siobhan @SiobhanLeachman · Jan 16



Sometimes I get demoralised by the amount of paywalled content I come across. But then I get cheered up when I discover fabulously **generous** folk who just make my Wiki world go round. See the drawings by Dr. Tony Ayling.

(Image:CC BY-SA 1.0) [commons.wikimedia.org/wiki/Category:...](https://commons.wikimedia.org/wiki/Category:Tony_Ayling)



best practices to support open curation



- Be generous with your data
- Publish in open access outlets
- Engage with citizen scientists
- Avoid using walled gardens
- Avoid proprietary licenses

Thank you

Dario Taraborelli – @readermeter



D. Taraborelli (2019) *Open curation of knowledge on biodiversity: An introduction to Wikidata*.
doi.org/10.6084/m9.figshare.7780994

Acknowledgments

growth by Fabio Rinaldi [CC BY] from the Noun Project
Toolkit and tools by Todd Quackenbush [CC0] from Unsplash