

The background of the slide is a detailed Renaissance painting of the Tower of Babel. The tower is a massive, multi-tiered structure with intricate architectural details, including arches, windows, and scaffolding. It is built on a hillside overlooking a city and a body of water. In the foreground, a group of people in period clothing are gathered around a large stone block, some appearing to be in the process of building or dismantling it. The sky is filled with clouds, and a large, dark, stormy cloud is visible in the upper left corner.

An Ambitious Wikidata Tutorial

Emw

WikiConference USA

Washington, D.C.

2015-10-10 (Updated 2015-10-13)

Wikidata is a free knowledge base that can be read and edited by humans and machines.

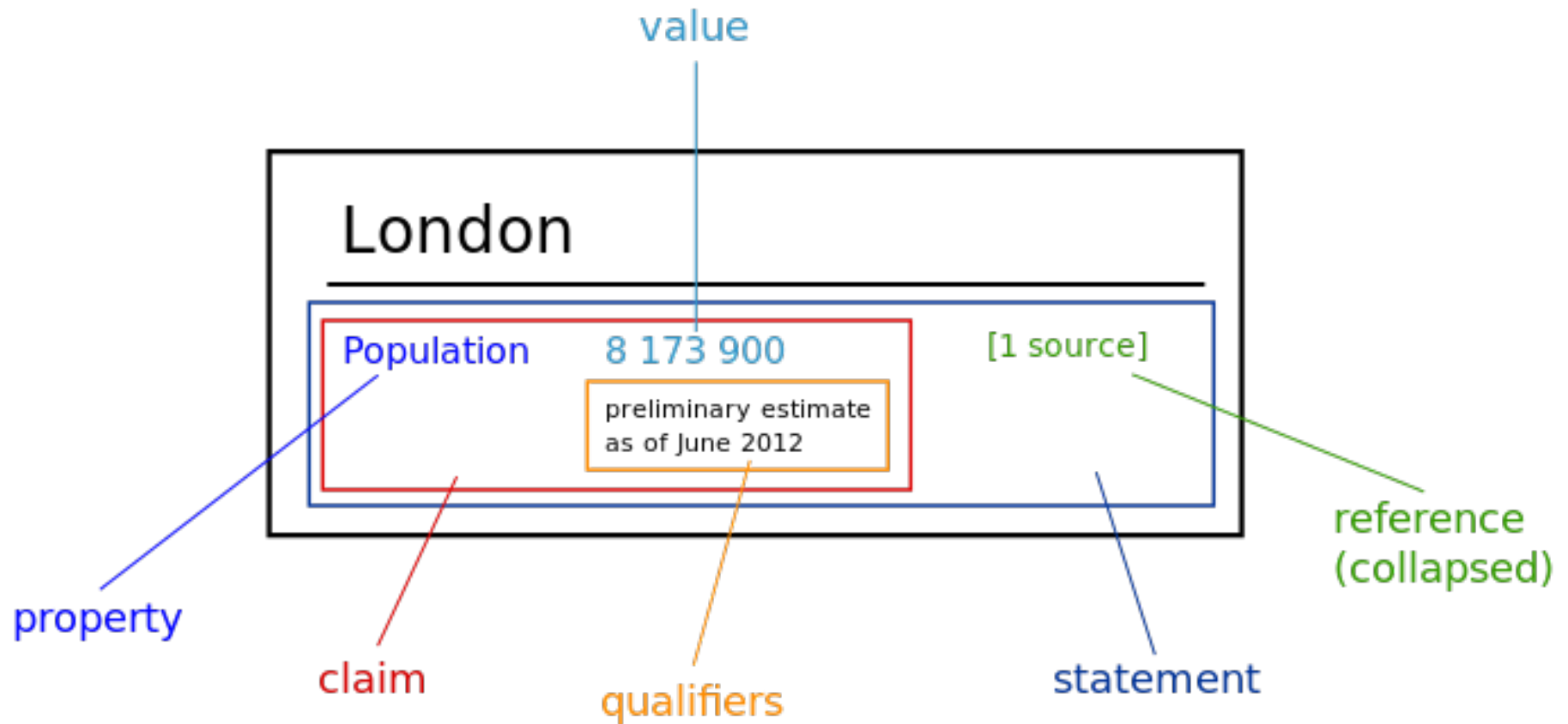
Wikidata's goals

- Centralize interwiki links
- Centralize infoboxes
- Provide an interface for rich queries
- Structure the sum of all human knowledge

What you'll learn from this talk

- Wikidata vocabulary
- How to edit Wikidata
- Where to find things
- Projects built with Wikidata
- Querying with SPARQL, etc.
- Wikidata API programming
- RDF and OWL exports
- Ontologies

Elements of a Wikidata statement



Example: Washington, D.C. (Q61)

population

658,893

point in time


1 July 2014

determination method

estimation

▼ 1 reference

reference URL

<http://www.census.gov/popest/data/cities/totals/2014/SUB-EST2014.html> 

author

United States Census Bureau

Items and properties

- Each item and property has its own page
- Items
 - Represent subjects: Barbara McClintock, Challenger disaster
 - Have identifiers like Q199654, Q921090
 - 14,875,838 items as of 2015-10-05
- Properties
 - Represent attribute names: occupation, cause of
 - Have identifiers like P106, P828
 - 1,805 properties as of 2015-10-05

Statements and claims

- Claims
 - Claims are “triplets”
 - Formally: subject, predicate, object
 - In Wikidata: item, property, value
 - Example: Barbara McClintock, occupation, scientist
- Statements
 - A claim is only part of a statement
 - Statements also include:
 - References
 - Ranks

Qualifiers, ranks, references

- Qualifiers

- Qualifiers are properties used on *claims* rather than items
- “Bethesda **population 56,527 point in time (P585) 1960**”

- Ranks

- Preferred, normal, deprecated
- Useful to mark outdated claims

- References

- Source of claim; provenance
- “... *stated in* (P248) 1960 United States Census”

More on Wikidata vocabulary

<https://www.wikidata.org/wiki/Wikidata:Glossary>

Wikidata link on Wikipedia

Wikipedia articles have a [Wikidata item](#) link in the left navigation panel.

Getting to Wikidata from 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

Article [Talk](#)

[Read](#) [Edit](#) [View history](#)

Barbara McClintock

From Wikipedia, the free encyclopedia

This article is about the American scientist. For the American illustrator, see [Barbara McClintock \(illustrator\)](#).

Barbara McClintock (June 16, 1902 – September 2, 1992) was an American scientist and **cytogeneticist** who was awarded the 1983 **Nobel Prize in Physiology or Medicine**. McClintock received her **PhD** in **botany** from **Cornell University** in 1927. There she started her career as the leader in the development of **maize cytogenetics**, the focus of her research for the rest of her life. From the late 1920s, McClintock studied **chromosomes** and how they change during reproduction in maize. She developed the technique for visualizing maize chromosomes and used microscopic analysis to demonstrate many fundamental genetic ideas. One of those ideas was the notion of **genetic recombination** by **crossing-over** during **meiosis**—a mechanism by which chromosomes exchange information. She produced the first **genetic map** for maize, linking regions of the chromosome to physical traits. She demonstrated the role of the **telomere** and **centromere**, regions of the chromosome that are important in the conservation of **genetic information**. She was recognized among the best in the field, awarded prestigious fellowships, and elected a member of the **National Academy of Sciences** in 1944.

During the 1940s and 1950s, McClintock discovered **transposition** and used it to demonstrate that **genes** are responsible for turning physical characteristics on and off. She developed theories to explain the suppression and expression of genetic information from one generation of maize plants to the next. Due to skepticism of her research and its implications, she stopped publishing her data in 1953.

Later, she made an extensive study of the cytogenetics and **ethnobotany** of maize **races** from South America. McClintock's research became well understood in the 1960s and 1970s, as other scientists confirmed the mechanisms of genetic change and **genetic regulation** that she had demonstrated in her maize research in the 1940s and 1950s. Awards and recognition for her contributions to the field followed, including the Nobel Prize in Physiology or Medicine, awarded to her in 1983 for the discovery of genetic **transposition**; she is the only woman to receive an unshared **Nobel Prize** in that category.^[1]

Contents [hide]

- 1 Early life
- 2 Education and research at Cornell
- 3 University of Missouri
- 4 Cold Spring Harbor

Barbara McClintock



Barbara McClintock shown in her laboratory.

Born	Eleanor McClintock June 16, 1902 Hartford, Connecticut, US
Died	September 2, 1992 (aged 90) Huntington, New York, US
Nationality	American
Fields	Cytogenetics
Institutions	University of Missouri Cold Spring Harbor Laboratory

Wikidata search

Instant search suggests items that have *labels or aliases* matching your keyword.

Search by label



Main Page Discussion Read View source View history

English Create account Log in

Bethes

Bethesdakirche
Wikipedia disambiguation page

Bethesda (Bethesda, Maryland)
unincorporated community in Montgomer...

Bethesda Softworks
American video game publisher

Bethesda
Wikimedia disambiguation page

Bethesda, Ohio

Walter Reed National Military Medic...
tri-service US military medical center in ...

Bethesda Station
Washington DC metro station

more

containing...
Bethes

Welcome to Wikidata
the free knowledge base with 14,879,446 data items that anyone can edit
Introduction • Project Chat • Community Portal • Help

open

multilingual

free

Welcome!
Wikidata is a free linked database that can be read and edited by both humans

Learn about data
New to the wonderful world of data? [Develop and improve your data literacy](#)

Search by alias: “flu” -> influenza



- Main page
- Community portal
- Project chat
- Create a new item
- Item by title
- Recent changes
- Random item
- Help
- Donate

- Print/export
- Create a book
- Download as PDF
- Printable version

- Tools
- What links here
- Related changes
- Special pages
- Permanent link
- Page information

Main Page

Discussion

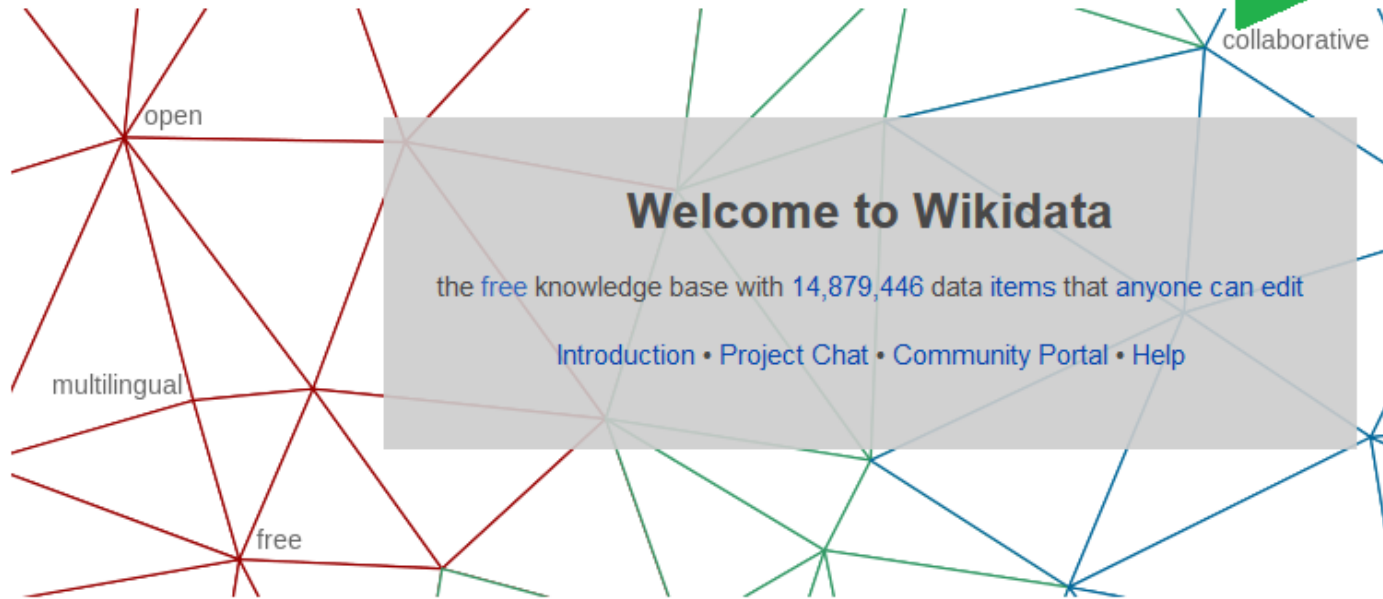
Read

View source

View history

flu

English Create account Log in



- influenza (*flu*)**
infectious disease affecting birds and m...
 - Flu**
Wikipedia disambiguation page
 - zinc finger, MYND-type containing 1...**
human gene
 - fluorine**
chemical element with the atomic numbe...
 - flute**
musical instrument of the woodwind fa...
 - carbon monoxide (*Flue gas*)**
colorless, odorless, and tasteless toxic ...
 - Flurin (*Flurin (first name)*)**
male given name
- more
- containing...
flu

Welcome!

Wikidata is a free linked database that can be read and edited by both humans

Learn about data

New to the wonderful world of data? [Develop and improve your data literacy](#)

Finding properties

- Is there a property for “number of windows”?
- What was the ID of that property, again?
- Search
 - In main site search box, prefix search term with “P:”
 - “P:number of”, “P:occupation”
 - Instant search doesn't work for properties, only items
- Browse
 - https://www.wikidata.org/wiki/Wikidata:List_of_properties
 - ^ bookmark this!

Let's edit Wikidata.

Barbara McClintock

<https://www.wikidata.org/wiki/Q199654>



- employer (P108):
Cold Spring Harbor Laboratory
start time (P580): December 1941
- member of (P463):
National Academy of Sciences
start time (P580): 1944
American Academy of Arts and Sciences
start time: 1959
Royal Society
start time: 1989
- award received (P166):
Nobel Prize in Physiology or Medicine
for work (P1686): mobile genetic elements
National Medal of Science
point in time (P585): 1971
- birth name (P1477): Eleanor McClintock

Area? Height? GDP per capita?

- Quantities with units recently made possible!
- area (P2046)
- height (P2048)
- mass (P2067)
- cost (P2130)
- GDP per capita (P2132)
- total debt (P2133)

Quantities: Lots of low-hanging fruit

- Not yet on Wikidata:
 - Area of Washington, D.C.
 - Height of Abraham Lincoln
 - Height of United States Capitol
 - Length of Mississippi River
 - GDP per capita of the United States of America

^ Add these!

Built on Wikidata

- **Histropedia**

300,000 timelines and 1.5 million events

<http://histropedia.com/timeline/1fr22b0p8s/Empires>

- **Reasonator**

Wikidata knowledge tailored for readers

J.S. Bach: <http://tools.wmflabs.org/reasonator/?q=Q1339>

- **Gene Wiki**

Every human gene is now on Wikidata

<http://blog.wikimedia.de/2014/10/22/establishing-wikidata-as-the-central-hub-for-linked-open-life-science-data/>

(More info today in Open Biomedical Knowledge session at 2:15 PM)

Histropedia

v1.7

Follow us [f](#) [in](#) [g+](#) [t](#)

[Timeline](#) [About](#) [Contact](#) [How it works](#) [Join Community](#)

[Feedback](#)



[Top 20 Timelines](#)

Start searching for your topic...

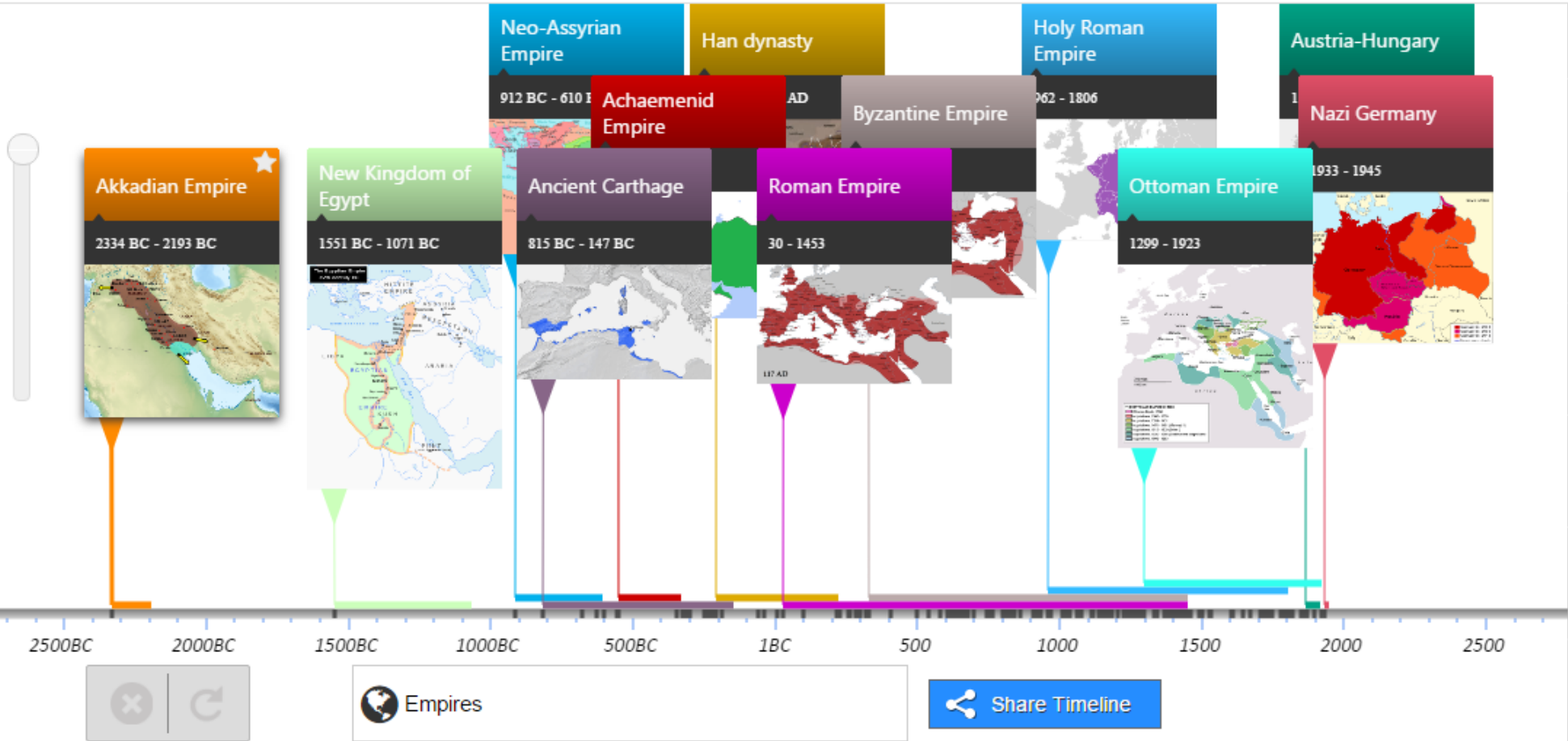


Compact



Save

Sign In



<http://histropedia.com/timeline/1fr22b0p8s/Empires>

Tools

- Wikidata API
 - <https://www.wikidata.org/w/api.php>
- Querying:
 - **Autolist:**
<http://tools.wmflabs.org/autolist/autolist1.html>
 - **Wikidata Query Service (new!):**
<https://query.wikidata.org>
- Software framework: **Wikidata Toolkit**
 - https://www.mediawiki.org/wiki/Wikidata_Toolkit
 - <https://github.com/Wikidata/Wikidata-Toolkit>

Wikidata API

Quick Python demo

Querying in Wikidata

List of politicians who died of cancer

Pseudo-query:

occupation: politician AND cause of death: cancer

occupation: P106

politician: Q82955

cause of death: P509

cancer: Q12078

Wikidata query in Autolist:

claim[106:82955] AND claim[509:12078]

http://tools.wmflabs.org/autolist/autolist1.html?q=claim[106:82955]%20AND%20claim[509:12078]

FOR EDITING WIKIDATA, please use this tool's successor, [AutoList 2!](#)

This tool can create a live list of items based on a [Wikidata Query](#). Check the [API documentation](#) to construct a query. Check out [some lists](#) the community finds useful, or add your own!

Query

Category on .

Show query or category subset of both superset of both


Language : [Permalink](#) [Embed](#) [Download](#) (Share on: [Twitter](#) | [Email](#))

Properties : (comma-separated, numerical values)

I have given OAuth authorisation to [WiDaR](#)

Item list (56 items)

1-50 | [51-56](#)

#	Item	Description	Wikipedia(s)
1	Hugo Chávez	52nd President of Venezuela	en.wikipedia  ace , af , als , am , ang , an , ar , arwikinews , arz , ast , ay , az , azwikiquote , bat_smg , ba , bcl , be_x_old , be , bg , bn , br , bs , ca , cawikinews , cawikiquote , cbk_zam , ckb , commons , cs , cswikiquote , cy ,

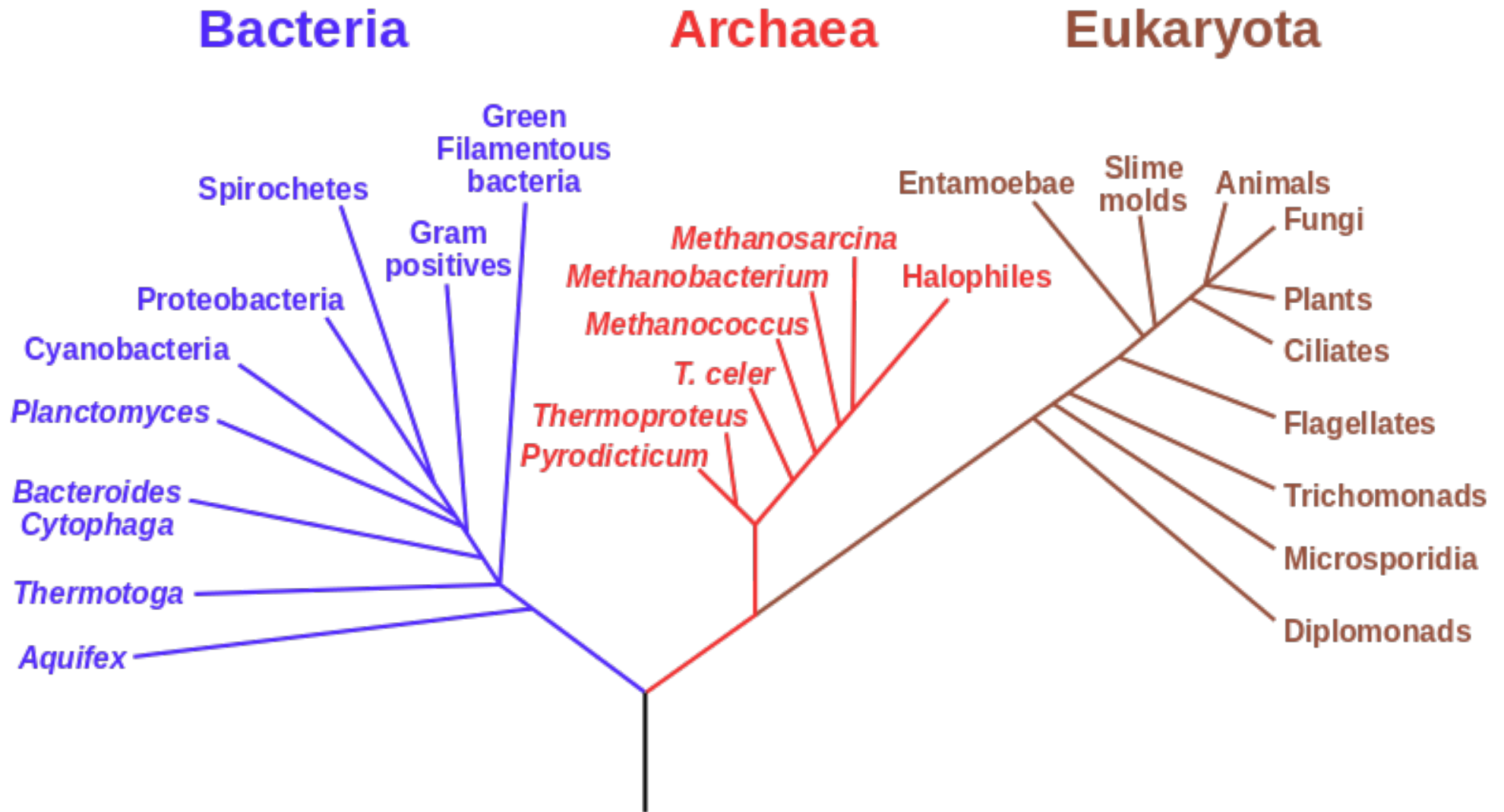
Only 56 politicians have died of cancer?

Seems rather low...

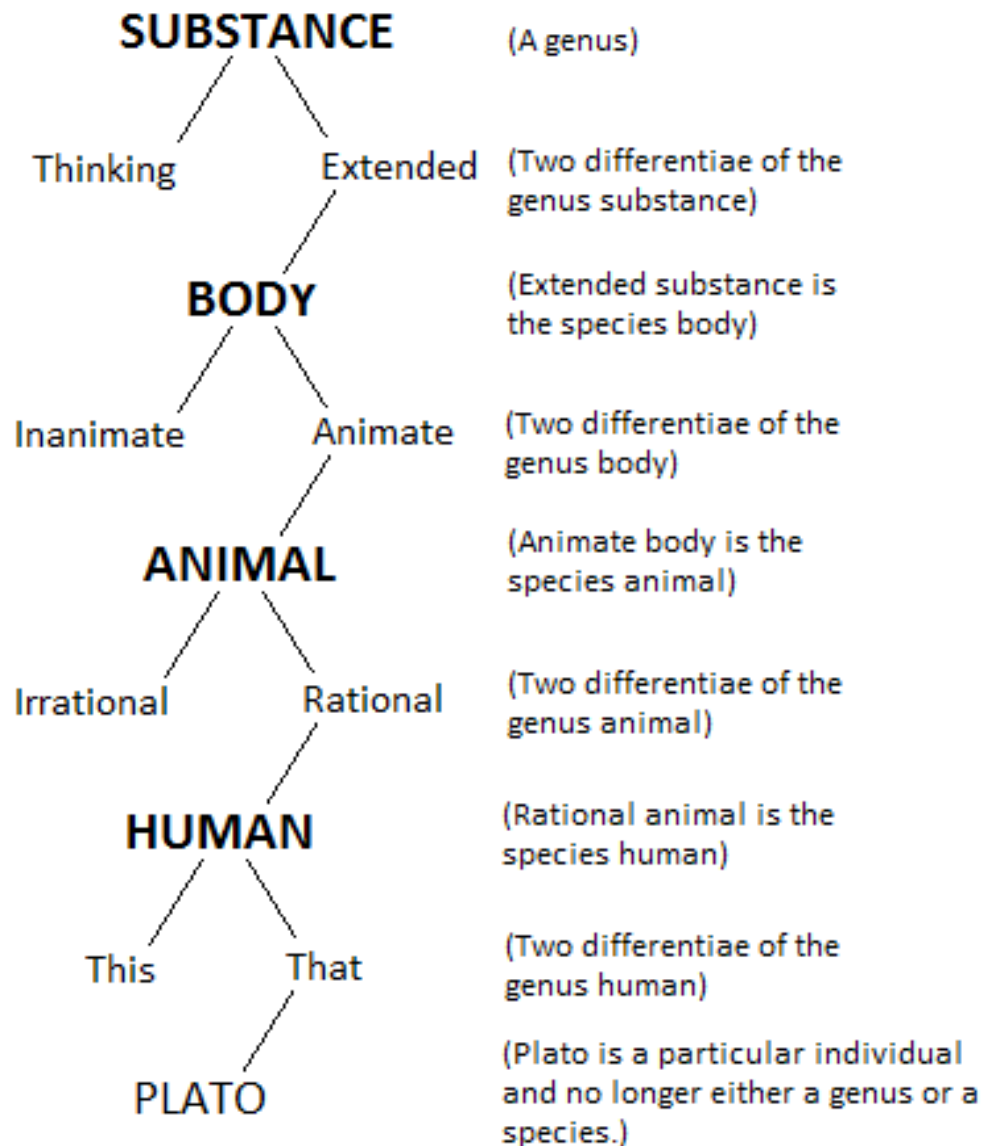
Classification on Wikidata

- Taxonomy of knowledge
- Enables powerful inference, novel applications
- Interesting philosophical, design, and engineering issues

Phylogenetic Tree of Life



Tree of Porphyry



Classes and instances

- Plato *is a* human *is a* animal
- Plato *instance of* human *subclass of* animal
- Instance: concrete object, individual
- Class: abstract object

Classification on Wikidata

- instance of (P31)
 - `rdf:type` in RDF and OWL
 - Most popular Wikidata property
- subclass of (P279)
 - “all instances of A are also instances of B”
 - `rdfs:subClassOf` in RDF and OWL

Examples

- USS Nimitz *instance of* Nimitz-class aircraft carrier
Nimitz-class aircraft carrier *subclass of* aircraft carrier
- 2012 Cannes Film Festival *instance of* Cannes Film Festival
Cannes Film Festival *subclass of* film festival
- an individual charm quark *instance of* charm quark
charm quark *subclass of* quark

^ Many “leaf nodes” in Wikidata's taxonomic hierarchy are not instances.
(There are no items about individual quarks on Wikidata!)

https://www.wikidata.org/wiki/Help:Basic_membership_properties

Bad smells

Item has many *instance of* or *subclass of* claims

Items typically satisfy a **huge** number of *instance of* claims:

- Fido *instance of* dog
- Fido *instance of* English Pointer
- Fido *instance of* faithful animal
- ...

Solution: use one class for *instance of*, put other class knowledge into normal properties

- Fido *instance of* dog
- Fido *breed*: English Pointer
- Fido *known for*: faithfulness
- ...

Bad smells

subclass of claim that is nonsensical when interpreted as “All instances of A are also instances of B”

Example:

dog subclass of pet

But not all dogs are pets!

feral dog *subclass of* dog true

feral dog *subclass of* pet false

∴ dog *subclass of* pet false

Solution: put “pet” knowledge about dogs into claim that does not apply to all instances of dog. E.g. “dog *has role* pet”. (*Has role* would not be transitive.)

Classification on Wikidata

- Last but not least: part of (P361)
 - Third basic membership property
 - Top-level “part-whole” relation
- *subclass of* and *part of* are both transitive; *instance of* is not transitive
- Transitive relation:
 - A subclass of B*
 - B subclass of C*
 - ∴ A subclass of C*

https://www.wikidata.org/wiki/Help:Basic_membership_properties

subclass of (P279)
enables machines to infer
conceptual hierarchy

Recall

Query reports that only 56 politicians have died of cancer

Problem

Only matches the precise claim “cause of death: cancer”

Omits results that have:

- cause of death: lung cancer
- cause of death: lymphoma
- cause of death: leukemia

Solution

Include causes of death that are a *subclass of* cancer

- In Autolist: <http://tinyurl.com/ovgjqd8>
- Also possible in SPARQL in new Wikidata Query Service

SPARQL

- SPARQL: semantic query language for databases
- Wikidata recently added official support
 - Query UI: <https://query.wikidata.org>
 - Examples:
https://www.mediawiki.org/wiki/Wikibase/Indexing/SPARQL_Query_Examples

Example SPARQL query

List of politicians who died of cancer:

```
PREFIX wd: <http://www.wikidata.org/entity/>
```

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

```
SELECT ?politician ?cause ?politician_label ?cause_of_death_label WHERE {  
  ?politician wdt:P106 wd:Q82955 .    # find items that have "occupation (P106): politician (Q82955)"  
  ?politician wdt:P509 ?cause .      # with a P509 (cause of death) claim  
  ?cause wdt:P279* wd:Q12078 .      # ... where the cause is a subclass of (P279*) cancer (Q12078)  
  # ?politician wdt:P39 wd:Q11696 .  # Uncomment this line to include only U.S. Presidents  
  
  OPTIONAL {?politician rdfs:label ?politician_label filter (lang(?politician_label) = "en") .}  
  OPTIONAL {?cause rdfs:label ?cause_of_death_label filter (lang(?cause_of_death_label) = "en").}  
}  
ORDER BY ASC (?politician)
```

Live demo: <http://tinyurl.com/nh7jc2p>

http://tinyurl.com/nh7jc2p

```
1 PREFIX wd: <http://www.wikidata.org/entity/>
2 PREFIX wdt: <http://www.wikidata.org/prop/direct/>
3 |
4 SELECT ?politician ?cause ?politician_label ?cause_of_death_label WHERE {
5   ?politician wdt:P106 wd:Q82955 .      # find items that have "occupation (P106): politician (Q82955)"
6   ?politician wdt:P509 ?cause .      # with a P509 (cause of death) claim
7   ?cause wdt:P279* wd:Q12078 .      # ... where the cause is a subclass of (P279*) cancer (Q12078)
8   # ?politician wdt:P39 wd:Q11696 .  # Uncomment this line to include only U.S. Presidents
9
10  OPTIONAL {?politician rdfs:label ?politician_label filter (lang(?politician_label) = "en") .}
11  OPTIONAL {?cause rdfs:label ?cause_of_death_label filter (lang(?cause_of_death_label) = "en").}
12 }
13 ORDER BY ASC (?politician)
```

Execute

Clear

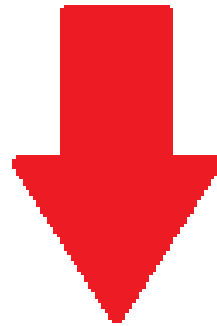
Add prefixes

Total results: 595, duration: 1065 ms. Click on * to explore.

[Generate short URL](#)

politician	cause	politician_label	cause_of_death_label
wd:Q1027427 *	wd:Q189588 *	John R. Fellows	stomach cancer
wd:Q1028400 *	wd:Q3242950 *	Károly Grósz	kidney cancer
wd:Q10320767 *	wd:Q189588 *	Luiz Gushiken	stomach cancer
wd:Q10376143 *	wd:Q47912 *	Sérgio Guerra	lung cancer
wd:Q1064774 *	wd:Q47912 *	Charles Hayes	lung cancer
wd:Q10664 *	wd:Q5526839 *	Neville Chamberlain	gastrointestinal cancer

Be sure to add an asterisk (*) to get the subclass tree!



?cause wdt:P279* wd:Q12078 .

Autolist vs. Wikidata Query Service

	Autolist	Wikidata Query Service
URL	https://tools.wmflabs.org/autolist/	https://query.wikidata.org
Syntax	WDQ (custom, but succinct)	SPARQL (W3C standard)
Support status	Unofficial	Official, beta
Release date	2013-09-17	2015-09-07
Developer	Magnus Manske	Stas Malyshev
Source code	https://bitbucket.org/magnusmanske/wikidataquery	https://github.com/wikimedia/wikidata-query-rdf/
License	GPL 2+	Apache 2.0
Technology	C++	Java, Blazegraph

How to: Explore RDF/OWL dumps locally

- Get the most recent dumps:
<http://tools.wmflabs.org/wikidata-exports/rdf/>
- Small, interesting: wikidata-taxonomy.nt.gz
- Download and install Protege:
<http://protege.stanford.edu/>

Wikidata RDF dumps 20150928

This page provides RDF dump files generated from the [Wikidata dump of 2015-09-28](#) (note that this dumpfile might be deleted in the future). All dump files have been generated using [Wikidata Toolkit](#).

Complete data dumps

- [wikidata-terms.nt.gz](#) (250,042,801 triples, 1.9GB)
RDF dump of all item labels, descriptions, and aliases in all languages.
- [wikidata-properties.nt.gz](#) (116,327 triples, 2.0MB)
RDF dump of all Wikidata property definitions, including datatypes, labels, descriptions, and aliases.
- [wikidata-statements.nt.gz](#) (361,343,257 triples, 4.2GB)
RDF dump of all statements, complete with references and qualifiers.
- [wikidata-sitelinks.nt.gz](#) (153,235,561 triples, 1.1GB)
RDF dump of all site link information. This includes all links from Wikidata to Wikipedia articles as well as to other MediaWiki project sites.

Simplified and derived dumps

- [wikidata-simple-statements.nt.gz](#) (104,589,127 triples, 811.1MB)
RDF dump with simplified versions of many statements, using just one triple per statement. References are omitted. Statements with qualifiers are not included.
- [wikidata-taxonomy.nt.gz](#) (697,753 triples, 3.0MB)
OWL/RDF dump the class hierarchy of Wikidata. Statements of property "subclass of" (P279) that have no qualifiers are exported using `rdfs:subClassOf`. All items that are used like classes in "subclass of" (P279) or "instance of" (P31) are declared as OWL classes.
- [wikidata-instances.nt.gz](#) (14,347,680 triples, 62.1MB)
OWL/RDF dump of all class membership information in Wikidata. Statements of property "instance of" (P31) that have no qualifiers are exported using `rdf:type`. Relevant class declarations are found in [wikidata-taxonomy.nt.gz](#).
- [wikidata-property-taxonomy.nt.gz](#) (1,854 triples, 6.8KB)

Protege

- <http://protege.stanford.edu/>
- The browser of the Semantic Web
- Good for small- to medium-sized ontologies

A free, open-source ontology editor and framework for building intelligent systems

Protégé is supported by a strong community of academic, government, and corporate users, who use Protégé to build knowledge-based solutions in areas as diverse as biomedicine, e-commerce, and organizational modeling.

[DOWNLOAD NOW](#)[USE WEBPROTÉGÉ](#)

TRUSTED BY OVER **244,817** USERS



Querying cancer types in wikidata-taxonomy.nt.gz

The screenshot shows the Protege software interface. The title bar reads "OntologyID(Anonymous-2) : [C:\Users\Eric\Desktop\wikipedia\wikidata\an_ambitious_wikidata_tutorial\src\wikidata-taxonomy.nt]". The menu bar includes File, Edit, View, Reasoner, Tools, Refactor, Window, and Help. Below the menu is a toolbar with navigation arrows and a dropdown menu showing "OntologyID(Anonymous-2)". To the right is a search box labeled "Search for entity".

The main workspace contains several tabs: Annotation Properties, Individuals by class, OWLViz, DL Query, OntoGraf, Ontology Differences, and SPARQL Query. Below these is a row of active ontology views: Active Ontology, Entities, Classes, Object Properties, and Data Properties.

The SPARQL query editor shows the following query:

```
SPARQL query:
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?cancer_type WHERE {
    ?cancer_type rdfs:subClassOf* <http://www.wikidata.org/entity/Q12078> .
}
```

The results pane shows a table with the following data:

cancer_type
Q12078
Q7315926
Q2509220
Q20751413
Q589612
Q5275615
Q2110267
Q20878803

At the bottom of the interface is an "Execute" button and a status bar with the text "To use the reasoner click Reasoner" and two checkboxes: "Start reasoner" (unchecked) and "Show Inferences" (checked).

Open questions: Modeling causes

What caused or causes:

- The Space Shuttle *Challenger* explosion?
- The dinosaurs to die?
- Malaria? Cancer?
- The American Civil War?

Causation on Wikidata

- https://www.wikidata.org/wiki/Help:Modeling_causes
- *has cause* (P828) (alias underlying cause): thing that ultimately resulted in the effect
- *has immediate cause* (P1478): thing that proximately resulted in the effect
- *has contributing factor* (P1479): thing that significantly influenced the effect, but did not directly cause it

What caused the American Civil War?

American Civil War (Q8676)

has cause:

- slavery in the United States (Q118382) (preferred rank)
- states' rights (Q48527) (deprecated rank)

...

has immediate cause:

- Battle of Fort Sumter (Q543165)
- United States presidential election, 1860 (Q698842)

...

has contributing factor:

- caning of Charles Sumner (Q5032419)
- Dred Scott v. Sandford (Q690462)
- Bleeding Kansas (Q331377)
- Uncle Tom's Cabin (Q2222)

...

Thank you!

<https://www.wikidata.org/wiki/User:Emw>