Scaling Wikidata Query Service

unlimited access to all the world's
knowledge for everyone is hard

WIKIMANIA
SINGAPORE

Lydia Pintscher - @nightrose

Wikidata Query Service?!

WIKIMANIA
SINGAPORE

The basics

The Wikidata Query Service is…

- A critical part of Wikidata
- Querying relations that unlock the true power of Wikidata
- Running Blazegraph

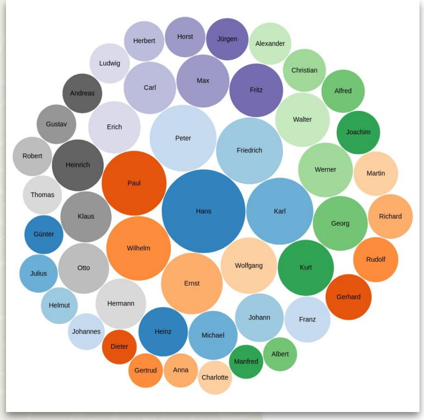
WIKIMANIA
SINGAPORE

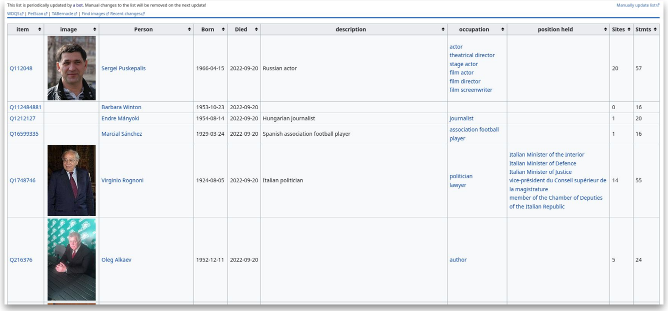
Who uses the Query Service?

WIKIMANIA
SINGAPORE

Wikidata Editors

- Understand and maintain parts of Wikidata
- Advocacy work and workshops
- Show the world what they worked on





This list is periodically updated by a bot. Manual changes to the list will be removed on the next update!

Manually update list

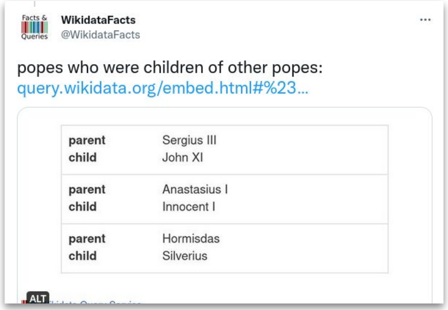
WDQS | PetScan | TABernacle | Find images Recent changes

item	image	Person	Born	Died	description	occupation	position held	Sites	Stmts
Q112048		Sergei Puskepalis	1966-04-15	2022-09-20	Russian actor	actor theatrical director stage actor film actor film director film screenwriter		20	57
Q112484881		Barbara Winton	1953-10-23	2022-09-20				0	16
Q1212127		Endre Mányoki	1954-08-14	2022-09-20	Hungarian journalist	journalist		1	20
Q16599335		Marcial Sánchez	1929-03-24	2022-09-20	Spanish association football player	association football player		1	16
Q1748746		Virginio Rognoni	1924-08-05	2022-09-20	Italian politician	politician lawyer	Italian Minister of the Interior Italian Minister of Defence Italian Minister of Justice vice-président du Conseil supérieur de la magistrature member of the Chamber of Deputies of the Italian Republic	14	55
Q216376		Oleg Alkaev	1952-12-11	2022-09-20		author		5	24

WIKIMANIA
SINGAPORE

Knowledge seekers/sharers

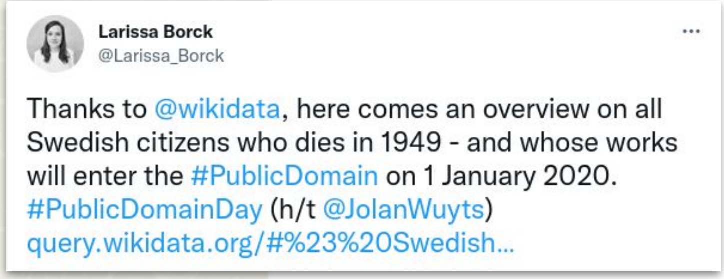
- Use queries to satisfy curiosity
- Use queries to share something curious with the world (e.g. via social networks)



WikidataFacts
@WikidataFacts

popes who were children of other popes:
query.wikidata.org/embed.html#%23...

parent	Sergius III
child	John XI
parent	Anastasius I
child	Innocent I
parent	Hormisdas
child	Silverius

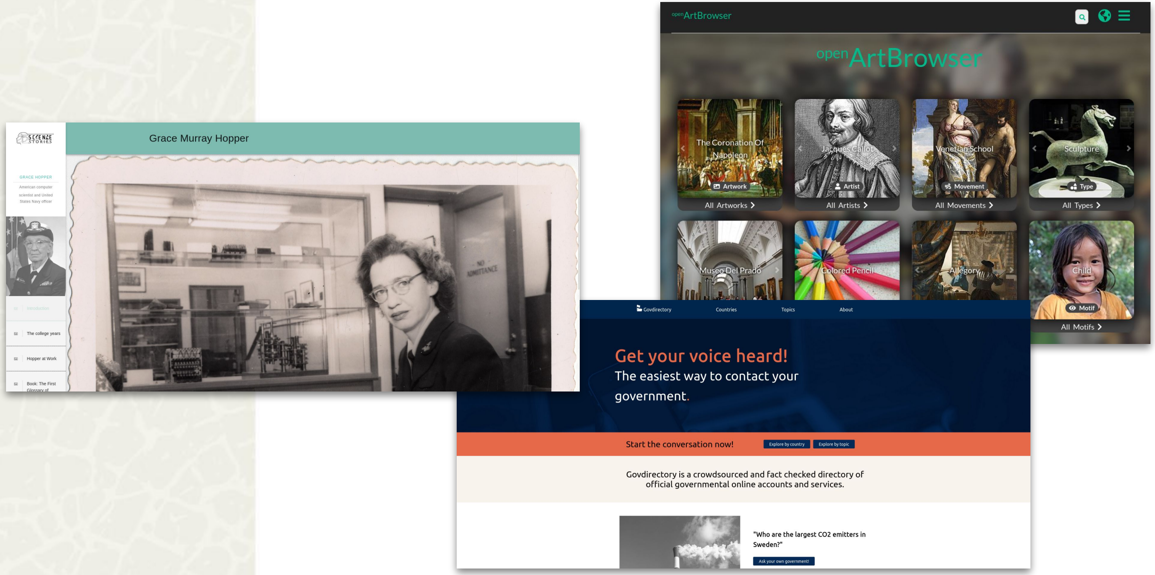


Larissa Borck
@Larissa_Borck

Thanks to @wikidata, here comes an overview on all Swedish citizens who dies in 1949 - and whose works will enter the #PublicDomain on 1 January 2020. #PublicDomainDay (h/t @JolanWuyts) query.wikidata.org/#%23%20Swedish...

Small and medium sized re-users

- Use queries to power their applications and services



Grace Murray Hopper

GRACE HOPPER

American computer scientist and United States Navy officer

Get your voice heard!

The easiest way to contact your government.

Start the conversation now!

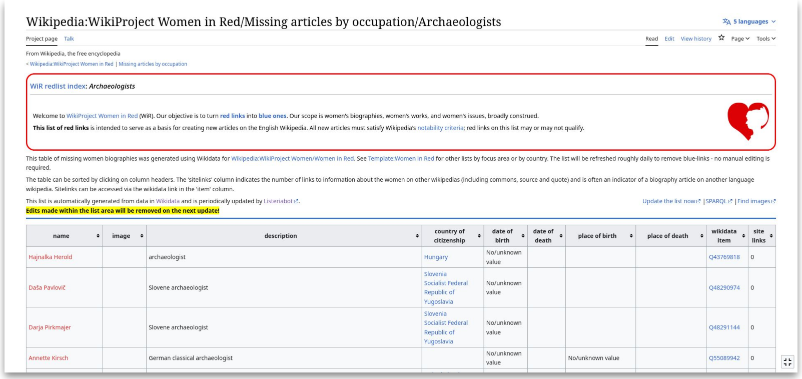
Govdirectory is a crowdsourced and fact checked directory of official governmental online accounts and services.

"Who are the largest CO2 emitters in Sweden?"

WIKIMANIA SINGAPORE

Wikimedia projects

- Structure work in wiki projects and power campaigns, especially around missing content (e.g. Women in Red)
- Better understand the content of the project



Wikipedia:WikiProject Women in Red/Missing articles by occupation/Archaeologists

From Wikipedia, the free encyclopedia

< Wikipedia:WikiProject Women in Red | Missing articles by occupation

WiR redlist index: *Archaeologists*

Welcome to WikiProject Women in Red (WiR). Our objective is to turn **red links** into **blue ones**. Our scope is women's biographies, women's works, and women's issues, broadly construed.

This list of red links is intended to serve as a basis for creating new articles on the English Wikipedia. All new articles must satisfy Wikipedia's notability criteria; red links on this list may or may not qualify.

This table of missing women biographies was generated using Wikidata for Wikipedia:WikiProject Women/Women in Red. See Template:Women in Red for other lists by focus area or by country. The list will be refreshed roughly daily to remove blue-links - no manual editing is required.

The table can be sorted by clicking on column headers. The 'sitelinks' column indicates the number of links to information about the women on other wikipedias (including commons, source and quote) and is often an indicator of a biography article on another language wikipedia. Sitelinks can be accessed via the wikidata link in the 'item' column.

This list is automatically generated from data in Wikidata and is periodically updated by Listeriabot.

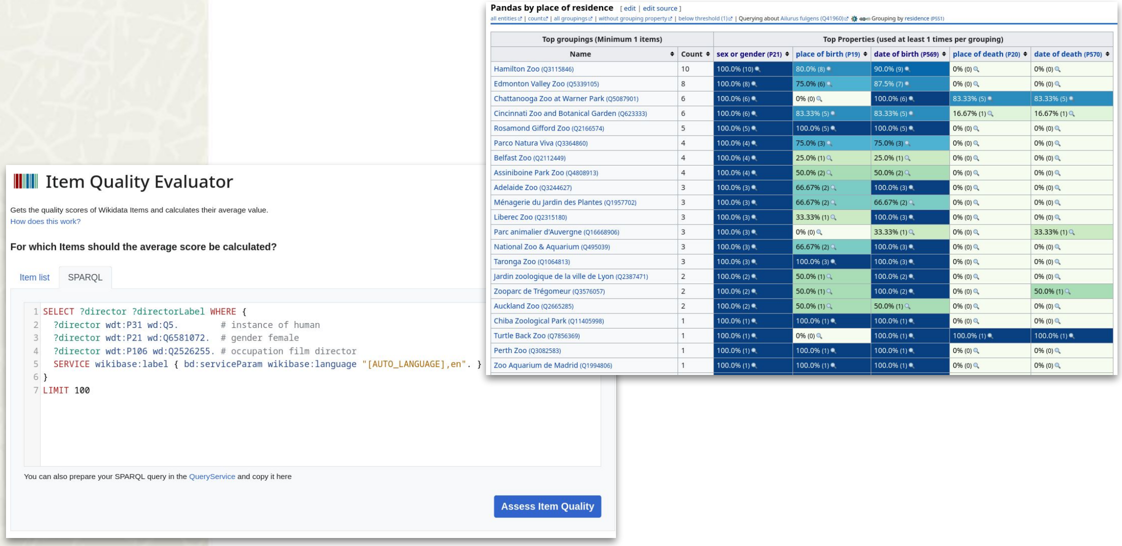
Update the list now | SPARQL | Find images

Edits made within the list area will be removed on the next update!

name	image	description	country of citizenship	date of birth	date of death	place of birth	place of death	wikidata item	site links
Hajnalka Herold		archaeologist	Hungary	No/unknown value				Q43769818	0
Daša Pavlovič		Slovene archaeologist	Slovenia Socialist Federal Republic of Yugoslavia	No/unknown value				Q48290974	0
Darja Pirkmajer		Slovene archaeologist	Slovenia Socialist Federal Republic of Yugoslavia	No/unknown value				Q48291144	0
Annette Kirsch		German classical archaeologist		No/unknown value		No/unknown value		Q55089942	0

Wikimedia development teams and tool builders

- Queries power tools (Listeria, Item Quality Evaluator, Integraality), often by getting lists of Items



Item Quality Evaluator

Gets the quality scores of Wikidata Items and calculates their average value.

How does this work?

For which Items should the average score be calculated?

```
SELECT ?director ?directorLabel WHERE {
  ?director wdt:P31 wd:Q5.         # instance of human
  ?director wdt:P21 wd:Q6581072.  # gender female
  ?director wdt:P106 wd:Q2526255. # occupation film director
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
}
LIMIT 100
```

You can also prepare your SPARQL query in the QueryService and copy it here

Pandas by place of residence [edit | edit source]

all entities | count | all groupings | without grouping property | below threshold (1) | Querying about Ailurus fulgens (Q41960) Grouping by residence (P551)

Top groupings (Minimum 1 items)		Top Properties (used at least 1 times per grouping)				
Name	Count	sex or gender (P21)	place of birth (P19)	date of birth (P569)	place of death (P20)	date of death (P570)
Hamilton Zoo (Q3115846)	10	100.0% (10)	80.0% (8)	90.0% (9)	0% (0)	0% (0)
Edmonton Valley Zoo (Q5339105)	8	100.0% (8)	75.0% (6)	87.5% (7)	0% (0)	0% (0)
Chattanooga Zoo at Warner Park (Q5087901)	6	100.0% (6)	0% (0)	100.0% (6)	83.33% (5)	83.33% (5)
Cincinnati Zoo and Botanical Garden (Q623333)	6	100.0% (6)	83.33% (5)	83.33% (5)	16.67% (1)	16.67% (1)
Rosamond Gifford Zoo (Q2166574)	5	100.0% (5)	100.0% (5)	100.0% (5)	0% (0)	0% (0)
Parco Natura Viva (Q3364860)	4	100.0% (4)	75.0% (3)	75.0% (3)	0% (0)	0% (0)
Belfast Zoo (Q2112449)	4	100.0% (4)	25.0% (1)	25.0% (1)	0% (0)	0% (0)
Assiniboine Park Zoo (Q4808913)	4	100.0% (4)	50.0% (2)	50.0% (2)	0% (0)	0% (0)
Adelaide Zoo (Q3244627)	3	100.0% (3)	66.67% (2)	100.0% (3)	0% (0)	0% (0)
Ménagerie du Jardin des Plantes (Q1957702)	3	100.0% (3)	66.67% (2)	66.67% (2)	0% (0)	0% (0)
Liberec Zoo (Q2315180)	3	100.0% (3)	33.33% (1)	100.0% (3)	0% (0)	0% (0)
Parc animalier d'Auvergne (Q16668906)	3	100.0% (3)	0% (0)	33.33% (1)	0% (0)	33.33% (1)
National Zoo & Aquarium (Q495039)	3	100.0% (3)	66.67% (2)	100.0% (3)	0% (0)	0% (0)
Taronga Zoo (Q1064813)	3	100.0% (3)	100.0% (3)	100.0% (3)	0% (0)	0% (0)
Jardin zoologique de la ville de Lyon (Q2387471)	2	100.0% (2)	50.0% (1)	100.0% (2)	0% (0)	0% (0)
Zooparc de Trégomeur (Q3576057)	2	100.0% (2)	50.0% (1)	100.0% (2)	0% (0)	50.0% (1)
Auckland Zoo (Q2665285)	2	100.0% (2)	50.0% (1)	50.0% (1)	0% (0)	0% (0)
Chiba Zoological Park (Q11405998)	1	100.0% (1)	100.0% (1)	100.0% (1)	0% (0)	0% (0)
Turtle Back Zoo (Q7856369)	1	100.0% (1)	0% (0)	100.0% (1)	100.0% (1)	100.0% (1)
Perth Zoo (Q3082583)	1	100.0% (1)	100.0% (1)	100.0% (1)	0% (0)	0% (0)
Zoo Aquarium de Madrid (Q1994806)	1	100.0% (1)	100.0% (1)	100.0% (1)	0% (0)	0% (0)

WIKIMANIA
SINGAPORE

The scale

- One of the largest SPARQL endpoints on the Internet
- 15 Billion triples from 105 Million Items with almost 1.5 Billion statements + smaller number of Properties and Lexemes
- 700k edits per day on Wikidata
- About 5000 requests per minute

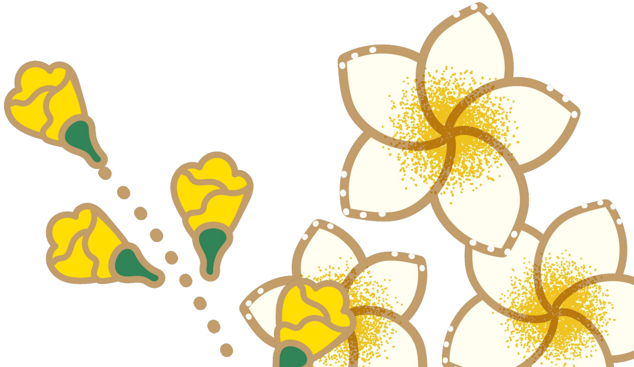
WIKIMANIA
SINGAPORE

Current (interconnected) challenges

WIKIMANIA SINGAPORE

Keeping up with data size

- Wikidata keeps growing
- Blazegraph has no sharding support -> larger disks and memory size required
- Internal Blazegraph limitations for number of allocators



Keeping up with query and write load

- The query load can overload the system, leading to high response times and timeouts
- WDQS write loads are lagging behind Wikidata

WIKIMANIA SINGAPORE

Keeping it stable and secure

- Servers crashing leading to limited capacity and overload
- Blazegraph no longer actively developed / maintained

WIKIMANIA SINGAPORE

WIKIMANIA
SINGAPORE

Consequences

- Legitimate queries time out
- As the graph grows, queries that worked before, now no longer work
- Editors are restricting their editing work
- Editors and reusers are not getting useful new functionality

How have we addressed the problem?

What we’ve done

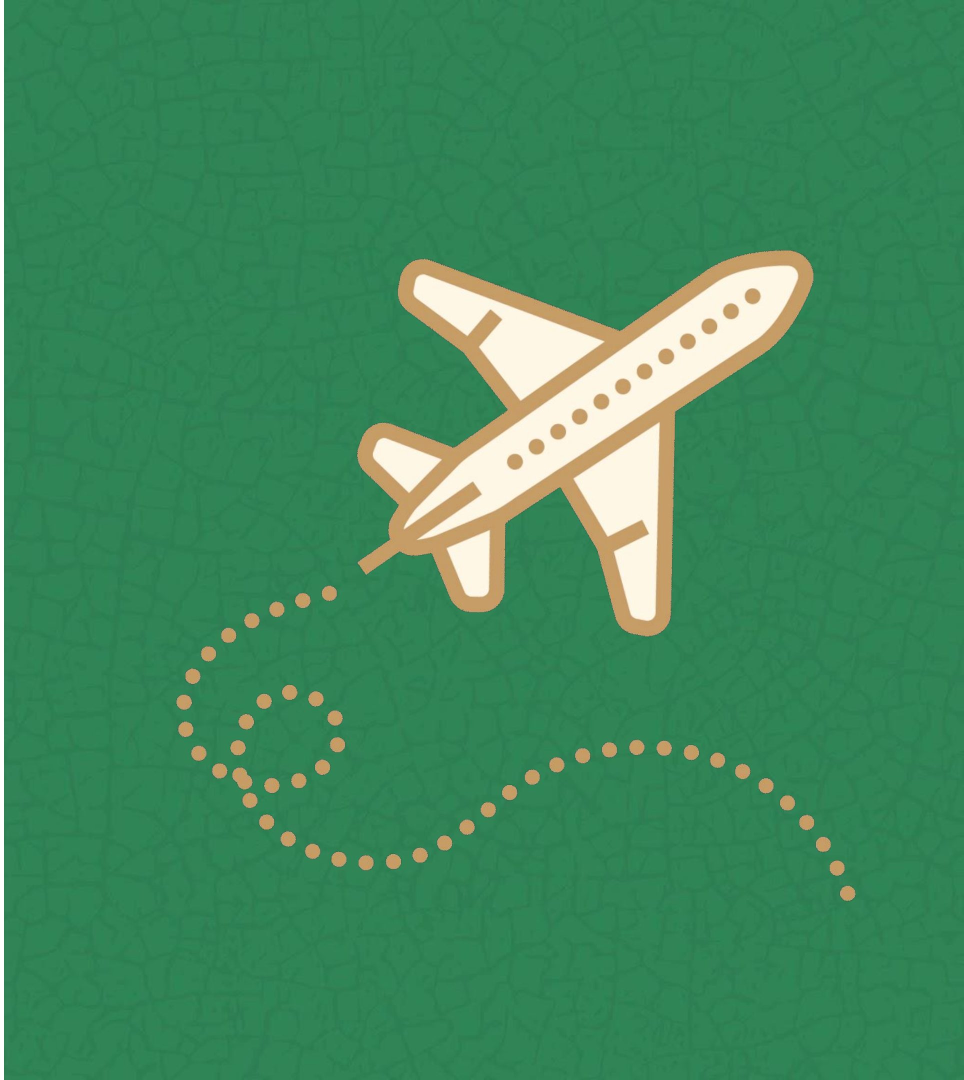
- Introduced a new streaming updater to cope with more edits/min
- Made a disaster mitigation plan
- Got an overview of alternative backends
- Took pressure off the system
    - Built out the Wikibase Ecosystem and especially Wikibase Cloud
    - Developed the Wikibase REST API
    - Improved documentation for the different ways to access Wikidata’s data to help developers chose the right one for their usecase

What we are doing

- Thinking through splitting off a part of the graph into a separate Blazegraph instance while keeping the data in Wikidata
- Discussing the future of the scientific article corpus in Wikidata with the WikiCite community
- Reducing redundant data by introducing a new language code for multilingual content

WIKIMANIA
SINGAPORE

And in the future?



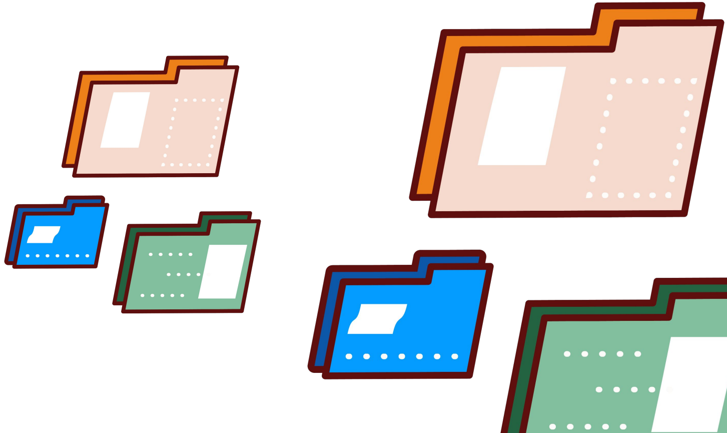
We need to continue addressing the different aspects of the problem

- A lot of data
- A lot of queries
- A lot of edits
- Unmaintained Blazegraph

WIKIMANIA
SINGAPORE

A lot of data

- Move large specialized data out of Wikidata (e.g. Wikibase Ecosystem)
- Continue to reduce redundant data (e.g. mul language code, automated descriptions, Lua modules for inverse relations)
- Split the graph



WIKIMANIA
SINGAPORE

A lot of queries

- Move people to more appropriate access methods to reduce the load
  - Make it easier to work with other existing access methods (e.g. subset dumps, Query Service on cloud providers)
  - Provide additional access methods
  - Automatically reroute queries to more appropriate systems
  - Automatically rewrite inefficient queries
  - Increase the incentives and pressure to move to other access methods

A lot of edits

- Reduce redundant data and the unnecessary edits that come with them

WIKIMANIA
SINGAPORE

Unmaintained Blazegraph

- Evaluate alternatives and move away from Blazegraph
- Evangelize for the development of new graph backends

More questions? Want to stay up-to-date?

Wikidata's weekly newsletter

Search Platform team office hours

lydia.pintscher@wikimedia.de

@nightrose

User:Lydia Pintscher (WMDE)

WIKIMANIA
SINGAPORE

WIKIMANIA
SINGAPORE