



Wikidata Query Service

Where are we? Where is it going?

Guillaume Lederrey

Engineering Manager
guillaume.lederrey@wikimedia.org

Lydia Pintscher

Wikidata Portfolio Lead
lydia.pintscher@wikimedia.de

David Causse

Software Engineer
dcausse@wikimedia.org

This session is recorded: Please mute your microphone and camera when you're not speaking.

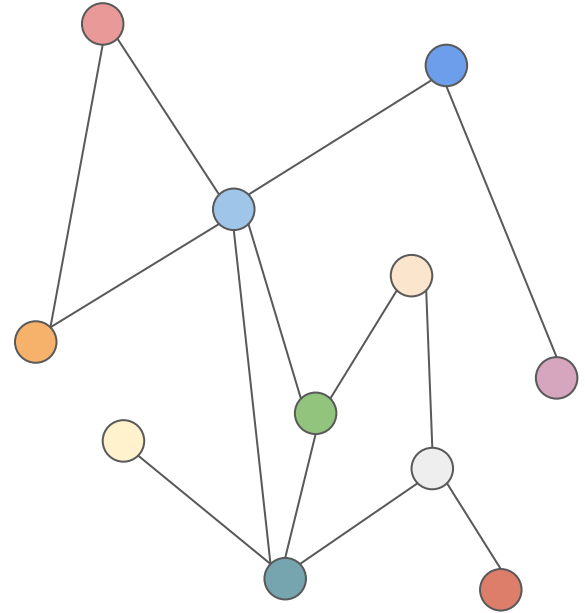
**DATA REUSE
DAYS** February 18-27, 2025



Background

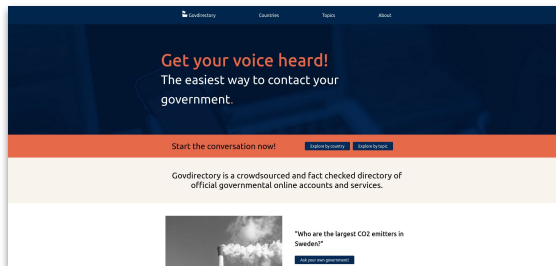
Wikidata Query Service is...

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



Who is using Wikidata Query Service?

- Wikidata and Wikimedia editors
- Knowledge seekers/sharers
- Small and medium-sized reusers
- Tool builders



Wikidata Query Service (powered by SPARQL) - Wikidata

Item	Image	Person	Born	Died	Description	Occupation	Position held	Site	Stats
Q112043		Serge Puskas	1965-04-11	2022-09-26	Russian actor	actor theatrical director stage actor film actor film director		20	17
Q11204481		Endre Molnár	1903-10-20	2022-03-20	Hungarian journalist	journalist		8	18
Q1212127		Endre Molnár	1904-10-14	2022-03-20	Hungarian football player	association football player		1	25
Q10050035		Marián Lešček	1929-03-24	2022-09-26	Slovak association football player	association football player		1	16
Q1748746		Václav Boudný	1924-08-05	2022-09-26	Czech politician	politician lawyer	Deputy Minister of Justice Deputy Minister of Culture Deputy Minister of Justice Vice president of the Chamber of Deputies of the Czech Republic	14	55
Q216033		Greg Kinnear	1952-12-11	2022-09-26		author		5	24

Pandas by place of residence [488 | with source]

Top groupings (minimum 1 items)

Name	Count	Sex	gender	place of birth	date of birth	place of death	date of death
Hamilton Zoo (Q2115846)	10	100.0%	♂	80.2%	0%	50.0%	0%
Edmonton Valley Zoo (Q2103105)	8	100.0%	♂	75.0%	0%	87.5%	0%
Chautauque Zoo at Algonk Park (Q2087901)	6	100.0%	♂	66.6%	0%	100.0%	0%
Chesham Zoo and Gardens (Q2062333)	6	100.0%	♂	33.3%	0%	16.67%	16.67%
Bosumard Gifford Zoo (Q2168576)	5	100.0%	♂	100.0%	0%	0%	0%
Parco Natura Viva (Q2148460)	4	100.0%	♂	75.0%	0%	75.0%	0%
Belfast Zoo (Q2124480)	4	100.0%	♂	25.0%	25.0%	0%	0%
Académie Zoologique (Q2108013)	4	100.0%	♂	50.0%	0%	50.0%	0%
Adelaide Zoo (Q2344627)	3	100.0%	♂	66.67%	0%	100.0%	0%
Ménagerie du jardin des Plantes (Q193752)	3	100.0%	♂	66.67%	0%	66.67%	0%
Liberec Zoo (Q2371761)	3	100.0%	♂	33.33%	0%	100.0%	0%
Parc animalier (Zoo) de Saint-Denis (Q2168680)	3	100.0%	♂	33.33%	0%	0%	33.33%
National Zoo & Aquarium (Q206356)	3	100.0%	♂	66.67%	0%	66.67%	0%
Taronga Zoo (Q2104810)	3	100.0%	♂	100.0%	0%	100.0%	0%
jardin zoologique de la ville de Lyon (Q2087471)	2	100.0%	♂	50.0%	0%	0%	0%
Zooпарк de Троицкое (Q2216527)	2	100.0%	♂	50.0%	0%	100.0%	0%
Académie Zoologique (Q2081400)	2	100.0%	♂	50.0%	0%	50.0%	0%
Chiba Zoological Park (Q2164586)	1	100.0%	♂	100.0%	0%	0%	0%
Turtle Back Zoo (Q2166369)	1	100.0%	♂	0%	0%	100.0%	100.0%
Perth Zoo (Q2082083)	1	100.0%	♂	100.0%	0%	100.0%	0%
Zoo Aquarium de Madrid (Q2166369)	1	100.0%	♂	100.0%	0%	0%	0%

Facts & Queries 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

WikiFix

Recently Edited

--	--	--	--	--	--	--

Highly Ranked (# Wikipedia Articles)

--	--	--	--	--	--	--

The scale

- One of the largest SPARQL endpoints on the Internet
- **16.6 Billion triples** from **114 Million Items** with **1.65 Billion statements** + smaller number of Properties and Lexemes
- **450k edits per day** on Wikidata
- About **7000 to 8000 requests per minute**

The challenges

- Keeping up with data size
- Keeping up with query and write load
- Keeping it stable and secure

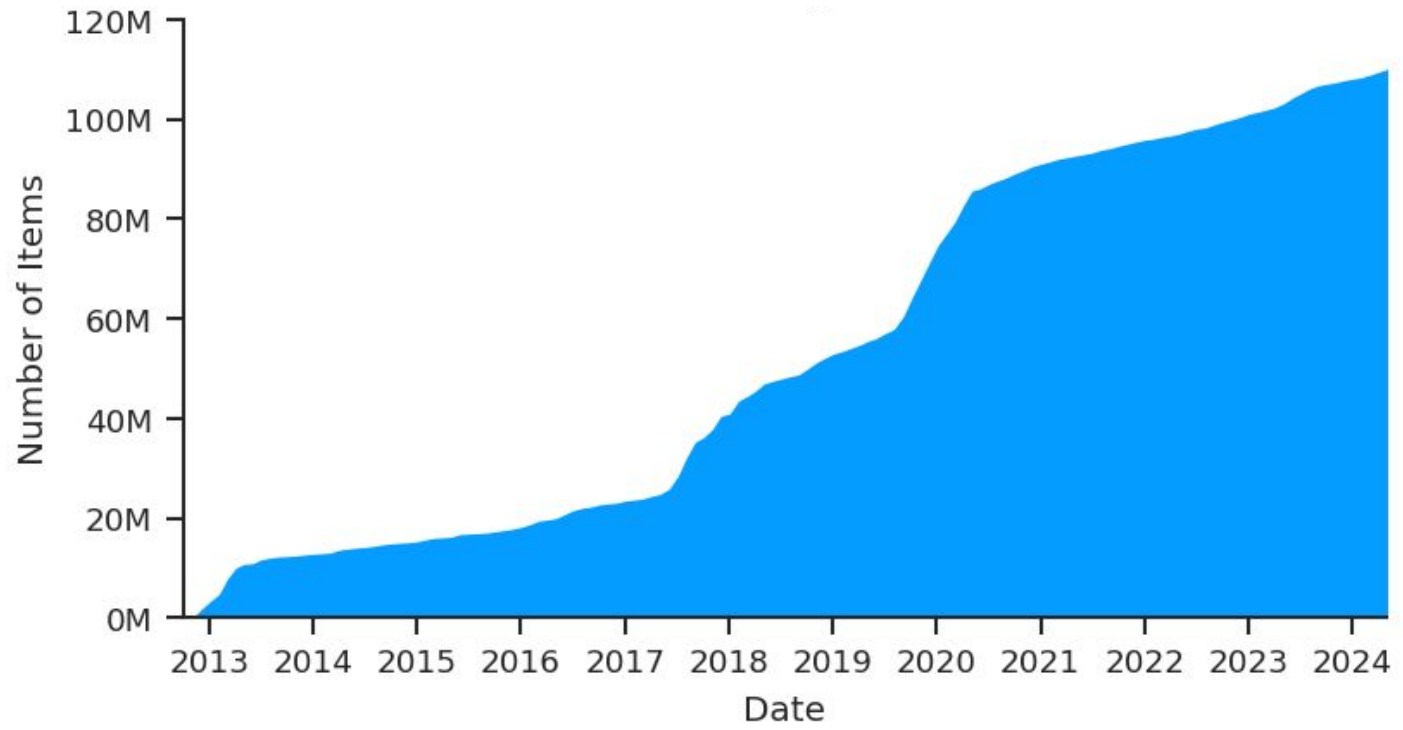


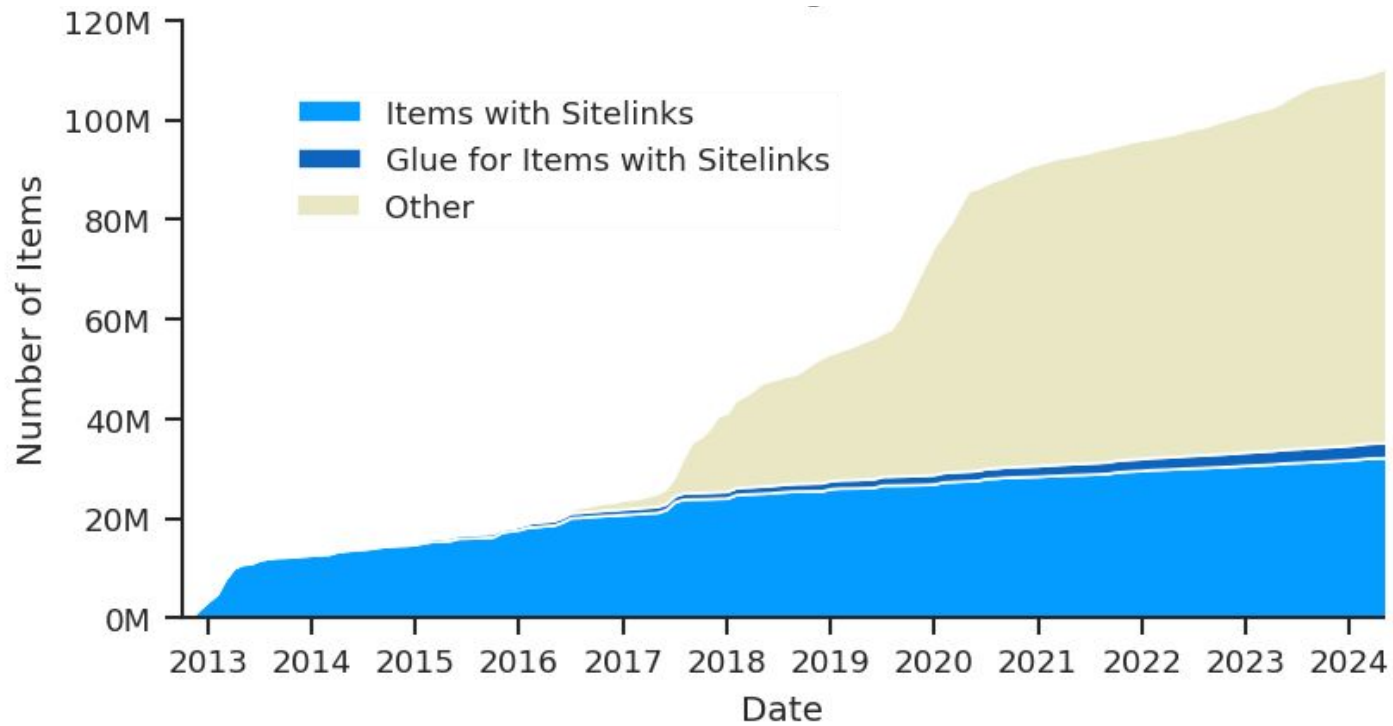
What have we done lately to address the problem?

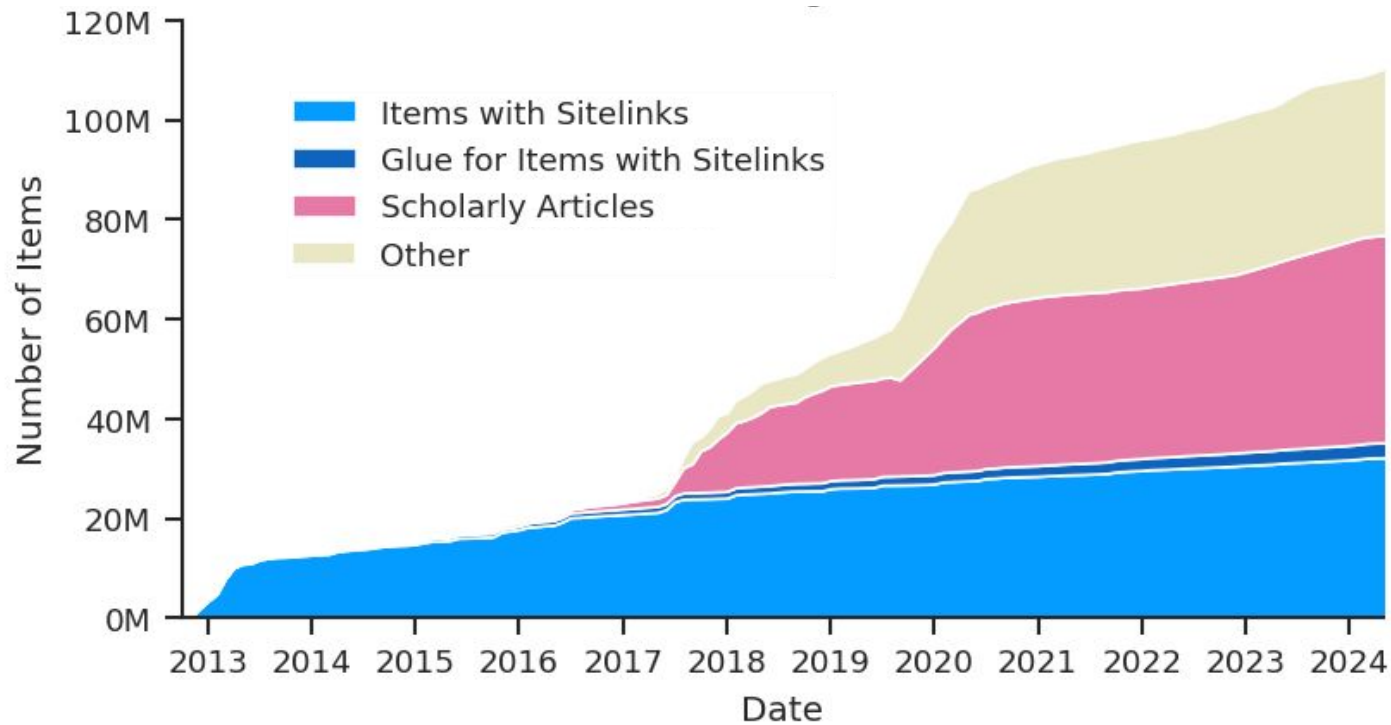
- Built out the **Wikibase Ecosystem** to make it a viable alternative for data that is out of scope for Wikidata
- Created a **developer portal** ([Wikidata:For developers](#)) to guide reusers to the most appropriate data access method
- Developed the **Wikibase REST API** to offer meaningful alternatives for some queries
- Released the **mul language code** to reduce redundant data - nearly 10 million Items already have labels and aliases migrated to mul

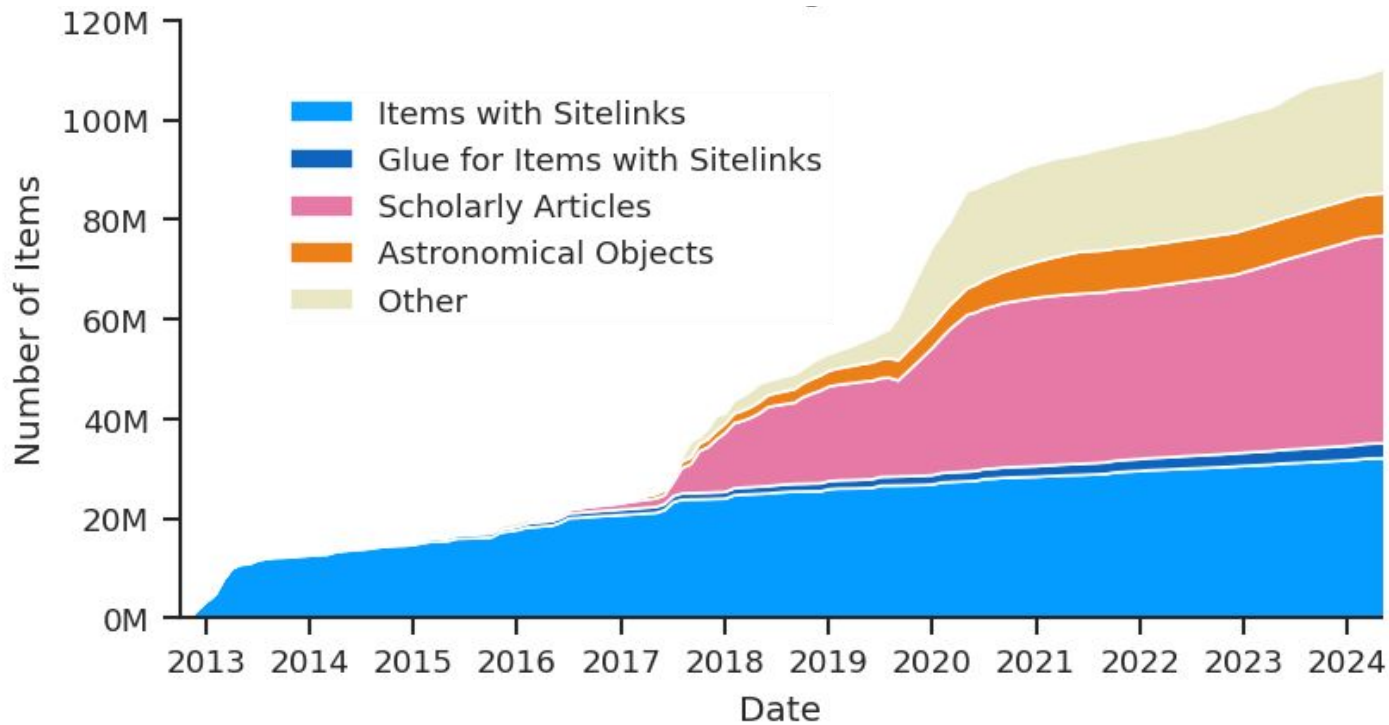
- Community testing and benchmarking of [alternative backends](#)
- Setting up and testing the [graph split](#)
- Exposed the [RDF update stream](#) publicly

- Research into the current data access methods and how to improve them
- Research to better understand the makeup of the data, queries and more











What is coming next?

3 ways forward (short- and longer-term)

- Offering alternative ways to access the data
- Getting the amount of data under control
- Finding alternatives for a better future

Offering alternative ways to access the data

- Additional and better ways to access the data (further improved REST API, subset dumps, ...)
- Wrapping up the data access methods research
- Developer outreach to make the different data access methods more widely known

Getting the amount of data under control

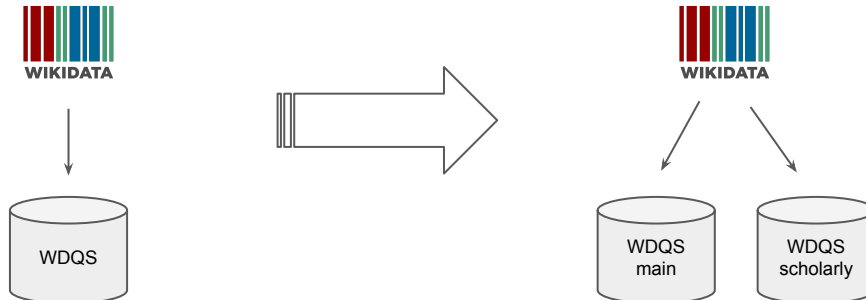
- Continue building out the Wikibase Ecosystem so more data can find a home there
- Continue reducing redundant data (automated descriptions?)
- Mass editing policy RfC and later notability policy
- Split the graph in the Query Service

In more languages			
Language	Label	Description	Also known as
default for all languages	No label defined	-	
English	opacaline C	chemical compound	
German	No label defined	chemische Verbindung	
Spanish	No label defined	compuesto químico	
Latin	No label defined	compositum chemicum	
Afrikaans	No label defined	chemiese verbinding	
Aragonese	No label defined	compuesto quimico	
Arabic	No label defined		مركب كيميائي
Asturian	No label defined	compuestu químicu	
Belarusian	No label defined	хімічнае злучэнне	
Belarusian (Taraskevica orthography)	No label defined	хімічнае злучэнне	
Bulgarian	No label defined	химично съединение	
Bangla	No label defined	রাসায়নিক যৌগ	
Breton	No label defined	kediad kimiek	
Catalan	No label defined	compost químic	
Czech	No label defined	chemická sloučenina	
Welsh	No label defined	cyfansoddydd cemegol	
Danish	No label defined	kemisk forbindelse	
Swiss High German	No label defined	chemische Verbindung	
Greek	No label defined	χημική ένωση	
Canadian English	opacaline C	chemical compound	
British English	opacaline C	chemical compound	
Esperanto	No label defined	kemia kombinaĵo	

The graph split

- Disruptive but necessary to keep the service up
- Currently in the transition and testing phase - switchover at the **end of March 2025**

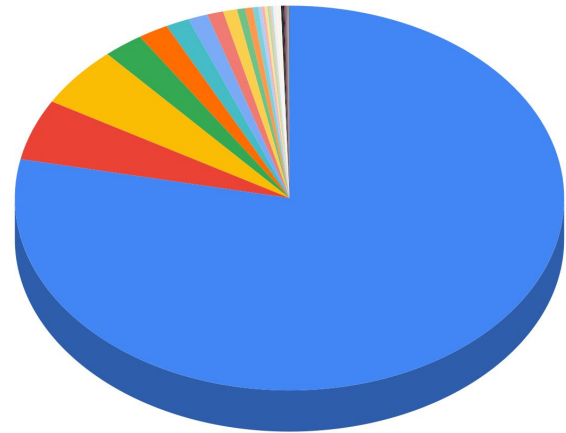
- Largest uniform corpus (scholarly articles) will reside in its own Blazegraph instance
- Editing remains in the same wiki
- Queries that touch scholarly articles will need to be adapted



Query sample to understand graph split impact

- Took a sample of 110k queries in Q3 2024
- 8% of queries touch scientific articles
- 72.2% of these queries come from one user agent (who does not mind), we are working with #2 and #3 isn't really affected

User agent distribution for queries touching scientific articles



Finding alternatives for a better future

- Test [alternative backends](#) and help them improve - [benchmarking](#) ongoing
- Migrate away from Blazegraph



How can you help?

- Help rewrite queries affected by the graph split on [Wikidata:Request a query](#)
- Test [alternative backends](#) and help them improve
- Adapt tools and queries you use to the upcoming graph split
- Migrate redundant labels and aliases to the new mul language code



Thanks for your attention!

Questions and problems: [Wikidata:Report a technical problem/WDQS and Search](https://www.wikidata.org/wiki/Wikidata:Report_a_technical_problem/WDQS_and_Search)

Get in touch with us:

Guillaume Lederrey

guillaume.lederrey@wikimedia.org
IRC: gehel

Lydia Pintscher

@nightrose
lydia.pintscher@wikimedia.de

David Causse

dcausse@wikimedia.org

**DATA REUSE
DAYS**
February 18-27, 2025