

MariaDB Vector: Making AI transparent on your own data Use case: Wikipedia editing with RAG

foss-north, Göteborg, Sweden Fri 14 Apr 2025

### Robert Silén

robert.silen<u>@mariadb.org</u> Community Advocacy, MariaDB Foundation



# Why transparency on AI on your own data?

- When accuracy, sources and accountability matter for the user
- When you need control and insight to intermediate steps

How: Use an open source database for your AI apps:

- Store all data in the same open source standard RDBMS: data, sources, vectors and intermediate results
- Single queries can combine vector & standard data
- Developers know the standard RDBMSes
- Numerous tools work with the standard RDBMSes
- Lack of vendor lock in
- Lower total cost



# Agenda

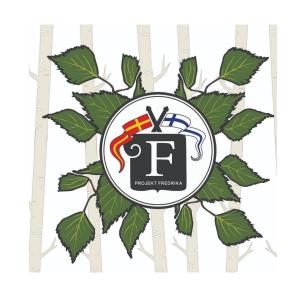
- 1. Use case: RAG for editing Wikipedia
- 2. What exactly **is** MariaDB Server?
- 3. What AI functionality does MariaDB have?
- 4. What are the main <u>use cases</u> for MariaDB Vector?
- 5. Steps in <u>creating RAG</u> with MariaDB Vector
- 6. So where is the advantage of using MariaDB?
- 7. <u>Technical details</u> (another TOC; on your own time)



# Projekt Fredrika

- Non-profit founded 2017 by Kaj Arnö
- Improve Swedish Finland on open data
- Wikipedia, Wikidata, Wikimedia Commons

 Fredrika Runeberg, wife of national poet Johan Ludvig Runeberg.
 Journalist, and first women author of Finland







- 2	Α	В	С	D	E	F	G	н	E	J	K	L	М	N	0	Р	Q	R	s	Т	U	٧	W	Х	Y	Z	AA	AB
1				TOTAL \	_	ANGUAGE (3	0d/1y):	576.920	.134.862	499.389	1.379.816	1.345.578	7.528.948	216.016	1.377.487	100.586	704.147	2.061	20.449	18.984	161.212	97.146	762.688	3.468	43.679	25.441	158.733	14.13
2	Q-kod	Artikel	Тур	Prio	Commons	s Svenska ar	ng_coun	sv_30d	sv_1y	fi_30d	fi_1y	en_30d	en_1y	de_30d	de_1y	fr_30d	fr_1y	ca_30d	ca_1y	cs_30d	cs_1y	es_30d	es_1y	et_30d	et_1y	he_30d	he_1y	hu_30
3	Q2815380	0 Kaj_(humorgrupp)	KAJ	1		https://sv	22	184464	195549	155452	160394	99124	99124	9179	9179	2127	2127	103	103	535	535	2246	2246	582	582	774	774	3
	Q1315611	1 Bara_bada_bastu	KAJ	1		https://sv	15	38905	38905	52303	52303	76262	76262	4999	4999	652	652									621	621	2
5	Q9624204	4 Axel_Åhman	KAJ	1	Category:	https://sv	3	80208	82979	23891	23891																	
3	Q1315613	3 Kevin_Holmström	KAJ	1		https://sv	3	72369	72369	25093	25093																	
7	Q1315530	0 Jakob_Norrgård	KAJ	1	Category:	https://sv	3	80341	80341	26880	26880																	
3	Q1461092	2 Finlandssvenska	Språk	1		https://sv	22	5977	24337	0	9108	10100	54574	809	5620			21	104			0	5047					
9	Q815342	Nessun_dorma	Musik			https://sv	37	2757	11121	0	6592	27664	459948	4076	79332	1761	34119	0	2991	419	7808	4465	70346			121	1480	
0	Q7654934	4 Österbottniska	Språk	2	!	https://sv	5	13134	16382	7984	9146	11216	13672															
1	Q2156306	6 Vörå	Ort	- 2		https://sv	25	38566	43268	74604	86316	23280	26725	1900	2159	285	350	0	0	111	111	227	313	150	150			
2	Q57036	Bastu	Kultur	- 2		https://sv	72	4995	13998	4363	28150																339	6
3	Q726673	Finlandssvenskar	Nation	- 2	!	https://sv	33	11489	34877	5450	36436	1539															786	
4	Q3188515	5 Professionella_pjasalappar	KAJ			https://sv	1	993	1157									6										
5	Q5702	Österbotten_(nutida_landskap)	Ort	- 2	!	https://sv	60	0	32073	7186	47272	1172						1									281	
6	Q1332333	3 Taco_hej_(me_Gusta)	KAJ	- 2	!	https://sv	1	1099	1099																			
7	Q1334446	6 Pa_to_ta_na_kako?	KAJ	2		https://sv	1	43	43									4										
8	Q9027	Svenska	Språk	3	1	https://sv	201	0	76456	2542	26566	3317				_4		æ	1777								755	4
9	Q33	Finland	Ort	3	1	https://sv	321	0	118268	46408	515465	21872					,			-	h						093	50
0	Q643535	Österbotten_(historiskt_landskap)	Ort	3	1	https://sv	23	1384	6014	613	6500	209			and it						13			107	1111			
1	Q1068564	4 Svenska_Österbotten	Ort	3	1	https://sv	1	4461	8462										N. Contraction	(2)								
2	Q1232329	9 Språk_i_Finland	Språk		i	https://sv	12	0	5764	1031	8689	504	86		Va S					-		•		10				
3	Q3189004	4 Kom_ti_byin	KAJ	3		https://sv	1	975	1234				-		100 mg/ (100 mg/)		_			1								
	Q3751292		10000												1-		_			A SECTION AND ADDRESS OF THE PARTY OF THE PA				A comment	2			
4		2 Svenskfinland	Ort	3		https://sv	6	1901	7457	165	1146	41				1			1									
	Q125080		Ort	3		https://sv https://sv				165 0	1146 86652	41 749					•		1		13						409	1
5	Q2E5000						6	1901	7457			41 749 78994					•				13/						409 195	
5 6	Q1184211	Vasa	Ort	-		https://sv	6 80	1901 0	7457 44156	0	86652		•							No.	13						409	70
5 6 7	Q1184211 Q235269	Vasa 1 Eurovision_Song_Contest_2025	Ort Musik	4		https://sv https://sv	6 80 48	1901 0 28205	7457 44156 60381	0 39018	86652 101964	78994									1						409	70
5 6 7 8	Q1184211 Q235269 Q4261412	Vasa 1 Eurovision_Song_Contest_2025 Pernilla_Karlsson	Ort Musik Musik	4		https://sv https://sv https://sv	6 80 48 27	1901 0 28205 1698	7457 44156 60381 2184	0 39018 2780	86652 101964 5485	78994 245									3						409	70
5 6 7 8	Q1184211 Q235269 Q4261412 Q945779	Vasa 1 Eurovision_Song_Contest_2025 Pernilla_Karisson 2 Tävlingen_för_ny_musik	Ort Musik Musik Musik	4		https://sv https://sv https://sv https://sv	6 80 48 27 13	1901 0 28205 1698 681	7457 44156 60381 2184 2151	0 39018 2780 3800	86652 101964 5485 45585	78994 245 253						a			A					3	409 195	11.700
5 6 7 8 9	Q1184211 Q235269 Q4261412 Q945779 Q3351781	Vasa 1 Eurovision_Song_Contest_2025 Pernilla_Karlsson 2 Tävlingen_för_ny_musik När_jag_blundar_(Pernilla_Karlsson-låt)	Ort Musik Musik Musik Musik	4		https://sv https://sv https://sv https://sv https://sv	6 80 48 27 13	1901 0 28205 1698 681 0	7457 44156 60381 2184 2151 1529	0 39018 2780 3800 1379	86652 101964 5485 45585 2314	78994 245 253									I						409	70
25 26 27 28	Q1184211 Q235269 Q4261412 Q945779 Q3351781 Q3431686	Vasa 1 Eurovision_Song_Contest_2025 Pernilla_Karisson 2 Tävlingen_för_ny_musik När_jag_blundar_(Pernilla_Karisson-låt) 1 Vie_X3M	Ort Musik Musik Musik Musik Kultur	4		https://sv https://sv https://sv https://sv https://sv	6 80 48 27 13 15	1901 0 28205 1698 681 0	7457 44156 60381 2184 2151 1529 1708	0 39018 2780 3800 1379 507	86652 101964 5485 45585 2314 3284	78994 245 253															409 195	70



MariaDB Foundation https://mariadb.org



## **KAJ** references

### References [edit]

- ^ KAJ med i Melodifestivalen utlovar dialekt och sauna-pop ≥ [KAJ in Melodifestivalen – promises dialect and sauna pop] (in Swedish). Sveriges Radio. 27 November 2024. Retrieved 24 February 2025.
- A "KAJ" Nästan unplugged. Series 3. Episode 2 (in Swedish). Alexander Theatre: Yle. 21 September 2018. Yle Fem. Retrieved 23 March 2025.
- 4. ^ Aldén, Nicke; Westerlund, Märta (4 August 2016). "Så här gjordes KAJs nya musikvideo i en tagning" ¿ [This is how KAJ's new music video was made in one takel (in Swedish). Yle. Retrieved 23 March 2025.
- 5. ^ Wilow, Felix (13 March 2025). "Marianne Eklöf Törnbloms hyllning till KAJ-Jakob" (Marianne Eklöf Törnblom's tribute to KAJ-Jakob] (in Swedish). Aftonbladet. Retrieved 23 March 2025. [Jakob Norrgård's performance of "Nessun dorma" with his own text about driving in Finland has become a success on social media with over 600,000 views on Youtube.]
- 6. A "KAJ: Nissan bromsa (Nessun dorma), storslagen aria på Alexandersteatern" & [KAJ: The Nissan braked (Nessun dorma), grand aria at the Alexander Theatre] (in Swedish). Yle. 21 September 2019. Retrieved 23 March 2025.
- ^ Isojoki, Jukka (13 June 2019). "Vörjeans är världens bästa rockabillyband från Vörå" [Vörjeans is the world's best rockabilly band from Vörå]. Yle (in Swedish). Retrieved 13 April 2025.
- 4 "KAJ has won 'Melodifestivalen' and will represent Sweden in Basel" ∠.
   Eurovision Song Contest. 8 March 2025. Retrieved 8 March 2025.
- 10. ^ "Axel Åhman om barndomen och fotbollen" 

   <sup>™</sup> [Axel Åhman on childhood and football] (Podcast). Yle. Retrieved 27 February 2025.
- A "Kevin Holmström är Vegas vinterpratare" 

   [Kevin Holmström is Vega's Winter Speaker] (in Swedish). Yle. 20 December 2023. Retrieved 27 February 2025.
- 13. ^ "Recension: Botnia Paradise är som en läcker godispåse från taxfreebutiken, och ingen förälder stoppar dig och säger att "nu räcker det" " [Z\* [Review: Botnia Paradise is like a delicious bag of sweets from the

- Ntinos, Fotios (11 March 2025). "Sweden: "Bara Bada Bastu" hits No. 1 on Spotify's Daily Global Viral Chart!" . Eurovision Fun. Archived from the original on 11 March 2025. Retrieved 11 March 2025.
- Yli-Ojanperä, Elina (16 March 2012). "När jag blundar Pernilla Karlsson laulaa äidistään veljensä sanoin ja sävelin" ¿. Yle Luovat sisällöt ja media (in Finnish). Retrieved 8 March 2025.
- A Karttunen, Camilla (11 March 2025). "KAJ'n kappale rikkoo Euroviisujen sääntöjä – Tässä voi olla ratkaisu" ... Iltalehti (in Finnish). Retrieved 11 March 2025.
- A a b c d Lüttge, Peter; Åhman, Erik; Fant, Izabella (2 March 2025). "I Vörå finns i dag bara stolta och lyckliga människor följ med på en resa genom KAJ-land" (2 (in Swedish). Yle. Retrieved 11 March 2025.
- A a b c d e f g h i Boo, Sofia (9 March 2025). "Det här är Kaj vinnarna av Melodifestivalen 2025" ¿ Göteborgs-Posten (in Swedish). Retrieved 11 March 2025.
- A "Kevin Holmström: "Ingenting är definitivt för mig" " ... Kyrkpressen (in Swedish). 24 January 2024. Retrieved 11 March 2025.

- A "Vinnarna i SFV:s Arvid Mörne-tävling 2017" 

   in Swedish). Svenska folkskolans vänner. 8 May 2017. Retrieved 11 March 2025.
- A <sup>a</sup> <sup>b</sup> Koivuranta, Riitta (10 March 2025). "Tämä kaikki KAJ:n jäsenistä tiedetään" <sup>b</sup>. Helsingin Sanomat (in Finnish). Retrieved 11 March 2025.
- A Lempinen, Jenna (10 March 2025). "Tämä kaikki KAJ-kolmikosta tiedetään – pitkiä parisuhteita ja yllättävät taustat" .
   Ilta-Sanomat (in Finnish). Retrieved 11 March 2025.
- Andrée, Benjamin (8 March 2025). "KAJ:s okända kärleksliv: "Mönster i gruppen" " ¿₹. Expressen (in Swedish). Retrieved 11 March 2025.
- 41. <sup>A a b c</sup> Holmberg, Bianca (16 June 2016). "Humorn är grunden dialekten skatten för KAJ" 

  [A left of the foundation dialect the treasure for KAJ] (in Swedish). Språkbruk. Retrieved 13 April 2025.
- 42. A "Jakob charmade Kevin och Axel genom att hjula och slå volter sedan föddes KAJ" [Z] [Jakob won over Kevin and Axel with his cartwheels and flips and thus, KAJ came into being] (in Swedish). Yle. 10 November 2017. Retrieved 13 April 2025.



# Uppslagsverket Nagu

- Nagu on Wikipedia Fredrika's first project
- Archipelago, South of Åbo
- Improved hundreds of articles with text and images, and over a thousand islands
- Organized "skrivstugor" with local experts
- Results in Wikipedia were good enough for a physical printed encyclopedia





# From themes and experts, to **sources**

- Reliant on experts, and even more so, on **sources**
- Fredrika's Wikipedian in Residence
  - 150 000+ edits on Wikidata
  - 10 000+ edits on 4500+ Wikipedia articles
  - 100+ articles created on Swedish Wikipedia
- How to scale efforts?



ÅLAND - NAGU - INGÅ - IDROTT - FÖRFATTARF - RIRGER CARISTEDT - ÖVRIGT

### En tillbakablick på 2024 Kai / 13.12.2024



dan var en del av oss samlade till influeds 13.12.2023, på Luciadagen. I år har vi träffat en

del av er under Kulturfondens Al-seminarium (1.10.2024) och spart hoppas träffa en del av er vid SLS 140-Aniubileum (S.2.2025) som vi fått inbiudan till

### Fredrika matar Al-draken Kai / 02.10.2024



Fredrika has under 2024 taxit

Linda Mannila för möiligheten att i ett blixttal berätta har Fredrika matar Al-draken, inför en hepeistrad publik om ett hundratal personer med intresse för Al och det svenska Finland. Projekt

### Al Sauna: Improving Wikipedia with Gen Al



the Finnish open culture organisation AvoinGLAM together with Wikimedia oundation organised an

of Finland (Swedish: Riksarkivet: Finnish: Kansallisarkisto) in Finland, Projekt Fredrika had the pleasure of being invited to speak...

#### THUBAKABUICK

### Projekt Fredrika mot ett Alfyllt år 2024



TILLBAKABLICK

med Wikipedia- och Wiki data-projekt tillsammans med SLS (Svenska littera tursällskapet i Finlandi

Det är intressant och öve

lar som läses mest på

och Kotus (Institutet för de inhamska språken) come incotour file att utwockly digitaly askingt file affärstidningen Forum. Samarbetet har synts i olika publikationer (se nadan)

### Sommarpratare 2023



att Yle Vega publicerar sina sommarpratare. Av eammal vana fast utan årets lista på sommarpratare på Wikipedia på

sv:Lista över sommarpratare under 2020talet#Sommar 2023, Det var skoi! Av 35 sommarpratare hade 9 artikel på svenska, 9 på

### Interviu med Ian Sten: att redigera Wikipedia var ju inte så svårt



digera på Wikipedia, [...] Under interviun nämnde Jan: [...] Intervjun på Twitter: [...] Jan Sten berättar om hans första erfarenheter att redigera Wikipedia. Projekt Fredrikas...

#### SLS material når ny publik via Roberts mest lästa 2022 Wikipedia



information når allt fler tack vare ett samarbete med Projekt Fredrika som jobbar för att lyfta fram det svenska Finland på Wikipedia, Samar

betet började i mars 2022 och har redan gett mycket synlighet för SLS material, SLS erundupodrae är att...

#### Ny hemsida med Hugo Robert, Kaj / 14.12.2022



n hakom wahhsaitan nmiaktfradrika fi felir

#### dynamiska Wordoress till statiska Hugo är vi ivriga att berätta om det, för via Hugos fördelar kan vi enklare befrämia Fredrikas mål -...

#### Lyckade skrivstugor: Våra lärdomar

kus är ofta att förbättra existerande artiklar på

Wikipedia, men vi har också skapat en hel del

plock ur statistiken för visningar år 2022 för

Kai. Robert / 01.12.2022



pedianer ska skriva nya ardianer anser vara i enlighet med Wikipedias kvalitetskrav och processer. Via försök och misstag tror vi oss på Fredrika ha kommit på

några nyckellärdomar – väsentliga...

#### Bidrag till Ukraina på Wikipedia

Kai. Robert / 16.09.2022

Veckans tävling på Wikipedia hade temat Ukraina och Sydafrika - två länder inbiudna till bokmässan i Göteborg denna höst. Vi

estämde oss för att hoppa med i tävlingen där deltagaren får poäng för alla bidrag till temats artiklar under tävlingstiden. Projekt Fredrika kombinerade det svenska

### Fredrika utför uppdrag på

Wikipedia Kai. Robert / 05.06.2022

Projekt Fredrika är en ideell förening, som i en förandrad värld anpassat sitt sätt att arbeta genom att övergå från en modell som

bygger på understöd till en modell som bygger

på betalda uppdrag. I denna blogginlaga beskriver vi denna nya modell, dels med tanke på

### Finlandssvenska tekniker



denna blogginlaga frågar nom någon form av digitalisering), hur de lämpligen avbildas på Wikipedia och annar öppen data, på svenska och andra språk.

### Ukrainas historia på Wikipedia

Robert / 16.03.2022 förståeligt högt. Vi samlade statistik på flera språk över några centrala Wiki-

pedia artiklar ur Ukrainas historia. Kanske det inspirerar till att utforska eller att förhättra dom. Siälva märkte vi att flera saknade Wikidata beskrivning på svenska, så vi åtgärdade saken...

#### Nato-intresset större i Finland och Sverige än inom Natoländer siälva

Kaj / 10.03.2022



för Nato är större i Finland och Sverige än inom att ands Nato-medlemeland Estland Lettland och Polen inherrinna. Intresset måtte genom lässtatistik på Wikipedia, rörande artikala am Nata Matadikan

### Källor på Wikipedia

### ASA-uppdrag



Under hösten 2021 johlys de Projekt Fredrika för att bättre avspegla Ålands sjö fartsutbildning of WikiFinlands skärgård på Wikipedia





Robert / 08.03.2022

Hur ser källorna ut för Vin terkriget på Wikipedia? Hur kan en kvantitativ analys om källorna hüllov

Robert / 01.12.2021

Finland är känt som de tusen sjöars och öars land.

# Vision: RAG for Wikipedia editing based on sources

- A human-AI co-editing system where AI helps identify gaps, biases, and outdated information in articles
- Suggests improvements grounded in verifiable, high-quality sources
- Editors remain in control AI becomes a transparent, explainable tool, not an autonomous agent with vetting queues for AI-suggested changes
- Respect existing Wikipedia guidelines as the main priority, in particular NPOV, encyclopaedic language style and high-quality source references

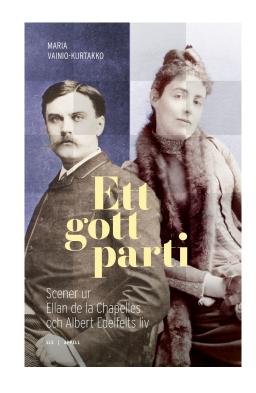


# RAG for Wikipedia editing

- 1. Start from identifying **great sources** (high quality books, articles).
- 2. Prepare by identifying which people, places and other Wikidata Q codes the source material is about

By using **NER** (*Named Entity Recognition*) and with **vectorization**: find closest matching chunks of source and Wikipedia articles by creating vectorized embeddings of them.

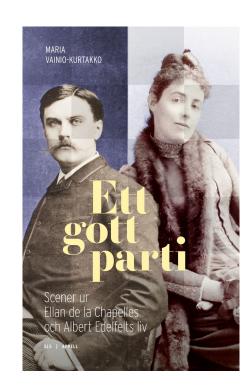
- 3. Generate Wikipedia improvement suggestions with RAG (Retrieval Augmented Generation): prompt an LLM with the source and wikipedia article, with a task and required output.
- 4. Review suggestions, and add to Wikipedia with the **source** reference.





## What was the result?

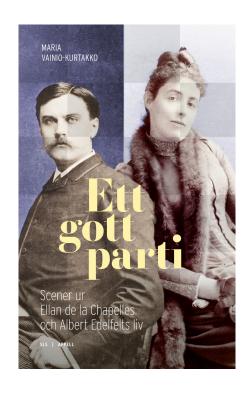
- We found 278 unique people och 265 places in the book
- We generated Wikipedia improvements suggestions on most mentioned
- The author of the book evaluated and
  - o approved % as such to be added to Wikipedia,
  - $\circ$  approved % after manual modifications, and
  - o dismissed %
- Challenges
  - Matching content to Wikipedia: is "Erik" king Erik XIV or Edelfelt's son? is "Alexandra" queen Alexandra of Denmark, or Albert's mother?
  - Filtering out bad suggestion: give AI the option to say "sources are not good enough"





## How was Al used in this use case?

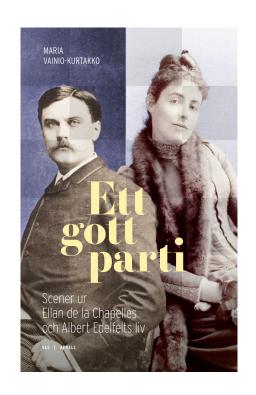
- Wrote python scripts using AI Cursor and Claude
- Chunkification of ebook to chapters, pages, paragraphs
- NER: SpaCy large model in Swedish combined with "entitiy fishing"
  - o https://spacy.io/models/sv
  - o https://github.com/kermitt2/entity-fishing
- Generated wikipedia suggestions with
  - Anthropic Claude best objective output!
  - OpenAI, ChatGPT
  - Mistral
  - Google Gemini





# Insights

- Chunkifications matters
  - One vector per document/article, chapter, page, paragraph or sentence
  - o Most relevant chunk ≠ chunks fed to LLM
- Vectorize whole Wikipedia?
  - All articles in all languages is a 76 GB zip file
  - MariaDB can handle the vectors, vector embedding a subset is cheaper in time and money
  - Pre Vectorized datasets can be available
- https://sv.wikipedia.org/wiki/
   Wikipedia:Projekt\_Fredrika/SLS-AI-pilot





## Future of RAG

- New, improved LLMs with less hallucination and better focus
- Do It Yourself (DIY) vs AI frameworks
- AI RAG Frameworks
  - Better prompts
  - Better chunkification strategy
  - AI-agent workflows
  - Graph database approach to extract claims
  - Multi-hop looping refine user query before vector search
- Every source is unique and RAG will need to be adopted for optimal results



# Thank you!

Contact details: <a href="mailto:robert.silen@mariadb.org">robert.silen@mariadb.org</a>

About:

https://mariadb.org/projects/mariadb-vector/

https://sv.wikipedia.org/wiki/Wikipedia:Projekt\_Fredrika/SLS-AI-pilot

