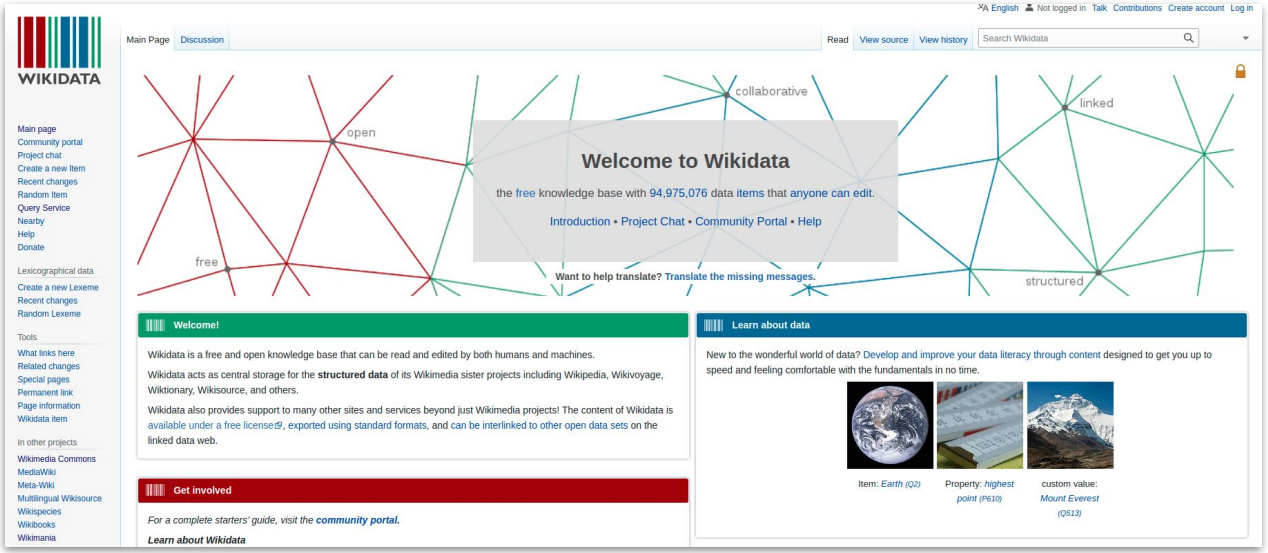
# Wikidata Knowledge Graph to Enable Equitable and Validated Generative AI

Lydia Pintscher & Philippe Saadé
Wikimedia Deutschland

WIKIMEDIA DEUTSCHLAND

# 01 What is Wikidata?

- One of the largest Wikimedia projects
- Data is used in a lot of technology you use every day
- Data available under CC0
- Made for humans and machines
- Multilingual
- Collaborative
- Free and open knowledge graph




WIKIMEDIA DEUTSCHLAND

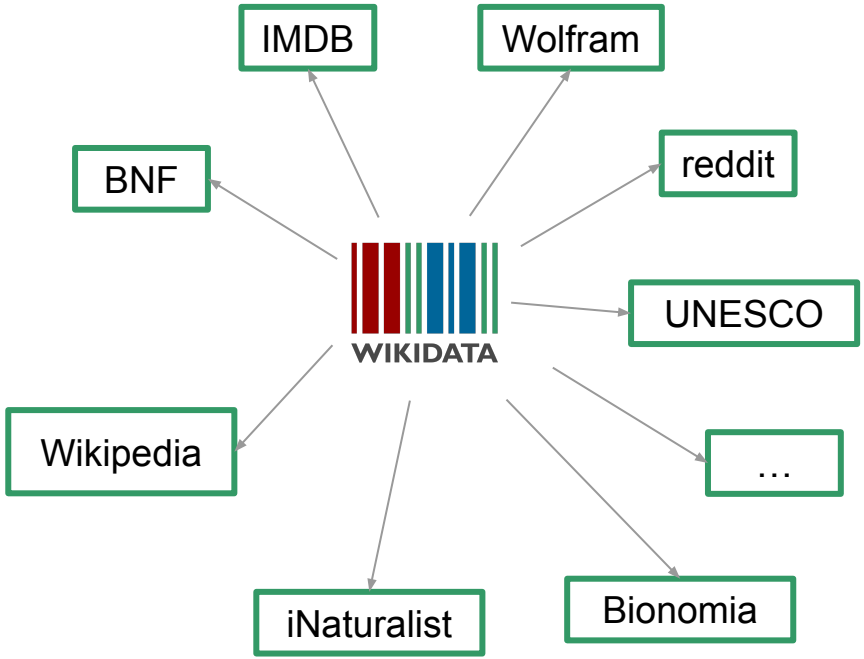
# At the heart of it: 24000 editors

WIKIMEDIA DEUTSCHLAND

Photo by Pierre-Selim Huard (CC BY 4.0)

# What makes Wikidata special?

- Anyone (you!) can be a part of it
- More nuanced modeling of the world and focusing on verifiability
- Multilingual (you'll find lots of Qs, Ps, Ls, …)
- Highly connected internally, to the other Wikimedia projects and to other databases, catalogs, etc. to open up a ton of additional data




WIKIMEDIA DEUTSCHLAND

WIKIMEDIA DEUTSCHLAND

# 02 Wikidata and AI

# Question answering

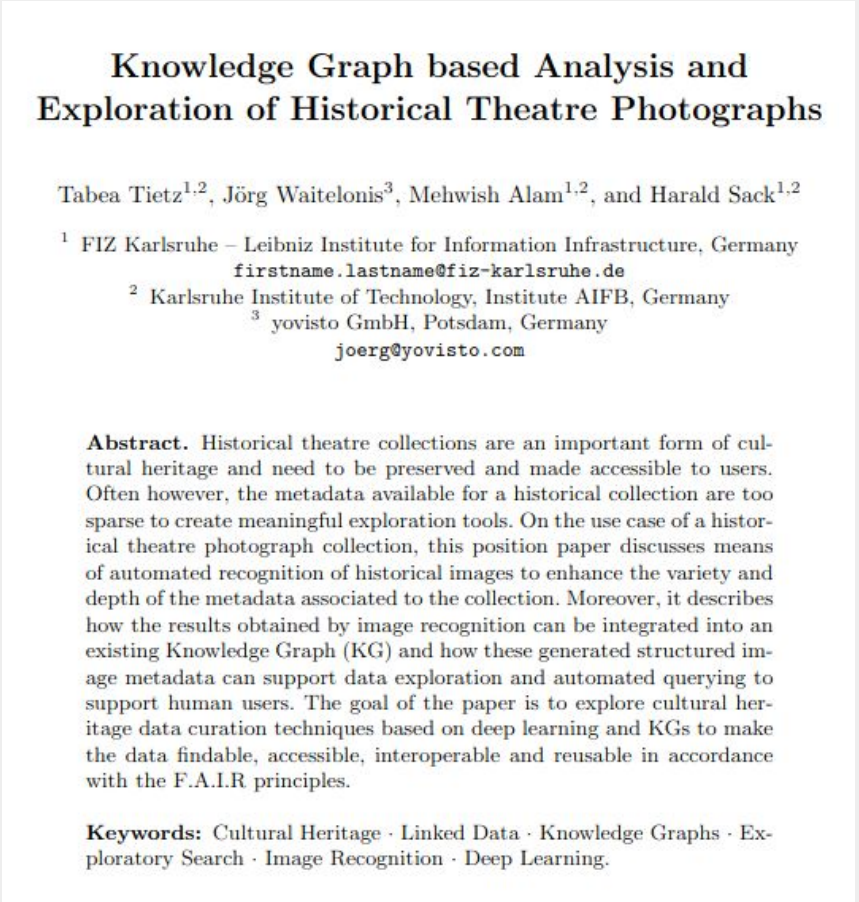
Wikidata acts as background knowledge to answer natural language questions




WIKIMEDIA DEUTSCHLAND

# Computer vision

Wikidata acts as a plausibility check for computer vision results



## Knowledge Graph based Analysis and Exploration of Historical Theatre Photographs

Tabea Tietz[1,2], Jörg Waitelonis[3], Mehwish Alam[1,2], and Harald Sack[1,2]

[1] FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Germany
firstname.lastname@fiz-karlsruhe.de
[2] Karlsruhe Institute of Technology, Institute AIFB, Germany
[3] yovisto GmbH, Potsdam, Germany
joerg@yovisto.com

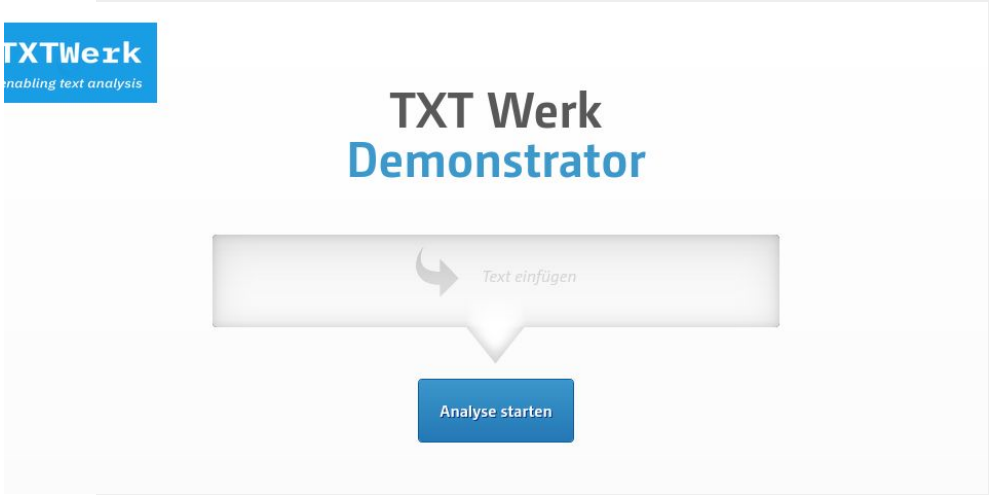
**Abstract.** Historical theatre collections are an important form of cultural heritage and need to be preserved and made accessible to users. Often however, the metadata available for a historical collection are too sparse to create meaningful exploration tools. On the use case of a historical theatre photograph collection, this position paper discusses means of automated recognition of historical images to enhance the variety and depth of the metadata associated to the collection. Moreover, it describes how the results obtained by image recognition can be integrated into an existing Knowledge Graph (KG) and how these generated structured image metadata can support data exploration and automated querying to support human users. The goal of the paper is to explore cultural heritage data curation techniques based on deep learning and KGs to make the data findable, accessible, interoperable and reusable in accordance with the F.A.I.R principles.

**Keywords:** Cultural Heritage · Linked Data · Knowledge Graphs · Exploratory Search · Image Recognition · Deep Learning.

WIKIMEDIA DEUTSCHLAND

# Named entity recognition

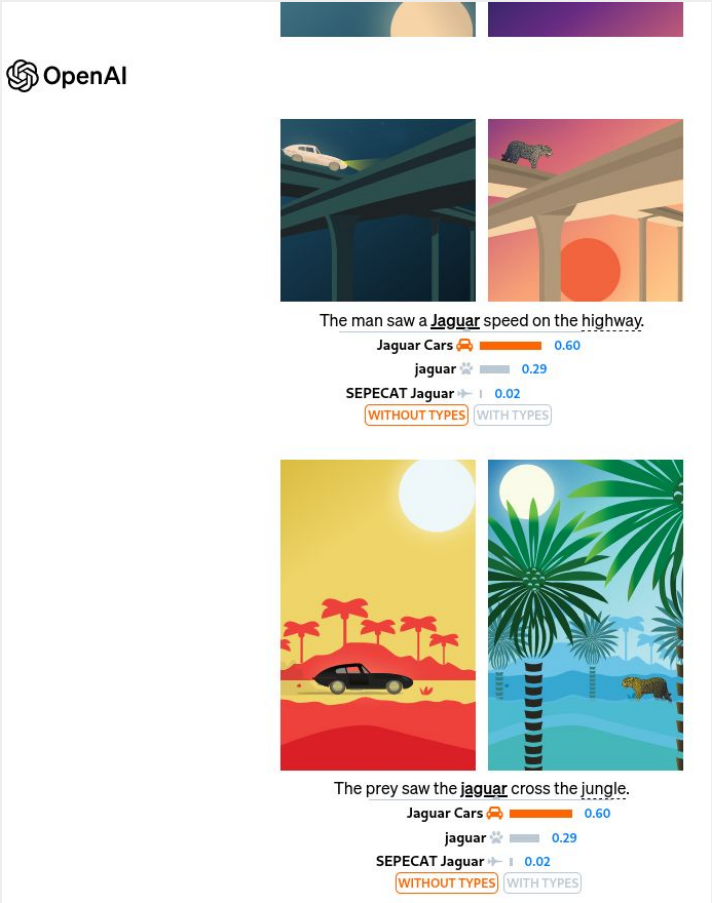
Wikidata provides important entities, stable identifiers for and information about them




WIKIMEDIA DEUTSCHLAND

# Entity disambiguation

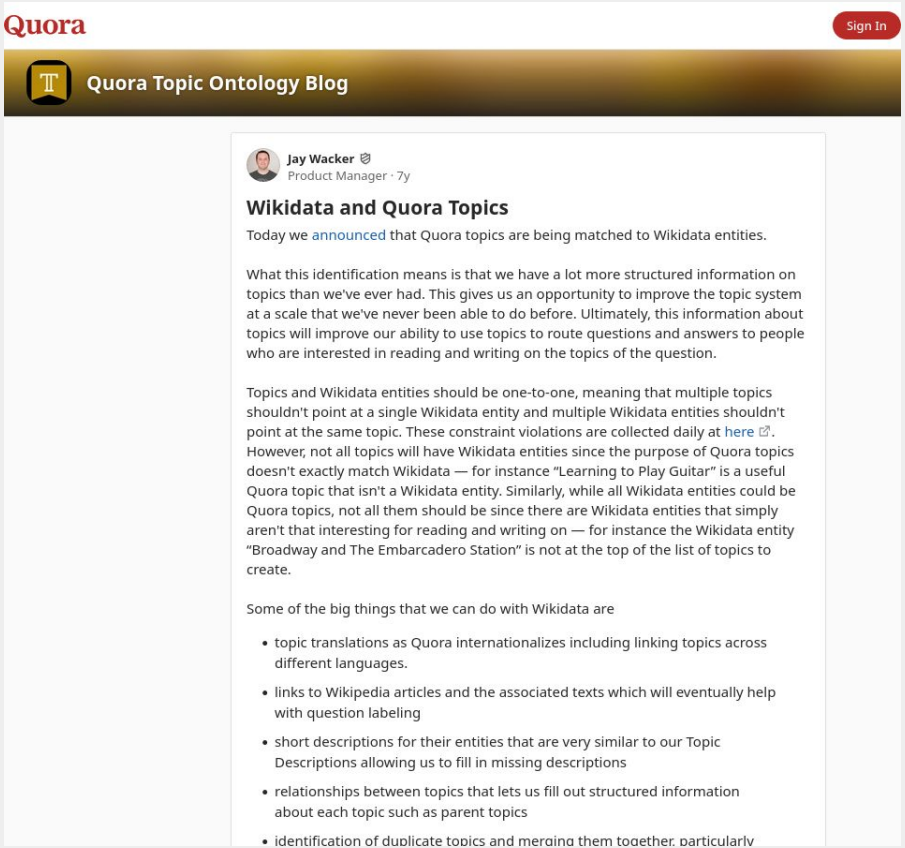
Wikidata provides important entities and information about them




WIKIMEDIA DEUTSCHLAND

# Classification

Wikidata provides relationships and hierarchies through its ontology




WIKIMEDIA DEUTSCHLAND

WIKIMEDIA DEUTSCHLAND

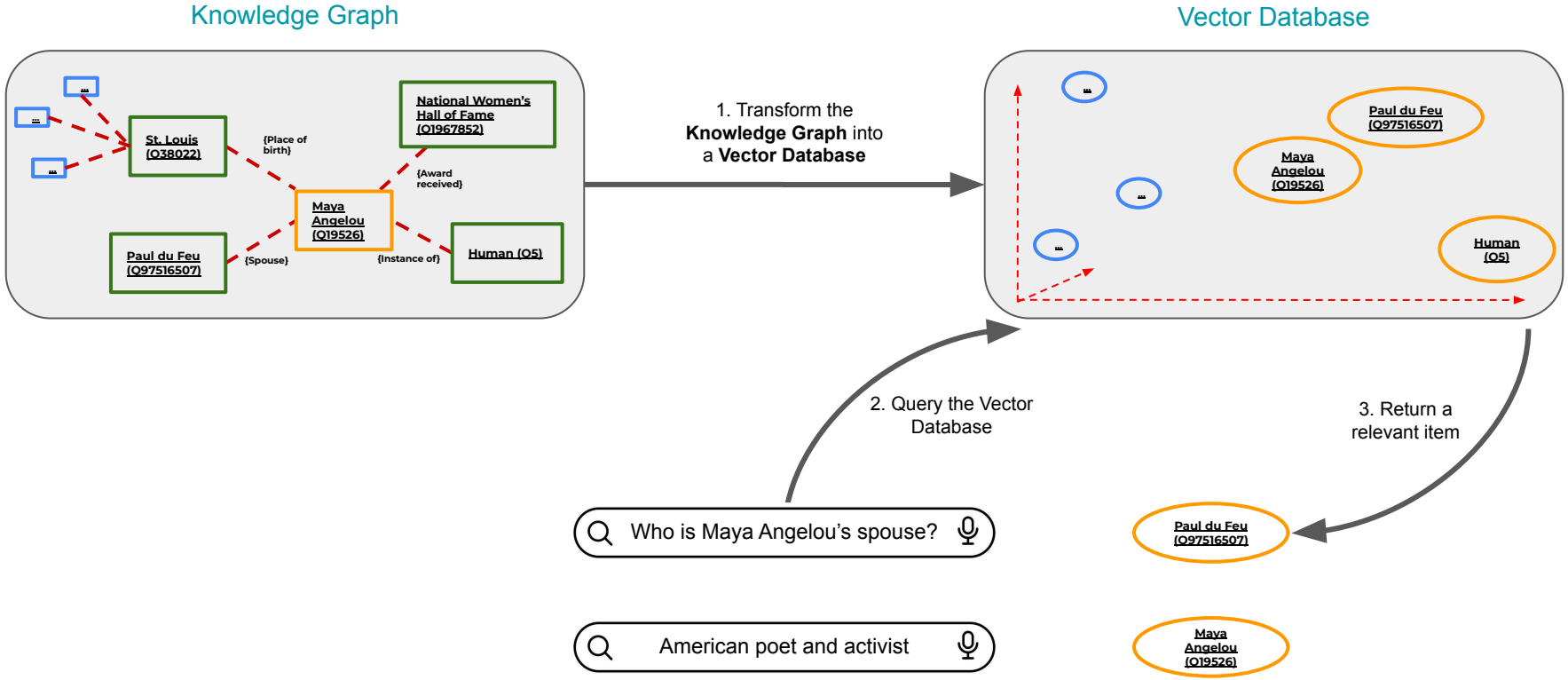
# 03 Wikidata Embedding Vector Database

WIKIMEDIA DEUTSCHLAND

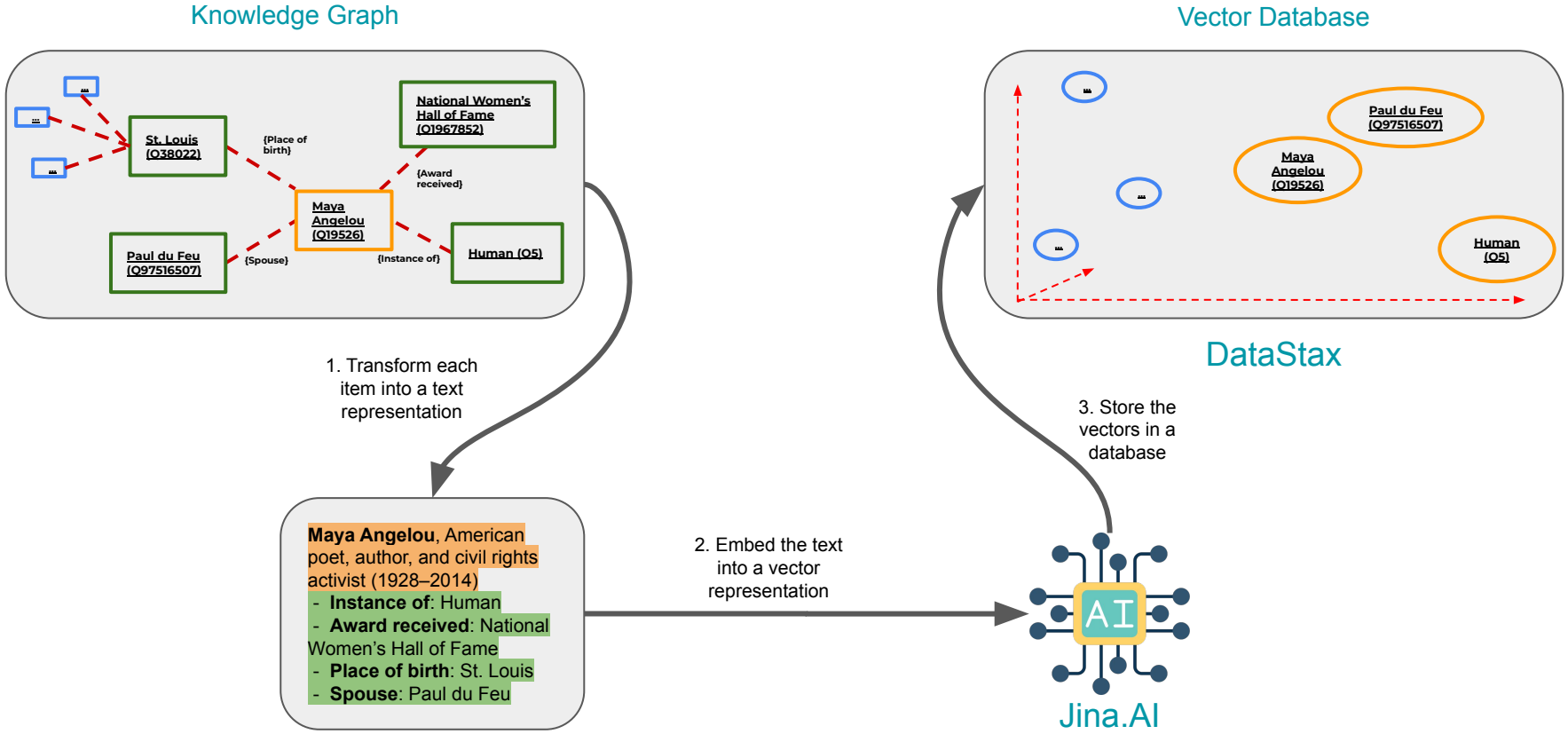
# Create a vectorised database of embeddings with Wikidata's data.

- Freely accessible vector search API
- Support the open-source machine learning community
- Promoting Global Access and Community Collaboration

# Wikidata Vector Database



# Wikidata Vector Database



WIKIMEDIA DEUTSCHLAND

The combination of a **Vector Database** and a **Graph Database** opens up numerous possibilities for future projects, including:

- Semantic search
- Information retrieval
- Integrating Wikidata with other knowledge graphs
- Clustering and classification tasks
- And more…

WIKIMEDIA DEUTSCHLAND

# → Retrieval Augmented Generation (RAG)

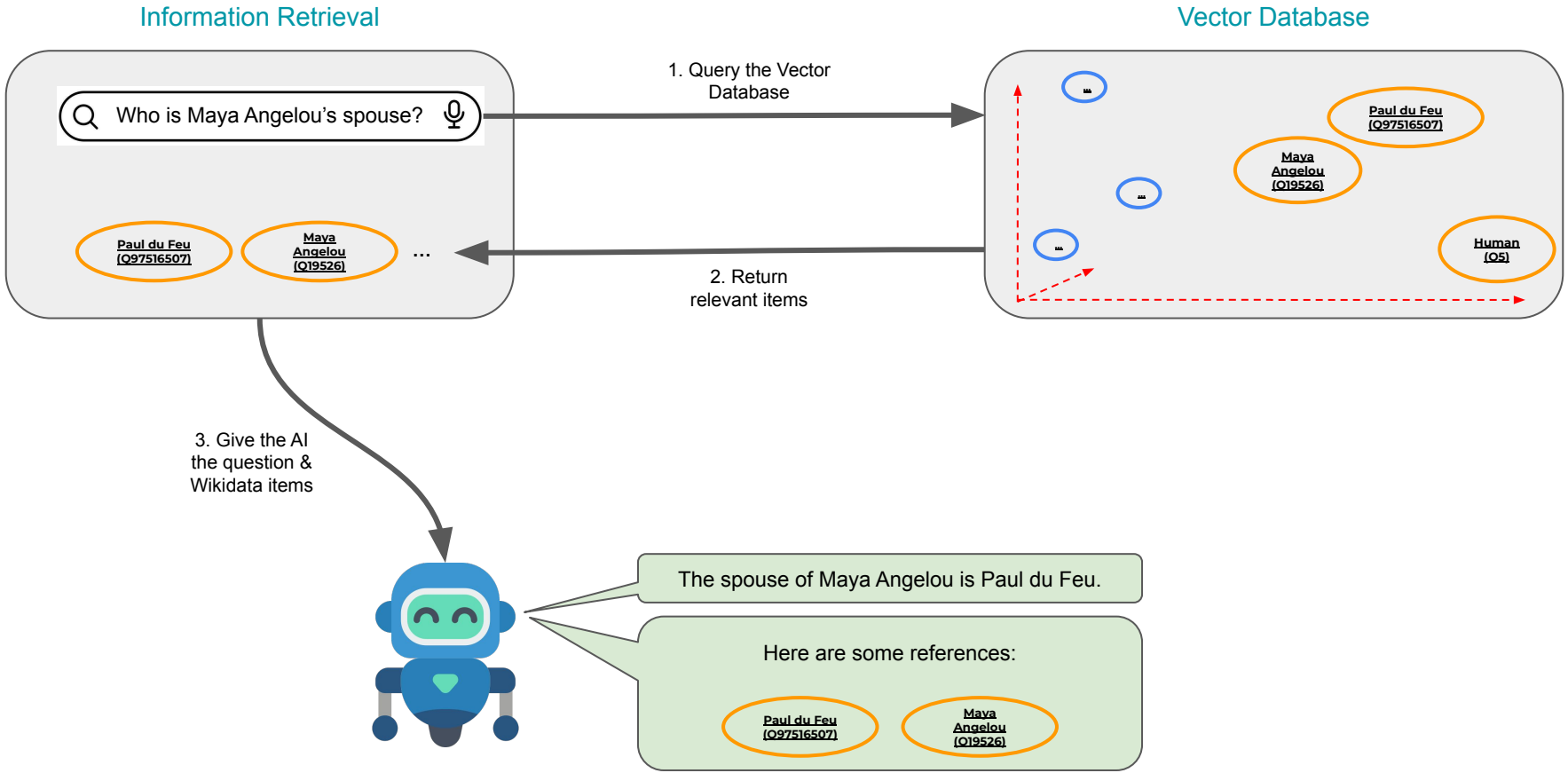
Retrieval Augmented Generation (RAG) is a technique that enables generative AI models to retrieve information, allowing them to respond to user queries by referencing a set of documents.

RAG addresses issues such as:

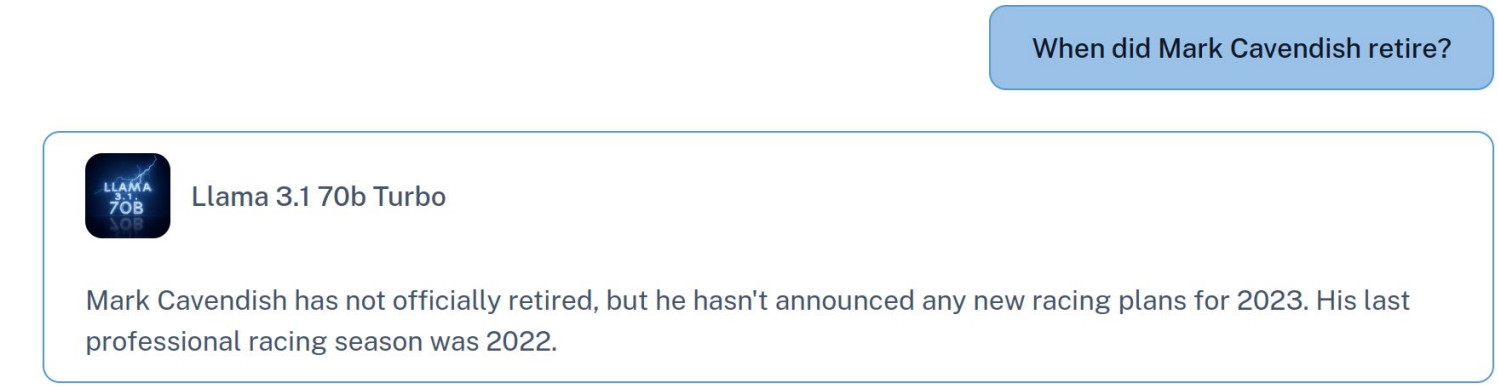
- Reduce "hallucinations" and increase reliability
- Provide up-to-date information

Source: Computerwoche

# Retrieval Augmented Generation (RAG)



# Retrieval Augmented Generation (RAG)




Large Language Model (Llama 3.1 70B Turbo):

- Knowledge Cutoff: December 2023
- No access to external resources
- Expensive to retrain with new data

# Retrieval Augmented Generation (RAG)

When did Mark Cavendish retire? powered by DATASTAX Clone data

AI Summary

Mark Cavendish retired on November 10, 2024.

Generative AI is experimental

**Reference Results** 1-10 of 10

Mark Cavendish (Q207713)

British professional road racing cyclist

Wikidata

Generate summary

Mark Cavendish (Q75248603)

(1941-1941)

Wikidata

Generate summary

WIKIMEDIA DEUTSCHLAND

Benefits of using RAG with Wikidata:

- Increases reliability and ensures up-to-date information
- Allows referencing and fact-checking against known data
- Supporting over 300 languages
- Enable underrepresented communities to make a bigger impact by contributing to Wikidata

# See you on Wikidata!

Philippe Saadé
philippe.saade@wikimedia.de

Lydia Pintscher
lydia.pintscher@wikimedia.de

WIKIMEDIA DEUTSCHLAND

WIKIDATA