

Wikidata Knowledge Graph to Enable Equitable and Validated Generative AI

Lydia Pintscher & Philippe Saadé
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

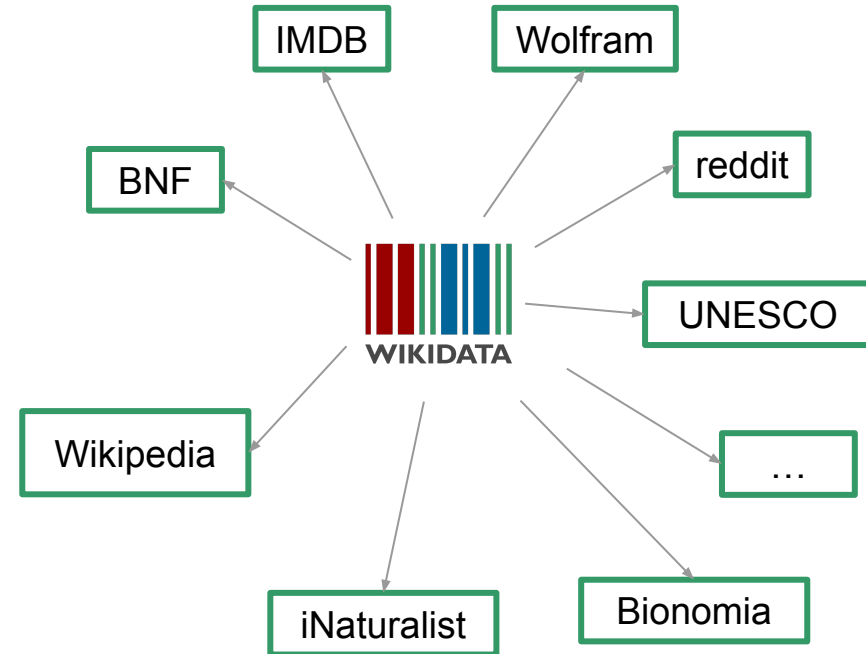
The screenshot shows the Wikidata main page. At the top, there's a navigation bar with 'Main Page' and 'Discussion' tabs, and a search bar. The main content area features a network graph with nodes and edges, labeled with terms like 'free', 'open', 'collaborative', 'linked', and 'structured'. A central grey box contains the text: 'Welcome to Wikidata', 'the free knowledge base with 94,975,076 data items that anyone can edit.', and 'Introduction • Project Chat • Community Portal • Help'. Below the graph, there are three sections: 'Welcome!' (describing Wikidata as a free and open knowledge base), 'Learn about data' (encouraging users to develop their data literacy), and 'Get involved' (providing a link to the community portal).

At the heart of it: 24000 editors



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



02

Wikidata and AI

Question answering

Wikidata acts as background knowledge to answer natural language questions



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 Waitelonis³, 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.

Named entity recognition

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

TXTWerk
enabling text analysis

TXT Werk Demonstrator



The screenshot shows the user interface of the TXT Werk Demonstrator. It features a large, light gray rectangular input field with a curved arrow icon on the left and the text 'Text einfügen' (Paste text) on the right. Below this input field is a blue button with the white text 'Analyse starten' (Start analysis).

Entity disambiguation

Wikidata provides important entities and information about them

OpenAI



The man saw a Jaguar speed on the highway.

Jaguar Cars 🚗 0.60

jaguar 🐆 0.29

SEPECAT Jaguar 🛩️ 0.02

WITHOUT TYPES WITH TYPES



The prey saw the jaguar cross the jungle.

Jaguar Cars 🚗 0.60

jaguar 🐆 0.29

SEPECAT Jaguar 🛩️ 0.02

WITHOUT TYPES WITH TYPES

Classification

Wikidata provides relationships and hierarchies through its ontology



Jay Wacker
Product Manager · 7y

Wikidata and Quora Topics

Today we [announced](#) that Quora topics are being matched to Wikidata entities.

What this identification means is that we have a lot more structured information on topics than we've ever had. This gives us an opportunity to improve the topic system at a scale that we've never been able to do before. Ultimately, this information about topics will improve our ability to use topics to route questions and answers to people who are interested in reading and writing on the topics of the question.

Topics and Wikidata entities should be one-to-one, meaning that multiple topics shouldn't point at a single Wikidata entity and multiple Wikidata entities shouldn't point at the same topic. These constraint violations are collected daily at [here](#). However, not all topics will have Wikidata entities since the purpose of Quora topics doesn't exactly match Wikidata — for instance “Learning to Play Guitar” is a useful Quora topic that isn't a Wikidata entity. Similarly, while all Wikidata entities could be Quora topics, not all them should be since there are Wikidata entities that simply aren't that interesting for reading and writing on — for instance the Wikidata entity “Broadway and The Embarcadero Station” is not at the top of the list of topics to create.

Some of the big things that we can do with Wikidata are

- topic translations as Quora internationalizes including linking topics across different languages.
- links to Wikipedia articles and the associated texts which will eventually help with question labeling
- short descriptions for their entities that are very similar to our Topic Descriptions allowing us to fill in missing descriptions
- relationships between topics that lets us fill out structured information about each topic such as parent topics
- identification of duplicate topics and merging them together, particularly

03

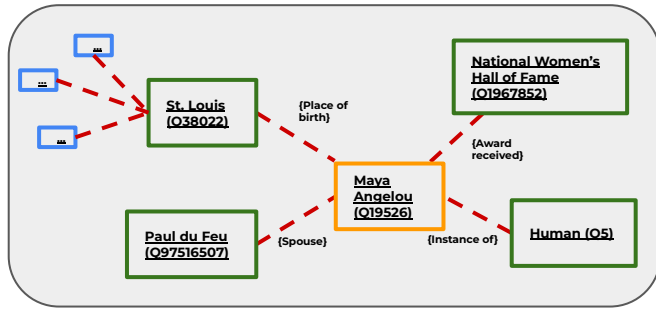
Wikidata Embedding Vector Database

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

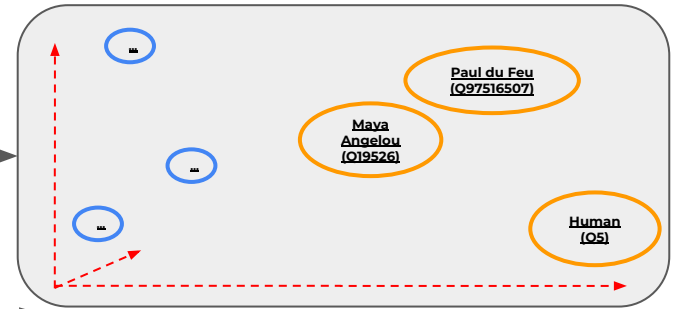
- Freely accessible vector search API
- Support the open-source machine learning community
- Promoting Global Access and Community Collaboration

Wikidata Vector Database

Knowledge Graph



Vector Database



1. Transform the Knowledge Graph into a Vector Database

2. Query the Vector Database

3. Return a relevant item

Q Who is Maya Angelou's spouse?

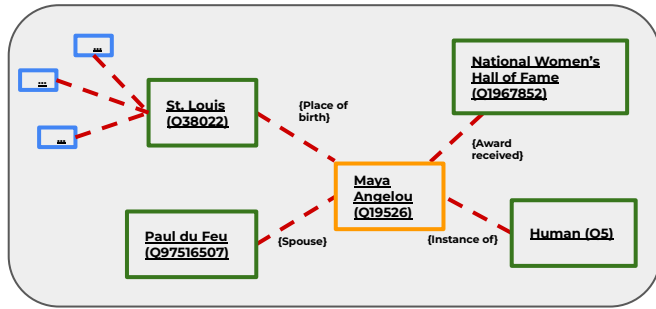
Q American poet and activist

Paul du Feu (Q97516507)

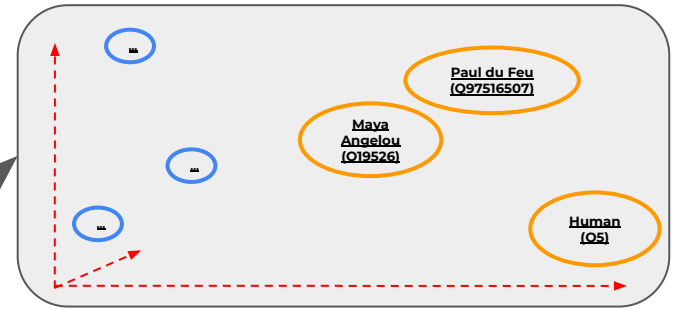
Maya Angelou (Q19526)

Wikidata Vector Database

Knowledge Graph



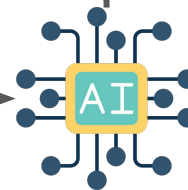
Vector Database



1. Transform each item into a text representation

Maya Angelou, American poet, author, and civil rights activist (1928–2014)
- Instance of: Human
- Award received: National Women's Hall of Fame
- Place of birth: St. Louis
- Spouse: Paul du Feu

2. Embed the text into a vector representation



Jina.AI

3. Store the vectors in a database

DataStax

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...

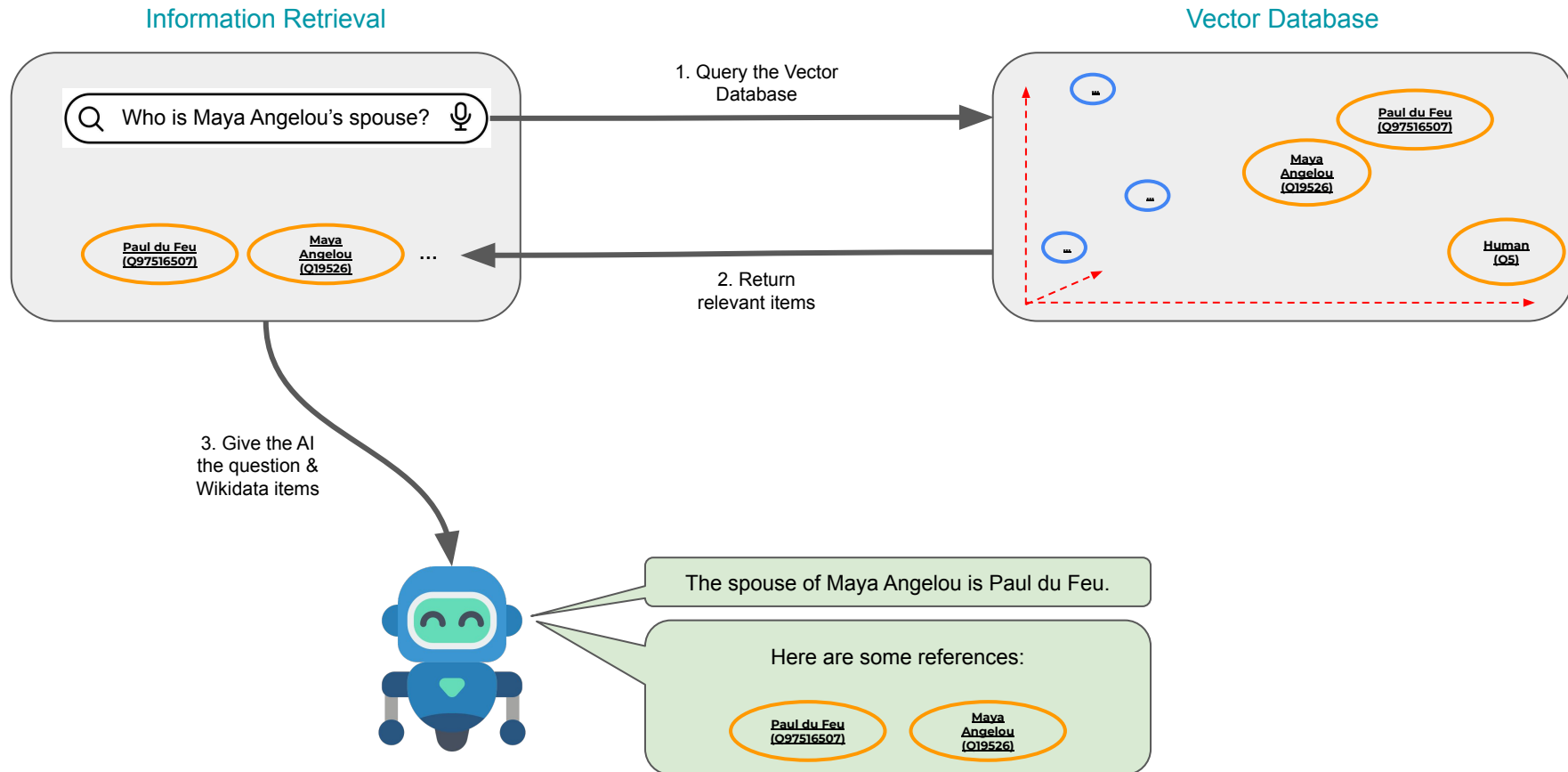
→ Retrieval Augmented Generation (RAG)

[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

Retrieval Augmented Generation (RAG)



Retrieval Augmented Generation (RAG)

When did Mark Cavendish retire?



Llama 3.1 70b Turbo

Mark Cavendish has not officially retired, but he hasn't announced any new racing plans for 2023. His last professional racing season was 2022.

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 **DATASTACK** ✕

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 🌟

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