MiKroloG: The Microdata Knowledge Graph

Teams

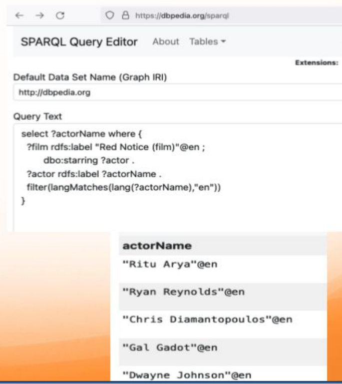
GDD

LACODAM

Knowledge Graphs

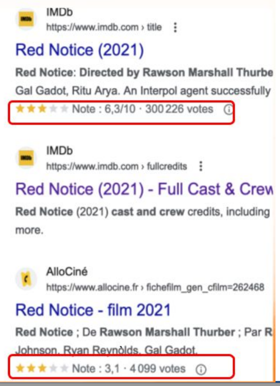
- Allow to answer questions and provide **correct and complete results**
- *Example:Give me the cast members of the film "Red Notice" directed by "Rawson Marshall Thurber".*

On DBPedia KG



Information is missing in Public KGs but in the Web

- Price, Reviews, Events, Me, My lectures, (not all) My Papers, are not in public Knowledge Graphs.
- *But they are on the web, potentially as* ***Microdata****, embedded in web pages.*

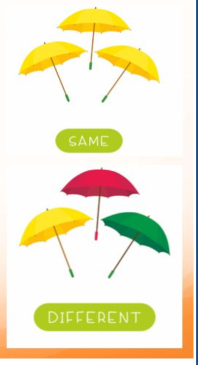


MiKroloG Objective

- **Search the Web With "Things"**
  - Extend public knowledge graphs with Microdata.
  - Search the Web as we do in KG: with queries that return correct and complete results.

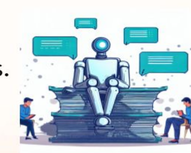
MiKroloG Challenge 1

- Microdata are defined following a standard schema -> schema.org.
- However, nobody knows how the schema is actually used by people.
- **General ideas:**
  - Microdata analysis.
  - Be able to observe Microdata and their evolution.



MiKroloG Challenge 2

- Providing a large KG with two uses cases:
  - Explore the KG with interactive queries. Need a quick answer.
  - Compute correct/complete results. Need fair query processing
- **General ideas:**
  - Exploration-> Approximate Query Processing (AQP)
  - Fairly processing -> Sliced query execution with Web preemption



MiKroloG Challenge 3

- Users don't know how to write SPARQL queries.
- They don't want to learn.
- They prefer to use natural language (NL).
- **General idea:**
  - Translate NL questions into SPARQL queries.

How Microdata are used ?

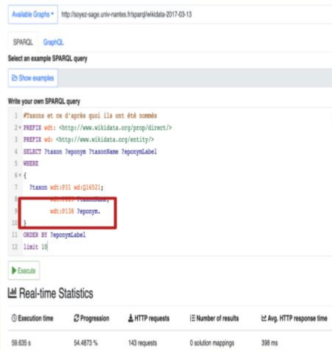
Minh Hoang Dang, Alban Gaignard, Hala Skaf-Molli and Pascal Molli. ***Schema.org: How is it used?*** Poster, accepted at the 22nd International Semantic Web Conference, ISWC 2023.

- Imagine you defined a schema, but you don't know how people use it.
  - *You've defined the class Person, but how many Persons have actually been created?*
  - *You've defined 60 attributes for the class Person, but how many attributes are actually filled in?*

Slicing Top-K Queries in time using Web Preemption

Julien Aimonier-Davat, **Hala Skaf-Molli and Pascal Molli.** *Processing SPARQL TOP-k Queries Online with Web Preemption* in (QuWeDa@ISWC2022).

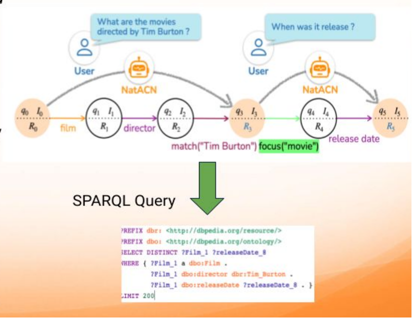
- We defined a preemptable Top-k operator that enables early pruning.
- It voids computing all results before keeping the top-K



Autopilot to help user querying KGs

Julie Boudebs, Sébastien Ferré and Peggy Cellier. ***NatACN: a Natural Language Interaction System.*** Poster at GDR TALN 2022.

- Imagine a user wants to know *films directed by Tim Burton and their release dates in DBPedia KG.*
- The autopilot guides the user during query construction :
  - start with simple queries, identify class, i.e. movie
  - focus navigation on this part of KG



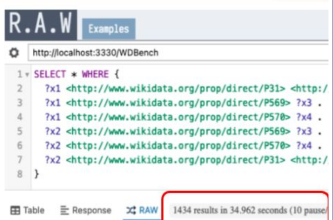
Recipes are better described than Drugs

Live demonstration at: https://schema-obs-demo.onrender.com/

Approximate query processing for SPARQL Servers

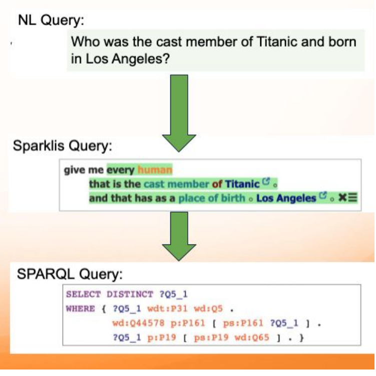
Julien Aimonier-Davat, **Minh-Hoang Dang, Brice Nédelec, Hala Skaf-Molli and Pascal Molli.** ***RAW-JENA: Approximate Query Processing for SPARQL Endpoints.*** *Demo accepted at the 22nd International Semantic Web Conference, ISWC 2023.*

- RAW-JENA integrates sample based approximate query processing directly in Apache JENA.
- Return a sample of results with an estimate of the cardinality of the complete result.
- After 35s RAW-JENA returns
  - 1434 random results
  - Estimate of 26M ± 3M results
  - Exact cardinality 25M results



Large Language Models for Querying KGs

- Instead of autopilot **use a CNL as an intermediate** when translating natural language questions to SPARQL queries
- Fine-tuning of large language models (LLMs) to translate NL to CNL

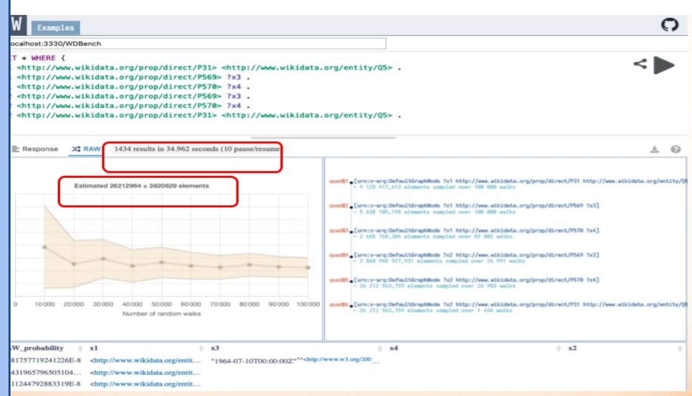


FedShop Synthetic Data Generator

Minh-Hoang Dang, Julien Aimonier-Davat, **Pascal Molli**, Olaf Hartig, **Hala Skaf-Molli**, and **Yotlan Le Crom.** ***FedShop: A Benchmark for Testing the Scalability of SPARQL Federation Engines***, *accepted at The 22nd International Semantic Web Conference, ISWC 2023.*

- Consider *N* web store selling products from a common catalog.
- Users search for products with attributes, reviews and similar products.
- Different configurations.
- FedShop is freely available online at:

https://github.com/GDD-Nantes/FedShop



-JENA is an open source extension of the Apache JENA at: s://github.com/GDD-Nantes/raw-jena

o of RAW-JENA at: https://youtu.be/We5-rG6uxN8

- **Prompt LLM:***"Given the question and the entities generate a $language query!"* where $language ∈ {Sparklis,SQUALL, SPARQL} is the target language
- **Question:** "Who was the cast born member of Titanic and born in Los Angeles ?
- **Entities (from WikiData KG):** (entityID,label, description,prob)
  - {'ID': 'Q44578', 'Label': 'Titanic', 'Description': '1997 film by James Cameron', 'Probability': 0.4532},
  - {'ID': 'Q25173', 'Label': 'Titanic', 'Description': 'British transatlantic passenger liner, launched and foundered in 1912', 'Probability: 0.3498'},
  - {'ID': 'Q65', 'Label': 'Los Angeles', 'Description': 'largest city in California, United States of America', 'Probability': 0.9623}]

J. Lehmann, P. Gattogi, D. Bhandiwad, **Sébastien Ferré**, S. Vahdat. ***Language models as controlled natural language semantic parsers for knowledge graph question answering***. ECAI 2023

Table 2. Evaluation Results on the Mintaka-CNL dataset for different combinations of language models and target languages. The last column shows the absolute difference between hits@1 score of the controlled natural language and SPARQL.

Model	Setting	Language	BLEU	METEOR	ROUGE	Exact Match	Hits@1	Δ Hits@1
GPT-3 Davinci	Fine-Tuning-400	SPARQL	0.60	0.57	0.51	0.09	0.18	-
		Sparklis	0.67	0.71	0.61	0.20	0.20	+0.02
		SQUALL	**0.72**	**0.73**	0.60	0.18	**0.36**	**+0.18**
GPT-3 Curie	Fine-Tuning-400	SPARQL	0.60	0.58	0.52	0.09	0.15	-
		Sparklis	0.67	0.71	**0.62**	0.19	0.19	+0.04
		SQUALL	0.71	0.70	0.60	0.16	0.26	+0.11
Llama 2 7B	Fine-Tuning-400	SPARQL	0.63	0.65	0.46	0.05	0.11	-
		Sparklis	0.62	0.68	0.59	0.19	0.19	+0.08
		SQUALL	0.71	0.72	0.59	0.13	0.23	+0.12
T5-Large	Fine-Tuning-400	SPARQL	0.35	0.47	0.24	0.00	0.03	-
		Sparklis	0.66	0.7	0.61	**0.21**	0.21	**+0.18**
		SQUALL	0.7	0.7	0.58	0.11	0.19	+0.16
GPT Neo	Fine-Tuning-400	SPARQL	0.38	0.33	0.36	0.03	0.04	-
		Sparklis	0.54	0.61	0.55	0.08	0.08	+0.04
		SQUALL	0.64	0.65	0.68	0.18	0.18	+0.14
GPT-2 XL	Fine-Tuning-400	SPARQL	0.31	0.35	0.24	0.04	0.06	-
		Sparklis	0.54	0.6	0.51	0.12	0.12	+0.06
		SQUALL	0.58	0.63	0.54	0.11	0.15	+0.09
BLOOM 1.7B	Fine-Tuning-400	SPARQL	0.38	0.43	0.29	0.02	0.04	-
		Sparklis	0.61	0.58	0.57	0.10	0.10	+0.06
		SQUALL	0.59	0.64	0.63	0.09	0.13	+0.09
GPT-2 0.12B	Fine-Tuning-400	SPARQL	0.28	0.22	0.18	0.00	0.00	-
		Sparklis	0.43	0.51	0.47	0.04	0.04	+0.04
		SQUALL	0.41	0.39	0.33	0.08	0.12	+0.12