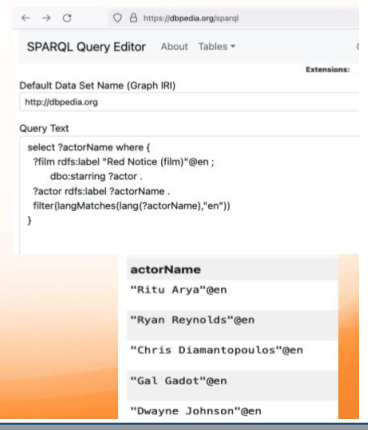
MiKroloG: The Microdata Knowledge Graph

Hala Skaf-Molli, GDD, LS2N
Pascal Molli, GDD, LS2N
Minh-hoang DANG (PhD student), GDD, LS2N
Alban Gaignard, Institut de Thorax
Sébastien Ferré, LACODAM, IRISA
Peggy Cellier, LACODAM, IRISA
Julie Boudebs (PhD Student), LACODAM, IRISA

Knowledge Graphs

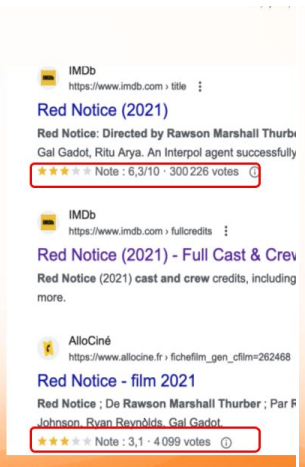
- Allow to answer questions and provide **correct and complete results**
- *Example:Give me the cast members of the film "Red Notice" directed by "Rawson Marshall Thurber".*

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.

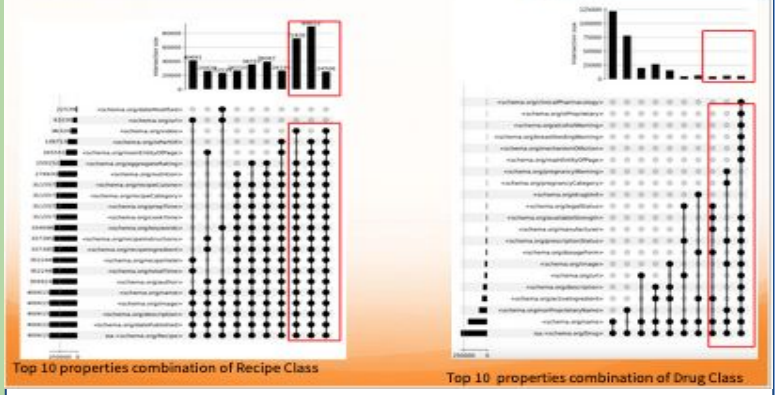
MiKroloG Challenge 1

- 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.

What is inside Microdata ?

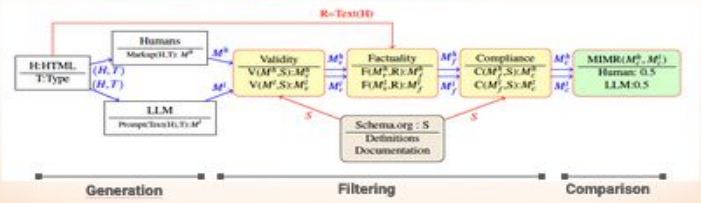
Live demonstration at:
https://schema-obs-demo.onrender.com/

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



Generating Schema.org Knowledge graphs with Large Language Models

Minh Hoang Dang, Thi Hoang Thi Pham, Pascal Molli., Hala Skaf-Molli and Alban Gaignard, *LLM4Schema.org: Generating Schema.org Markups with Large Language Models. Under Review, Semantic Web Journal*



Are LLMs better than Humans in generating Schema.org Knowledge graphs from web pages?

FedShop Synthetic Data Generator

- 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

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.*

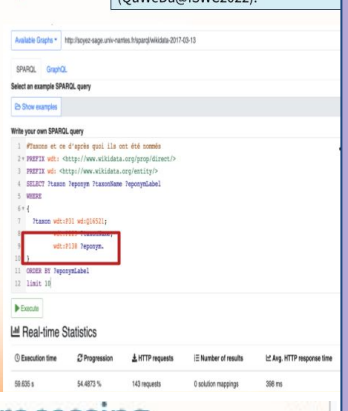
MiKroloG Challenge 2

- 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

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

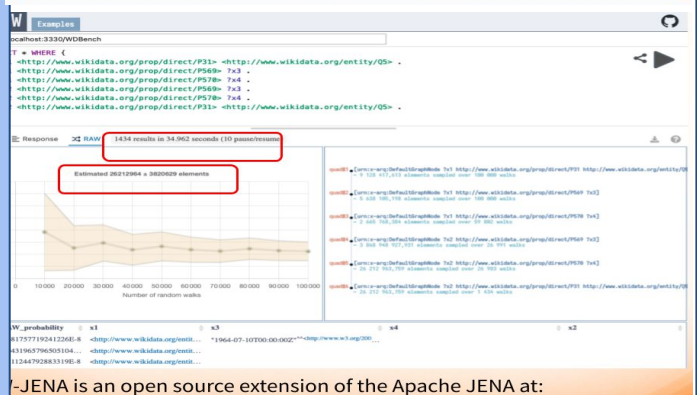


Approximate query processing for SPARQL Servers

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

- RAW-JENA that returns **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
- RAW-JENA integrates sample based approximate query processing directly in **Apache JENA.**

Q_{604} of WDBench [10].



-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

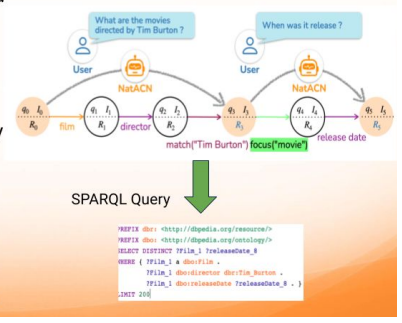
MiKroloG Challenge 3

- 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.

Autopilot to help user querying KGs

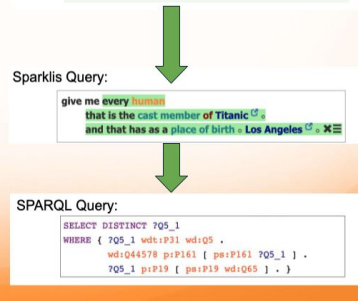
Julie Boudebs, Sébastien Ferré and Peggy Cellier. ***NatACN: a Natural Language Interaction System.*** Poster at GDR TALN 2022.

- 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



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



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