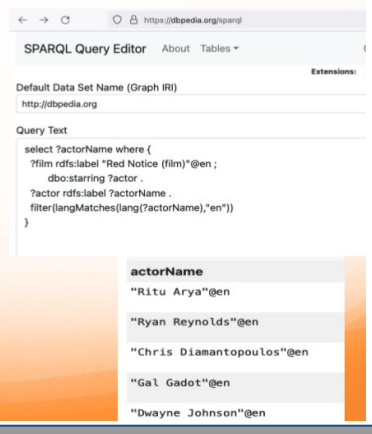
WanderLog: Exploring Large Knowledge Graphs With Sampling

Hala Skaf-Molli, GDD, LS2N
Pascal Molli, GDD, LS2N
Brice Nédelec, ingénieur de recherche
Alban Gaignard, Institut de Thorax
Sébastien Ferré, LACODAM, IRISA
Peggy Cellier, LACODAM, IRISA

Knowledge Graphs

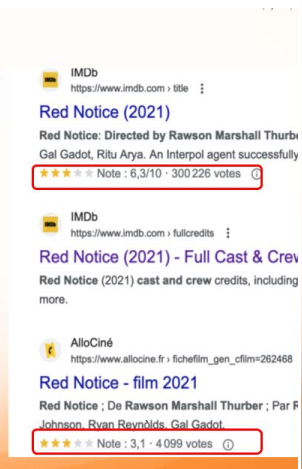
- Allow to answer questions and provide **correct and complete results**
- *Example:Give me the cast members of the film "Red Notice" directed by "Rawson Marshall Thurber".*

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.*



Knowledge graphs on the web-scale are huge and growing

- 86 billion facts collected from 33 million web domains in October 2021 and only 17 billion in 2013.
- Simply counting how many nodes are in the graph can take more than a few hours.
  - making interactive querying of the graph impractical.

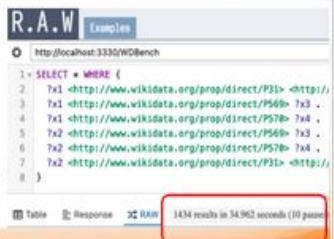
WanderLog Objective

- **Define a sampling execution model for SPARQL queries.**
  - Sampling is performed online by random walks with a fixed budget, guaranteeing users approximate results in a bounded time.

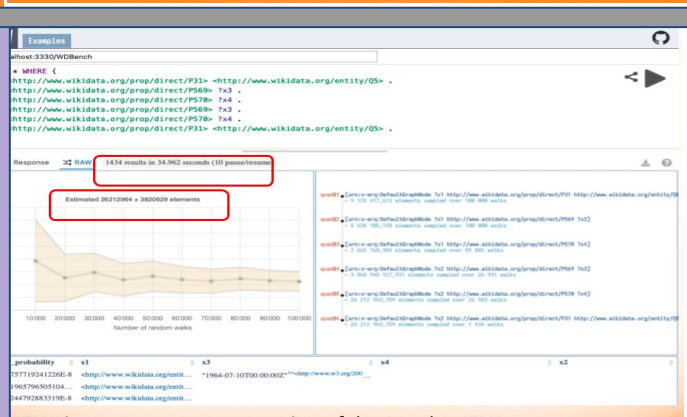
Approximate query processing for SPARQL Servers

- 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.**

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



Q_{604} of WDBench [10].



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

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

Federation engines do not scale

Hand-crafted query plan

	20 endpoints	200 endpoints
RSA	50 ms	1,5s
CostFed	~3s	> 1 hour (timeout)
FedUP	**244ms**	**12,4s**

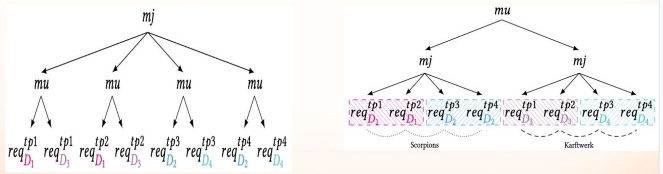
```
SELECT DISTINCT ?product ?localProductLabel WHERE {
    ?localProductXYZ owl:sameAs bsbm:Product43923 .

    ?localProductXYZ bsbm:productPropertyNumeric1 ?origProperty1 .
    ?localProductXYZ bsbm:productPropertyNumeric2 ?origProperty2 .
    ?localProductXYZ bsbm:productFeature ?localProdFeatureXYZ .

    ?localProdFeatureXYZ owl:sameAs ?prodFeature .
    ?localProdFeature owl:sameAs ?prodFeature .

    ?localProduct bsbm:productFeature ?localProdFeature .
    ?localProduct bsbm:productPropertyNumeric1 ?simProperty1 .
    ?localProduct bsbm:productPropertyNumeric2 ?simProperty2 .

    ?localProduct owl:sameAs ?product .
    ?localProduct rdfs:label ?localProductLabel .

    FILTER(bsbm:Product43923 != ?product)
    FILTER(?simProperty1 < (?origProperty1 + 20))
    FILTER(?simProperty1 > (?origProperty1 - 20))
    FILTER(?simProperty2 < (?origProperty2 + 70))
    FILTER(?simProperty2 > (?origProperty2 - 70))
}
```

Similar products in a federation of shops

nh-Hoang Dang, Julien Aimonier-Davat, Hala Skaf-Molli, and Yotlan Le Crom.Fedshop: A benchmark for testing the scalability of arql federation engines. 22th International Semantic Web Conference, ISWC 2023 .

Joins-over-Unions vs Unions-over-Joins

```
SELECT * WHERE {
    ?artist foaf:name ?name .                    #tp1@D1,D3
    ?artist foaf:based_near ?location .          #tp2@D1,D3
    ?location geo:parentFeature ?germany .       #tp3@D2,D4
    ?germany geo:name "Federal...Germany" . }    #tp4@D2,D4
```

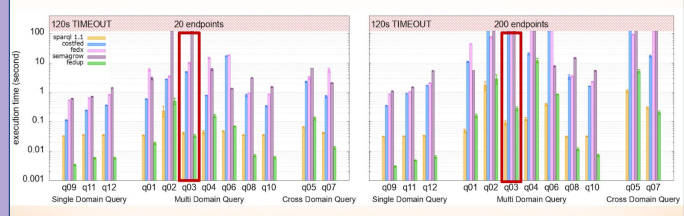


Joins-over-Unions: 8 calls to endpoints !

Unions-over-Joins: 4 calls to endpoints !
However, not in the search space of existing Federation Engines

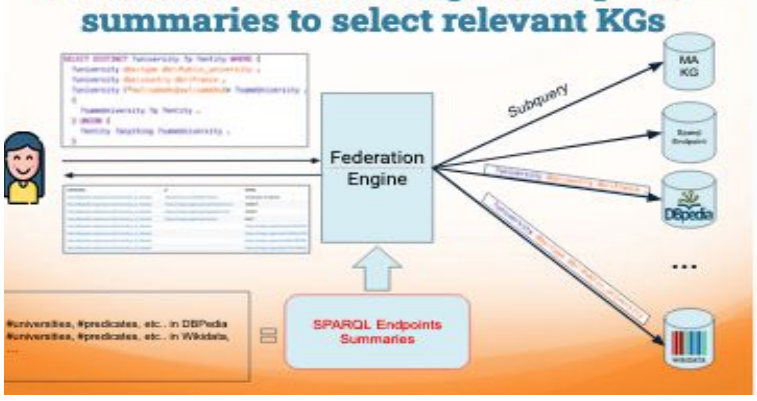
FedUP execution time for FedShop queries

Julien Aimonier-Davat, Brice Nédelec, Minh-Hoang Dang, Pascal Molli and Hala Skaf-Molli. ***FedUP: Querying Large-Scale Federations of SPARQL Endpoints.*** **The ACM WebConf, WWW 2024.**

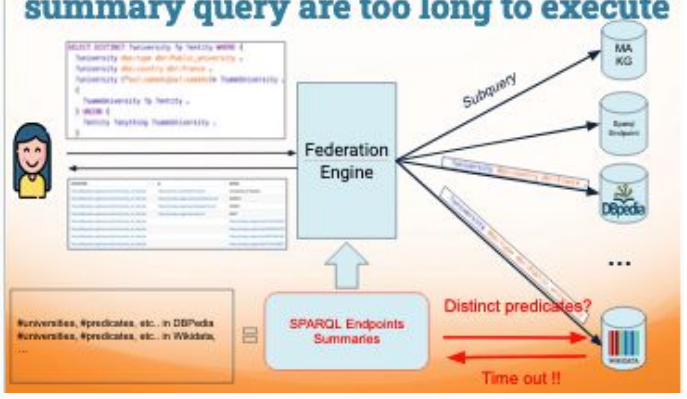


- FedUP outperforms other engines from 1 to 3 orders of magnitude
- For most queries, FedUP is closed to the handcrafted SPARQL 1.1

Efficient federation engines requires summaries to select relevant KGs

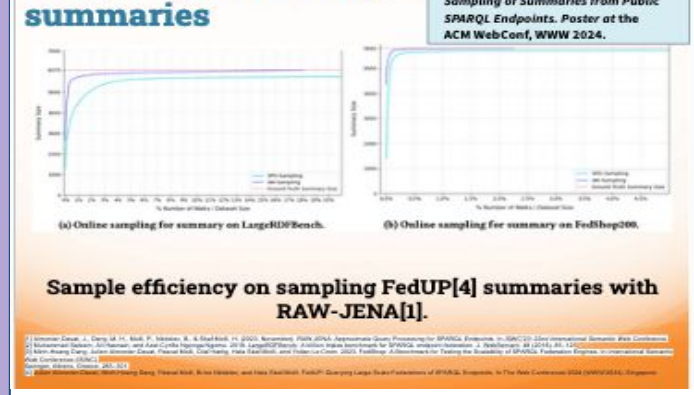


and building summaries online fails as summary query are too long to execute



Sampling for building summaries

Thi Hoang Thi Pham, Hala Skaf-Molli, Pascal Molli, Brice Nédelec ***Online Sampling of Summaries from Public SPARQL Endpoints. Poster at the*** **ACM WebConf, WWW 2024.**

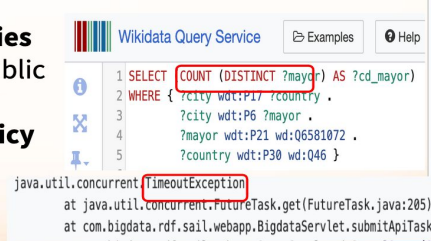


Sample efficiency on sampling FedUP[4] summaries with RAW-JENA[1].

Count-Distinct fails on public knowledge graphs

How many women lead cities in Europe?

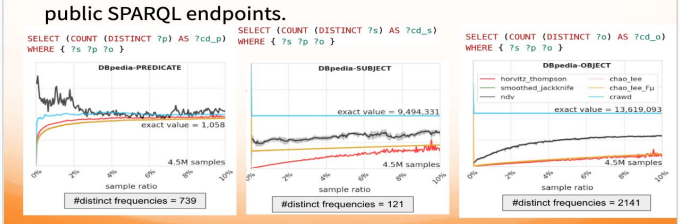
- **Count-Distinct queries** are interrupted by public SPARQL endpoint because **Fair use policy** (60s for Wikidata)
- How to have a **responsive** server
- that **efficiently** handles **Count-Distinct** queries?



Sampling for Count-distinct

Thi Hoang Thi Pham, Hala Skaf-Molli, Pascal Molli, Brice Nédelec ***CRAWD: Sampling-Based Estimation of Count-Distinct SPARQL Queries. International Semantic Web Conference, ISWC 2024.***

- **CRAWD** is a new Unbiased Count-distinct estimator for SPARQL.
- It drastically improves sampling efficiency, even on BGP queries.
- It makes Count-Distinct queries affordable in practice for online public SPARQL endpoints.



The performance gap grows as the gap between #distinct values and #distinct frequencies increases

Next Steps

- **Deploy wikidata with CRAWD** and **RAW-Jena**
  - Provide a responsive server that delivers results for all queries.
    - Create and maintain online summaries for efficient federation engines.
    - Query optimization
    - Source selection
- Time to integration sampling in SPARQL
  - We are starting discussion with the World Wide Web Consortium (W3C)