# Semagrow & KOBE

Antonis Troumpoukis (NCSR-Demokritos)

**COST Action Hackathon Querying Federations of Knowledge Graphs**
April 25-27, 2022, Izmir, Turkey

Semagrow federated query processor

- Semagrow is an open source dynamic data integration system:
    - presents to client applications a single, unified **SPARQL endpoint** that federates multiple data sources.
    - manages both **syntactic** and **semantic** heterogeneity.
- The federated data sources may serve data that use different vocabularies and codelists
    - Semagrow dynamically transforms responses from different sources to match the vocabularies used in the query.
- The federated data sources may offer non-SPARQL APIs (SQL, CassandraQL, etc).
    - Semagrow processes SPARQL queries and appropriately re-writes the sub-queries for each data source.
    - Semagrow fills in the missing expressivity, e.g. arbitrary joins for CQL sources
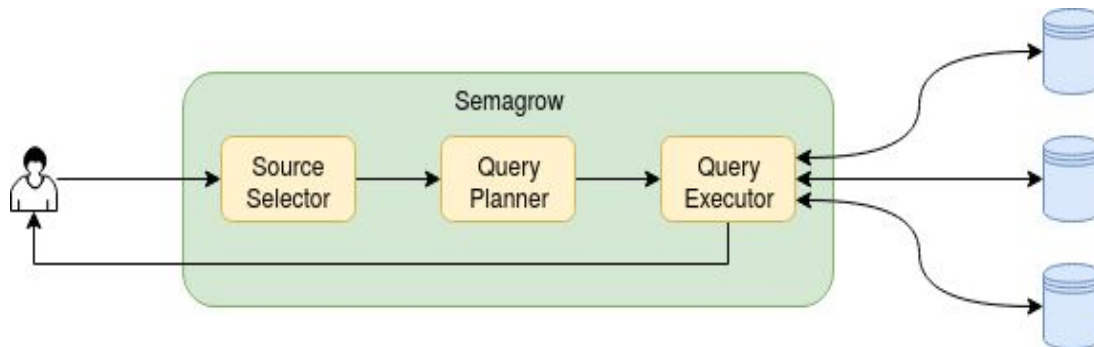
Geospatial extensions of Semagrow
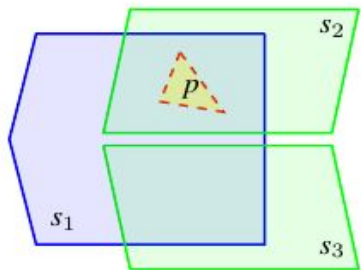
Example: Assume that we are interested in all green objects that their geometry is *within* geometry p. The source selector prunes s_1 because it contains *blue* objects and s_3 because its boundary is geospatially *disjoint from p*.

- A Geospatial source selector [1]
  - Annotate all federated sources with a bounding polygon
  - Use summary to filter out sources that refer to irrelevant areas
  - Geo datasets: divided by areas of responsibility, geographic grids

- A Geospatial Join Optimization for federated geospatial **within-distance** queries [2]
  - Augments subqueries prepared for each source with additional filters s.t. can be answered from the spatial index of the sources.
 - Filters out "too-far away" shapes using constructed rectangles.
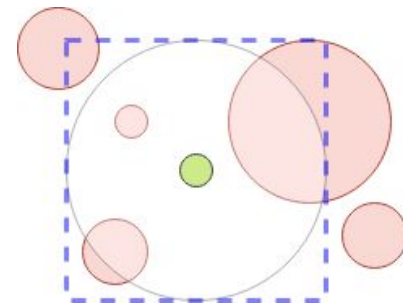
Example: To speed up the process of finding all red shapes within distance d from the given green shape, Semagrow inserts a condition that filters out all shapes that do not intersect with the blue rectangle.

[1] A. Troumpoukis, S. Konstantopoulos, N. Prokopaki-Kostopoulou: A Geospatial Source Selector for Federated GeoSPARQL Querying, Open Res Europe 2022, 2:48
[2] A. Troumpoukis, S. Konstantopoulos, N. Prokopaki-Kostopoulou: A Geospatial Join Optimization for Federated GeoSPARQL Querying, GeoLD2022@ESWC2022
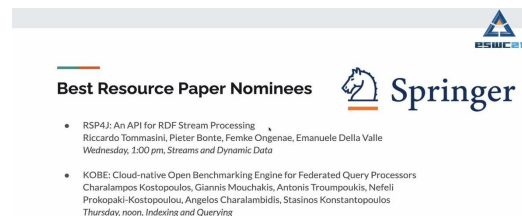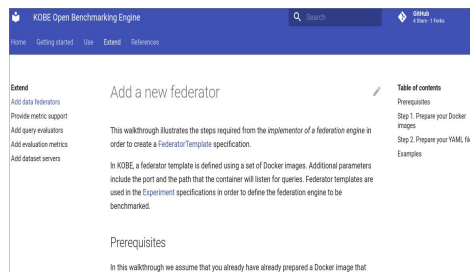
The KOBE Open Benchmarking Engine

- KOBE is a framework for benchmarking federated query engines.

- Features:
  - **Automation of the various tasks:**
    deployment, initialization of dataset servers and federation engines, experiment execution
  - **Reproducibility in different environments:**
    each component in its own Docker container
  - **Declarative specifications:**
    formalism that hides from the user the details of provisioning and orchestrating
  - **Simulating real-life scenarios:**
    network delays (dataset server latency limitations)
  - **Results presentation:**
    collection of logs and visualization in a WebUI
  - **Extensibility:**
    supports the integration of new benchmarks, new federators and new remote dataset servers
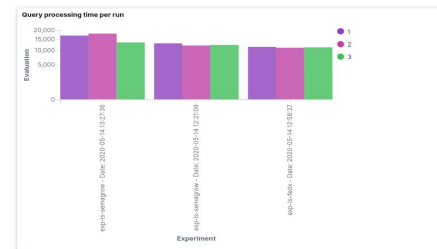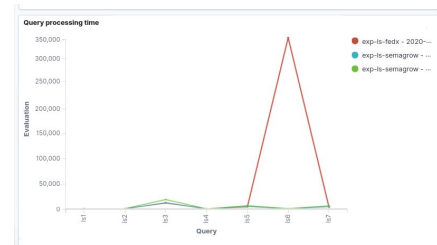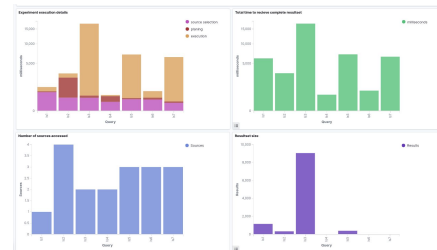
The KOBE Open Benchmarking Engine (cont.)

- *Dataset servers*: Virtuoso, LDF Servers, Strabon, *Federation Engines*: Semagrow, FedX, ??
- *Benchmarks*: Fedbench, LargeRDFBench, OPFbench, Geographica, Geofedbench.
- Detailed Documentation (step by step instructions for getting started, using and extending KOBE). https://semagrow.github.io/kobe/ (publicly available)

[3] C. Kostopoulos, G. Mouchakis, A. Troumpoukis, et al: KOBE: Cloud-Native Open Benchmarking Engine for Federated Query Processors. ESWC 2021
[4 C. Kostopoulos, G. Mouchakis, et al: KOBE: Cloud-native Open Benchmarking Engine for Federated Query Processors. ISWC (Demos/Industry) 2020

Current and Future work

- Geospatial extensions of Semagrow, Raster Data, Geospatial use-cases

- Federate TPF, REST APIs and other less-that-SPARQL data sources

- Integrate Federation engines, Dataset servers, and Benchmarks in KOBE

Thank you!

Visit us at: https://github.com/semagrow/