Geographica: Benchmarking Geospatial RDF Stores

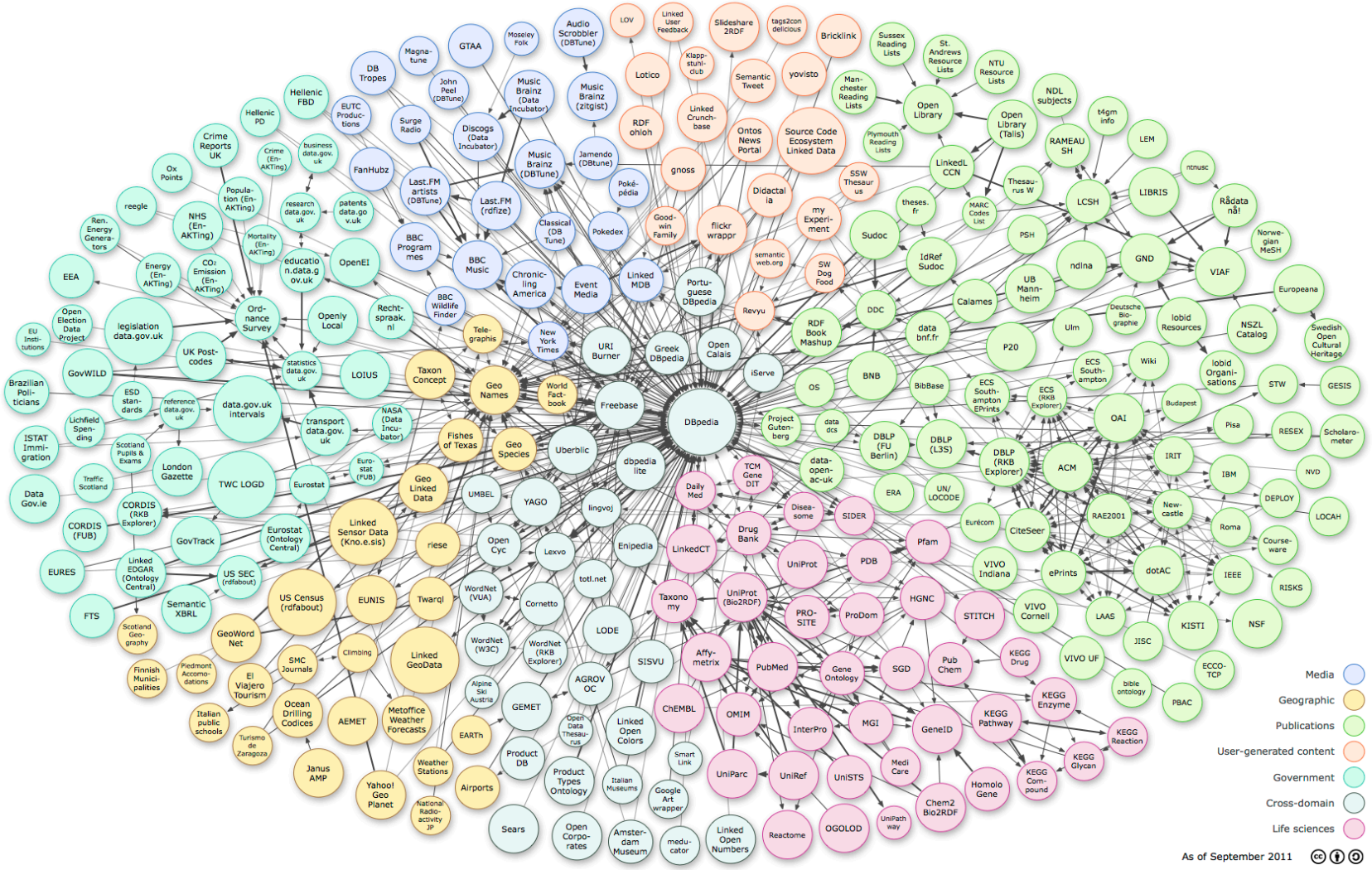
Kostis Kyzirakos
Database Architectures group, CWI

Joint work with
George Garbis and **Manolis Koubarakis**

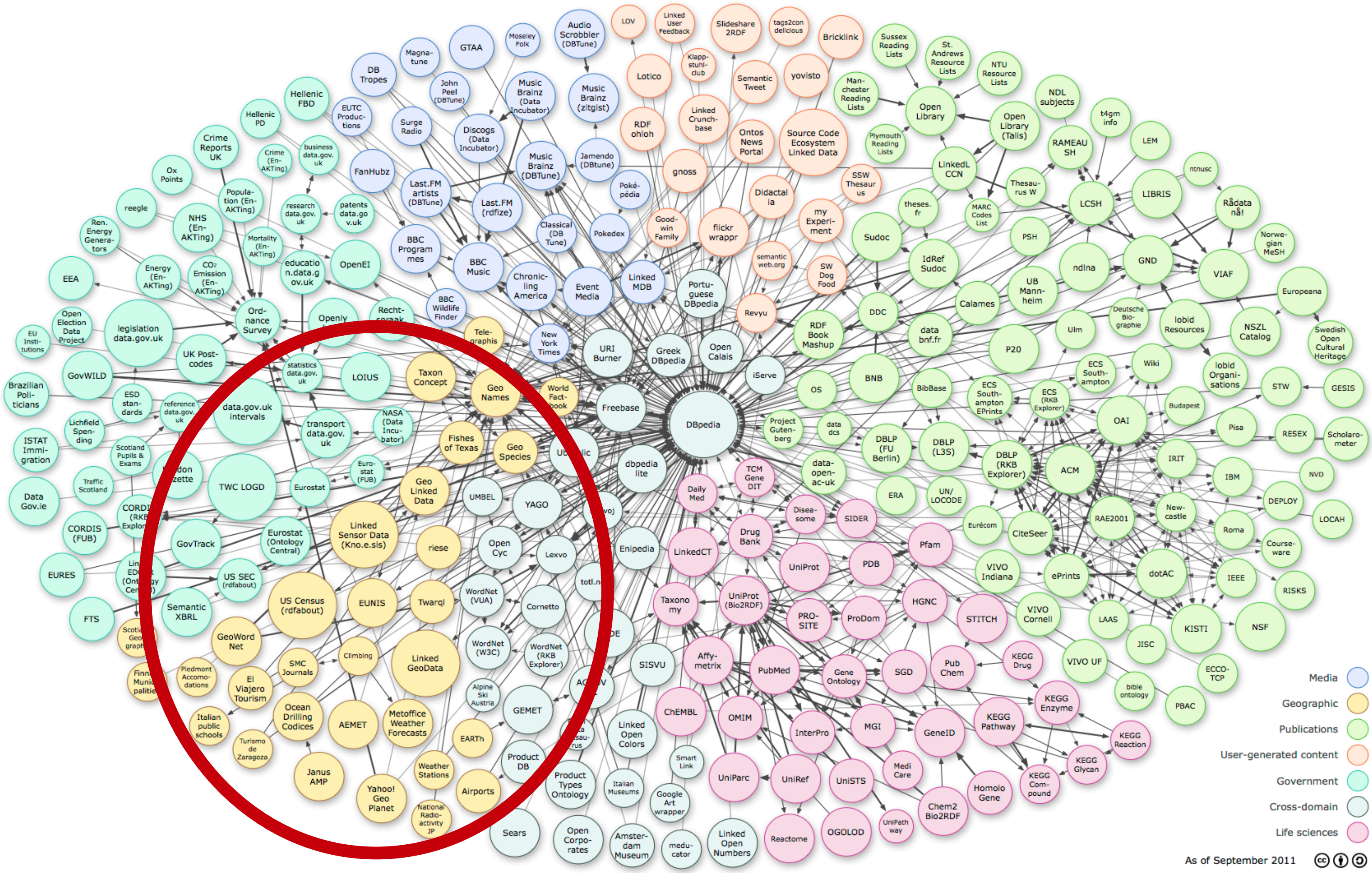
University of Athens
School of Science
Faculty of Informatics and Telecommunications

LDBC TUC, London
November 19, 2013

Linked Open Data Cloud



Linked Open Data Cloud

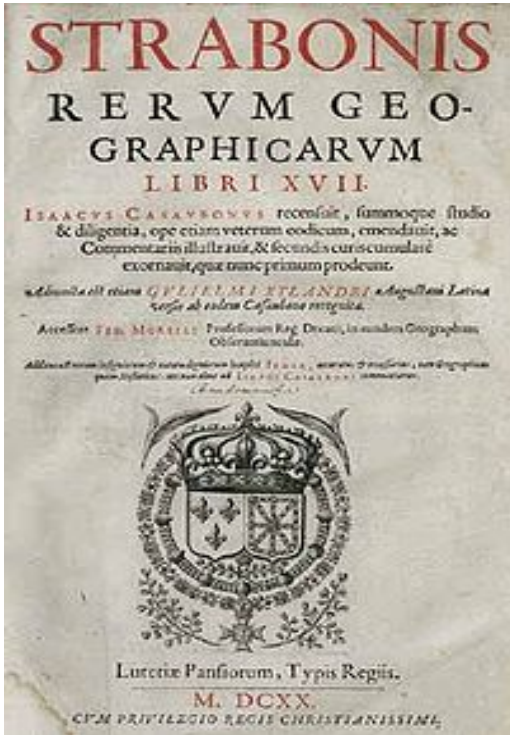


System	Language	Index	Geometries	CRS support	Geospatial Function Support
Strabon	stSPARQL/ GeoSPARQL*	R-tree-over-GiST	WKT / GML support	Yes	• OGC-SFA • Egenhofer • RCC-8
Parliament	GeoSPARQL*	R-Tree	WKT / GML support	Yes	•OGC-SFA •Egenhofer •RCC-8
Oracle 12c	GeoSPARQL	R-Tree, Quadtree	WKT / GML support	Yes	•OGC-SFA •Egenhofer •RCC-8
Brodt et al. **(RDF-3X extension)**	SPARQL	R-Tree	WKT support	No	OGC-SFA
Perry	SPARQL-ST	R-Tree	GeoRSS GML	Yes	RCC-8
AllegroGraph	Extended SPARQL	Distribution sweeping technique	2D point geometries	Partial	•Buffer •Bounding Box •Distance
OWLIM	Extended SPARQL	Custom	2D point geometries	No	•Point-in-polygon •Buffer •Distance
Virtuoso	SPARQL	R-Tree	2D point geometries	Yes	SQL/MM (subset)
uSeekM	GeoSPARQL	R-tree-over GiST	WKT support	No	OGC-SFA

The Benchmark Geographica

- Aim: measure the performance of **today's geospatial RDF stores**

- **Γεωγραφικά:** 17-volume geographical encyclopedia by Στράβων (AD 17)


STRABONIS
RERVM GEO-
GRAPHICARVM
LIBRI XVII.
ISAACVS CASAVBONVS recensuit, summoque studio & diligentia, ope etiam veterum exemplarium, emendavit, ac Commentariis illustravit, & secundis curis cumulatè excoluit, quæ nunc primùm prodeunt.
Lutetiæ Parisiorum, Typis Regiis.
M. DCXX.
CVM PRIVILEGIO REGIS CHRISTIANISSIMI.

Basic GIS Concepts and terminology

Theme: the information corresponding to a particular domain that we want to model. A *theme* is a set of geographic features.

Example: the countries of Europe

Basic GIS Concepts and terminology (cont'd)

Geographic feature or **geographic object**: a domain entity that can have various **attributes** that describe **spatial** and **non-spatial** characteristics.

Example: the country Greece with attributes

- Population
- Capital
- Geographical area
- Coastline
- Bordering countries

The Benchmark Geographica

- Organized around two workloads:
  - **Real-world** workload:
    - Based on existing linked geospatial datasets and known application scenarios
  - **Synthetic** workload:
    - Measure performance in a controlled environment where we can play around with properties of the data and the queries.

Real-World Workload

- **Datasets:** Real-world datasets for the geographic area of Greece playing an **important role in the LOD** cloud or **having complex geometries**
  - LinkedGeoData (LGD) for rivers and roads in Greece
  - GeoNames for Greece
  - DBpedia for Greece
  - Greek Administrative Geography (GAG)
  - CORINE land cover (CLC) for Greece
  - Hotspots

Real-World Workload
Datasets

Datasets	Size	Triples	# of Points	# of Lines (max/min/avg points/line)	# of Polygons (max/min/avg points/polygon)
GAG	33MB	4K	-	-	325 (15K/4/400)
CLC	401MB	630K	-	-	45K (5K/4/140)
LGD (only ways)	29MB	150K	-	12K (1.6K/2/21)	-
GeoNames	45MB	400K	22K	-	-
DBpedia	89MB	430K	8K	-	-
Hotspots	90MB	450K	-	-	37K (4/4/4)

Real-World Workload
Parts

- For this workload, Geographica has two parts:
  - **Micro part:** Tests primitive spatial functions offered by geospatial RDF stores
  - **Macro part:** Simulates some typical application scenarios

Real-World Workload
Micro Benchmark (1/2)

- **29 queries** that consist of **one or two triple patterns and a spatial function.**

- Functions included:
  - **Spatial analysis: `boundary`, `envelope`, `convex hull`, `buffer`, `area`**
  - **Topological: `equals`, `intersects`, `overlaps`, `crosses`, `within`, `distance`, `disjoint`**
    - As used in **spatial selections** and **spatial joins**
  - **Spatial aggregates: `extent`, `union`**

- Functions are applied to many **representative types of geometries** .

Example – spatial selection

- Find all points in Geonames that are contained in a given polygon.

```
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>

SELECT ?s1 ?o1
WHERE {
GRAPH <http://geographica.di.uoa.gr/dataset/geonames>
{ ?s1 <http://www.geonames.org/ontology#asWKT> ?o1 }

  FILTER(geof:sfWithin(?o1,"POLYGON((…))"^^geo:wktLiteral>)).
}
```

Example – spatial join

- Find all pairs of GAG polygons that overlap

```
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>

SELECT ?s1 ?s2
WHERE {
GRAPH <http://geographica.di.uoa.gr/dataset/gag>
{?s1 <http://geo.linkedopendata.gr/gag/ontology/asWKT> ?o1}

GRAPH <http://geographica.di.uoa.gr/dataset/clc>
{?s2 <http://geo.linkedopendata.gr/corine/ontology#asWKT> ?o2}

FILTER( geof:sfOverlaps(?o1, ?o2) )
}
```

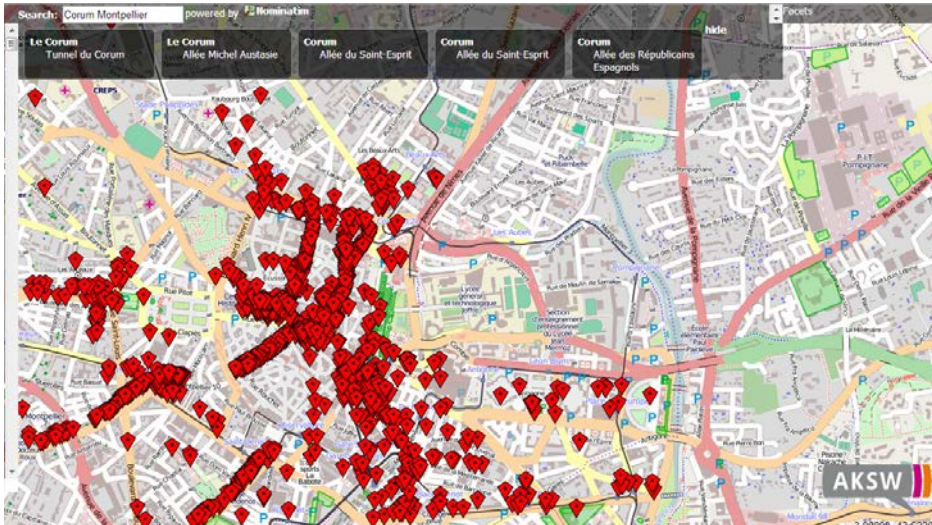
Real-World Workload

Macro part

- **Reverse Geocoding:** Attribute a street address and place to a given point.
  - Queries:
    - Find the closest populated place (from **GeoNames**)
    - Find the closest street (from **LGD**)

Real-World Workload
Macro part

- **Web Map Search and Browsing**
  - Queries:
    - Find the co-ordinates of a given POI based on thematic criteria (from **GeoNames**)
    - Find roads in a given bounding box around these co-ordinates (from **LGD**)
    - Find other POI in a given bounding box around these co-ordinates (from **LGD**)



Real-World Workload
Macro part

- **Rapid Mapping for Fire Monitoring:** representative of typical rapid mapping tasks carried out by space agencies in the case of an emergency

Synthetic Workload

Goal: Evaluate performance in a controlled environment with great precision over the thematic and spatial selectivity of queries.

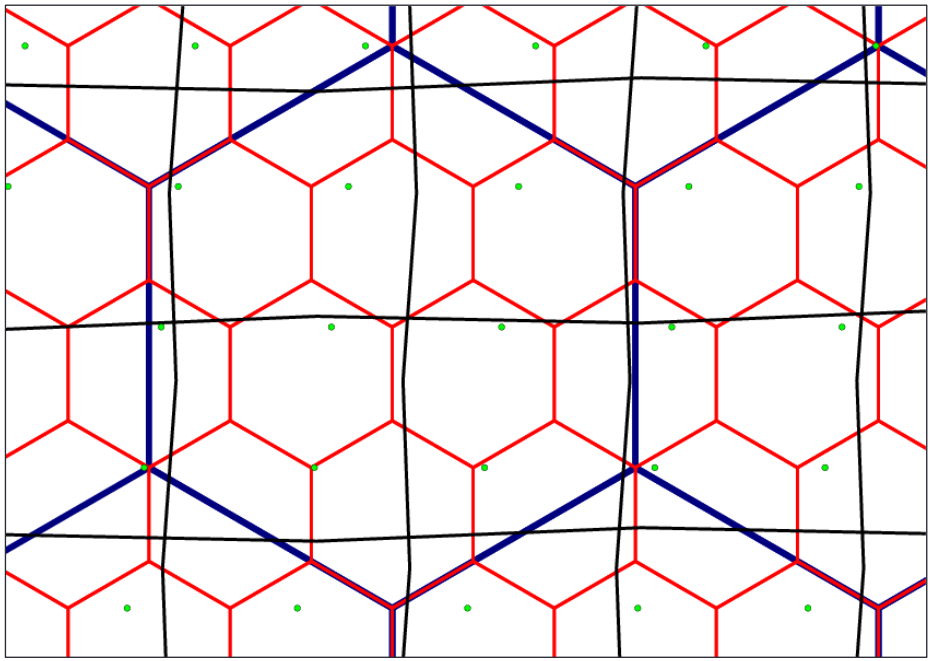
- **Thematic selectivity:** the fraction of the total features of a dataset that satisfy the non-spatial part of the query

- **Spatial selectivity:** the fraction of the total features of a dataset for which the tested topological relation holds

Synthetic Workload

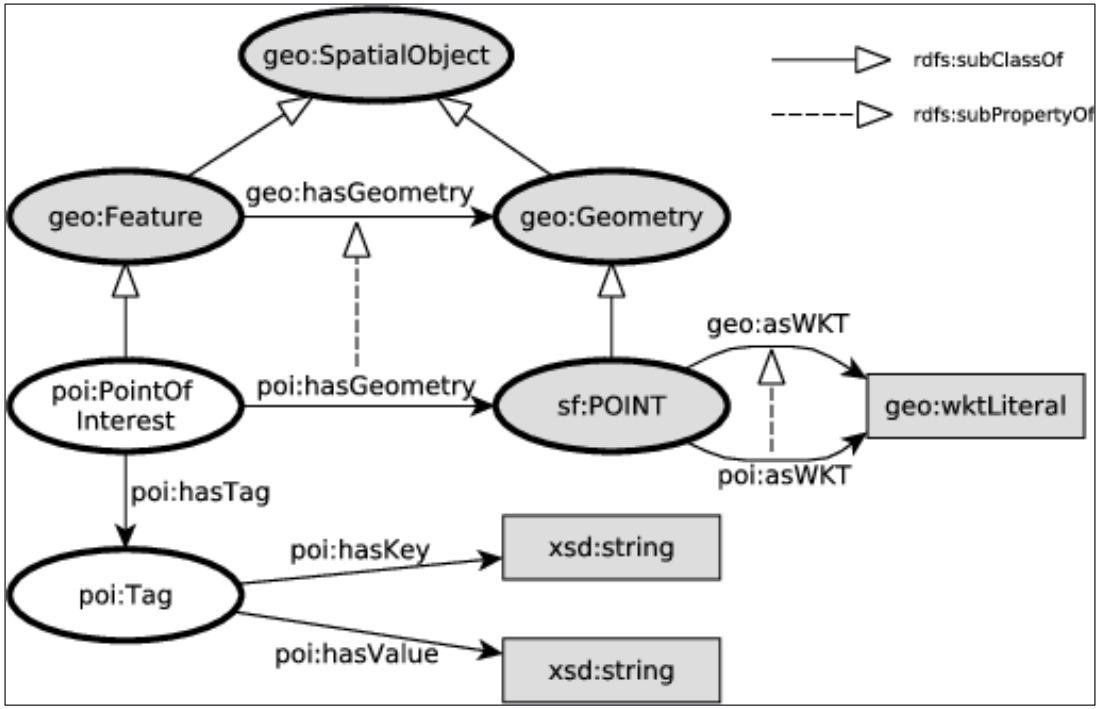
Generator: As in VESPA the produced datasets model features on a map:

- States in a country $((\frac{n}{3})^2)$
- Land ownership (n^2)
- Roads (n)
- POI (n^2)



Synthetic Workload
Ontology

- Based roughly on the ontology of OpenStreetMap and the GeoSPARQL vocabulary
- Tagging each feature with a key enables us to select a known fraction of features in a uniform way



Synthetic Workload
Query templates (1/2)

- Query template for spatial selections

```
SELECT ?s
WHERE {
?s ns:hasGeometry ?g .
?s c:hasTag ?tag .
?g ns:asWKT ?wkt .
?tag ns:hasKey "THEMA" .

FILTER(FUNCTION(?wkt, "GEOM"))}
```

- Parameters:
  - **ns:** specifies the kind of feature (and geometry type) examined
  - **THEMA:** defines the **thematic selectivity** of the query using another parameter **k**
  - **FUNCTION:** specifies the topological function examined
  - **GEOM:** specifies a rectangle that controls the **spatial selectivity** of the query

Synthetic Workload
Query templates (2/2)

- Query template for spatial joins

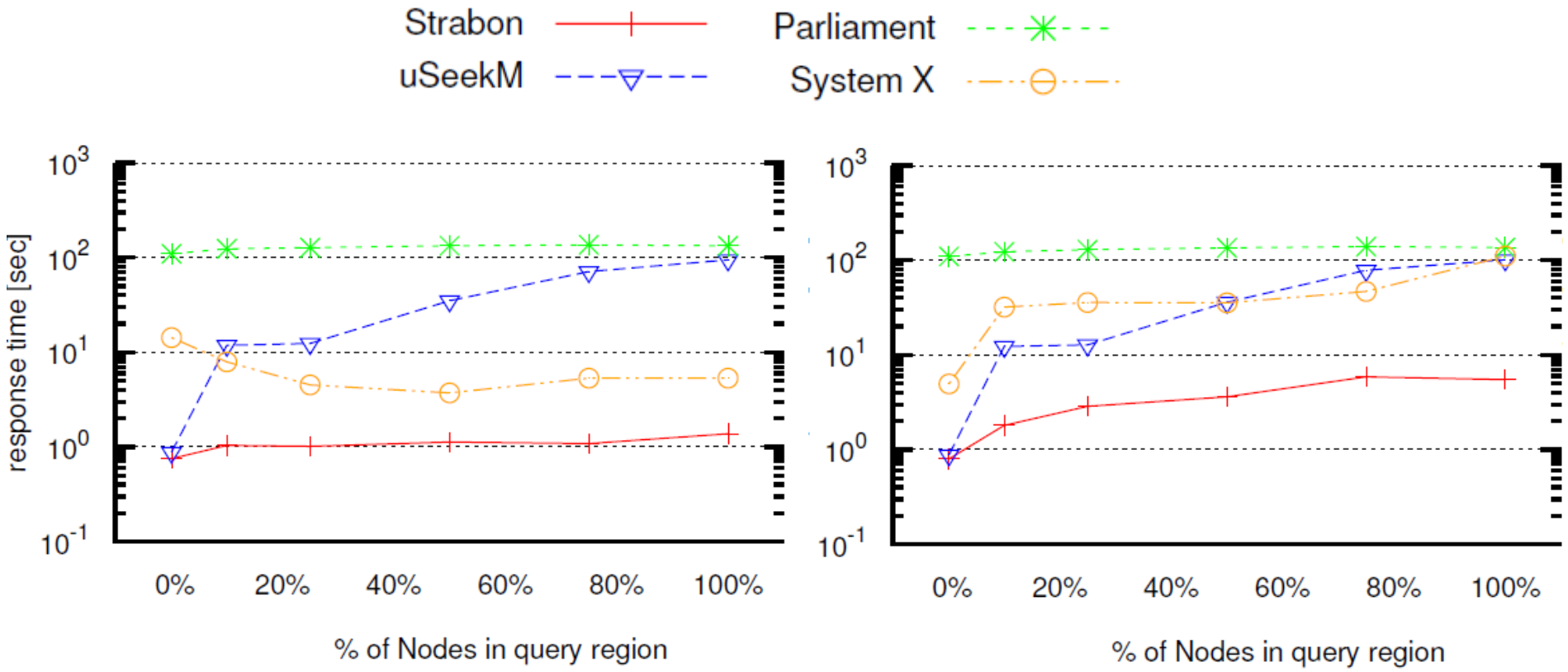
```
SELECT ?s1 ?s2
WHERE {
?s1 ns1:hasGeometry ?g1.
?s1 c:hasTag ?tag1 .
?g1 ns:asWKT ?wkt1 .
?tag1 ns:hasKey "THEMA" .

?s2 ns2:hasGeometry ?g2.
?s2 c:hasTag ?tag2 .
?g2 ns2:asWKT ?wkt2 .
?tag2 ns2:hasKey "THEMA'" .

FILTER(FUNCTION(?wkt1, ?wkt2))}
```

Findings

Current implementations:

- Always start by evaluating the spatial predicate
- Always evaluate spatial predicate at the end
- Take into account both the spatial and thematic selectivity of a query

Results
Synthetic Workload (Spatial Selections)



Intersects
Thematic Selectivity: 0.2%

Intersects
Thematic Selectivity: 100%

Future Work

- Produce a **larger dataset** for the real-world workload.
- **More real-world scenarios** can be added.
- Extend the synthetic generator to produce **non-uniform** datasets.
- Extend the benchmark to include
  - Directional queries
  - Nearest neighbour and reverse nearest neighbour queries
  - Multi-way spatial joins
- Contribute to **standardization efforts**.
  - Develop different profiles of a geospatial benchmark, e.g., point profile.
- Next target: **spatiotemporal RDF stores**
  - **Trajectories**

[GeoSPARQL, OGC standard 2012]

Questions?

[ISWC '13]

http://geographica.di.uoa.gr