Motivation

• Increased interest in publishing geospatial datasets as **linked data** (i.e., encoded in RDF and with semantic links to other datasets)

• Geospatial information might be:
  o **Quantitative** (e.g., exact geometric information)
  o **Qualitative** (e.g., topological relations)

... and express **knowledge** that is

  o **Complete**
  o **Incomplete** (or indefinite)
Ordnance Survey (UK)

SO16 0AS

73,546,231 triples
Global Administrative Areas (GADM)

9,896,532 triples

GADM version 2
2nd level subdivisions
Linked Geospatial Data

As of September 2011

62 billion triples
How do we manage (represent, store, query) this data efficiently?
Challenges: Theory

1. RDF extensions for **representing** and **querying** incomplete qualitative and quantitative geospatial information

- **GeoSPARQL**
  - Standard OGC query language for RDF data with geospatial information
  - Topological relations can be **expressed/queried**, but no reasoning is offered.

- We proposed **RDF**
  - Can work with any topological/temporal constraint language with/without constant symbols (e.g., RCC-5, RCC-8, IA)
  - Formal semantics and algorithm for computing certain answers
  - Preliminary complexity results for various constraint languages

**Open issue**

- **No** published algorithm for query processing when considering RCC-8 **and** constants
RDF\textsuperscript{i} by example

\texttt{gag:Region} rdfs:subClassOf geo:Feature.
\texttt{gag:WestGreece} rdf:type gag:Region.
\texttt{gag:Municipality} rdfs:subClassOf geo:Feature.
\texttt{gag:OlympiaMuni} rdf:type gag:Municipality.
\texttt{noa:Hotspot} rdfs:subClassOf geo:Feature.
\texttt{noa:hotspot} rdf:type noa:Hospot.
\texttt{noa:Fire} rdfs:subClassOf geo:Feature.
\texttt{noa:fire} rdf:type noa:Fire.

\texttt{gag:OlympiaMuni} geo:hasGeometry ex:oGeo.
\texttt{ex:oGeo} rdf:type sf:Polygon.
\texttt{ex:oGeo} geo:asWKT "POLYGON((..))"^^geo:wktLiteral.

\texttt{noa:hotspot} geo:hasGeometry ex:rec.
\texttt{ex:rec} geo:asWKT "POLYGON((..))"^^geo:wktLiteral.

\texttt{gag:WestGreece} geo:sfContains \texttt{gag:OlympiaMuni}.
\texttt{noa:hotspot} geo:sfContains \texttt{noa:fire}.
Query: Find fires inside the region of West Greece.

GeoSPARQL query:

```sparql
CERTAIN SELECT ?f
WHERE {
}
```
Query: Find fires inside the region of West Greece.

GeoSPARQL query:

```
CERTAIN SELECT ?f
WHERE {
}
```
Challenges: Theory

② Efficient computation of the entailment relation

\[ \phi \vdash \theta \]

- where \( \phi \) and \( \theta \) are quantifier-free first-order formulas of a constraint language expressing the topological relations of various frameworks (RCC-8, DE-9IM, etc.)
Challenges: Theory

Computing entailment is equivalent to checking consistency of formulas with constraint networks

- Constraint networks:
  - Spatial relations among regions
  - Regions might be constant ones (exact geometric information) or identified by a URI

- Most recent results considered basic and complete RCC-5 networks with polygonal regions

Open issue

- For RCC-8, deciding consistency is NP-complete
- No published algorithm for checking consistency
- Are there tractable cases?
Challenges: Practice

④ Scale to billions of triples

- Reasoners from QSR scale only up to hundreds of regions with complex spatial relations

How do they perform in our case?

- Setting:
  - Real linked geospatial datasets
  - No constants
  - Only base RCC-8 relations
  - Evaluation of consistency checking using the well-known path-consistency algorithm
Experimental evaluation

- Computation of the **complete** constraint network
- Running time: $O(n^3)$
- Memory requirements: $O(n^2)$

• $n \approx$ thousands to millions

Setup: Intel Xeon E5620, 2.4 GHz, 12MB L3, 48GB RAM, RAID 5, Ubuntu 12.04
Network structure

• We have started working on algorithms taking into account the structure of these networks:
  o Node degrees fit a power-law distribution
  o Network is sparse
Network structure (cont’d)

- Edges of three kinds:
  - non-tangential proper part
  - externally connected
  - equals

- Reflect networks composed of components with hierarchical structure
  - R-tree extensions (Papadias, Kalnis, Mamoulis, AAAI’99)

- Parallel algorithms combined with backward-chaining techniques for lazy query processing
  - Graph partitioning
  - Path compression data structures and indexes
Related work: Spatial

• Qualitative spatial reasoning
  – Efficient algorithms for **consistency checking** of constraint networks (**complex** spatial relations, **few** number of regions)
  – Does not consider **query processing**

• Description logic reasoners
  – **PelletSpatial**: RCC-8 reasoning (cannot handle disjunctions)
  – **RacerPro**: RCC-8 reasoning
Related work: Temporal

- Chaudhuri (VLDB’88)
- The knowledge representation language Telos (TOIS’90)
- Foundations of temporal constraint databases (Koubarakis, PhD thesis, ‘94)
- Qualitative temporal reasoning community (since 80s)
- SQL+i system (BNCOD‘96)
- Later system (IEEE’97)
- Hurtado and Vaisman (2006)
Conclusions

• What’s the **CHALLENGE**?

Implementing an **efficient** query processing system for **incomplete geospatial information** in RDF

• **The desired system** should:
  - **reason** about **qualitative** and **quantitative** spatial information that might be **incomplete**
  - be **scalable** to **billions** of triples in the most useful cases
Thank you
Dataset characteristics

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#triples</th>
<th>#regions</th>
<th>#RCC-8 relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADMGB</td>
<td>149 046</td>
<td>11 762</td>
<td>77 907</td>
</tr>
<tr>
<td>GAG</td>
<td>11 780</td>
<td>412</td>
<td>3023</td>
</tr>
<tr>
<td>NUTS-RDF</td>
<td>316 246</td>
<td>2236</td>
<td>3176</td>
</tr>
<tr>
<td>GADM-RDF</td>
<td>9 896 532</td>
<td>276 728</td>
<td>590 445</td>
</tr>
<tr>
<td>GADM-RDF-EUROPE</td>
<td>355 656</td>
<td>23 037</td>
<td>51 309</td>
</tr>
</tbody>
</table>