OLAP Manipulations on RDF Data following a Constellation Model

Rafik Saad   Olivier Teste   Cássia Trojahn

IRIT (UMR5505) & Universite Toulouse 2 Le Mirail (UTM2), France
srf.rafik@gmail.com,{olivier.teste,cassia.trojahn}@irit.fr

SemStats at ISWC 2013

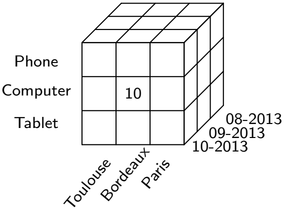
Outline

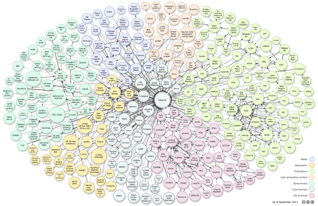
Outline

Context

- Linked Open Data (LOD) as large RDF interlinked data collection
- Emergent need to exploit LOD for analytical analysis
- *Online Analytical Processing* (OLAP) as a potential alternative for manipulating these data : aggregating, summarising and filtering

Multidimensional model





Dimensions : subjects of analysis (Product,Geography,Time)
Hierarchies : axes of analysis (Continent ← Country ← City)

Objectives

- Manipulate OLAP operations on RDF data without any ETL (Extract, Transform, Load) process
- Focus on RDF data described using the RDF Data Cube vocabulary
- Represent multiples hierarchies in a dimension
  - Country ← Region ← City
  - Area ← City
- Take into account the special case of *non-covering* hierarchies (at instance level)
  - Continent ← Country ← City (Toulouse)
  - Continent ← City (Monaco)

Outline

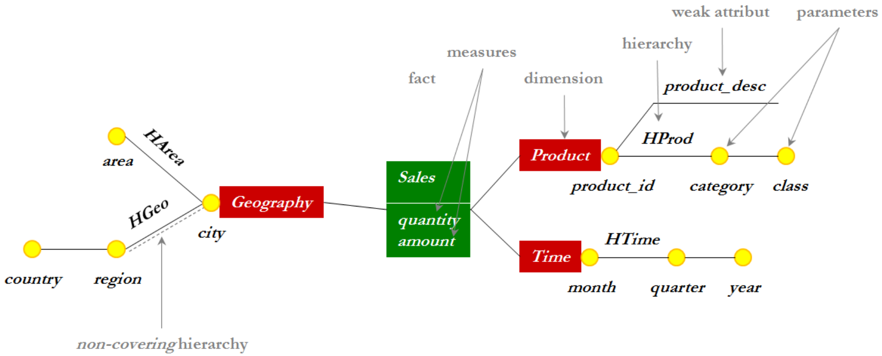
Proposed approach

- Formalise a multidimensional structure
  - following a constellation model [Ravat et al., 2008]
  - where facts and dimensions composed of multi-hierarchies
  - weak attributes complete the information on a hierarchy

- Define a mechanism for translating OLAP operations into SPARQL queries
  - based on a query algebra compliant with the constellation model
  - focus on main OLAP operations (*Drilldown*, *Rollup*, *Select*, *Rotate*)

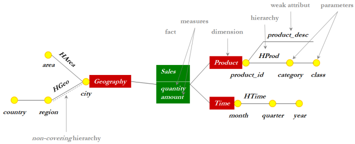
Constellation Model on RDF

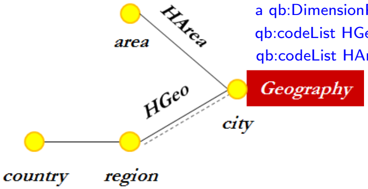
- Constellation schema as conceptual model for defining the elements of a multidimensional model in terms of RDF
- Definition of dimensions *D*, hierarchies *H*, facts *F* and constellation of facts *Cs*, using as basis the vocabularies RDF Data Cube, SKOS and RDFS



Constellation Model on RDF : dimensions

- A dimension models an axis of analysis and contains one or more hierarchies

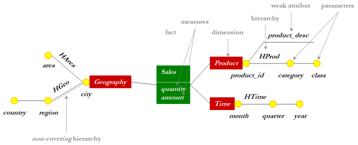


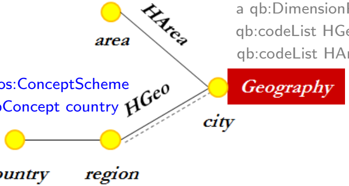


a qb:DimensionProperty
qb:codeList HGeo
qb:codeList HArea

Constellation Model on RDF : hierarchies

- A hierarchy represents levels of granularity from which measures are analysed





a qb:DimensionProperty
qb:codeList HGeo
qb:codeList HArea

a skos:ConceptScheme
a skos:hasTopConcept country

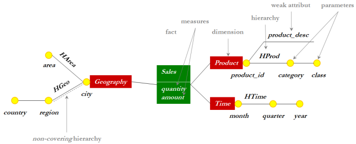
a rdfs:Class
a rdfs:subClassOf skos:Concept
skos:inScheme ex:HGeo
skos:broader ex:region

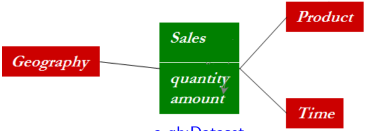
at instance level →

Toulouse a city
Midi-Pyrenes a region
Toulouse skos:broader Midi-Pyrenes

Constellation Model on RDF : facts

- A fact reflects the information to be analysed according to dimensions and measures
- Values of fact instances (observations) correspond to the lowest level of hierarchies for each dimension





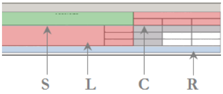
a qb:Dataset
a qb:MeasureProperty

at instance level →

fact1 a qb:Observation
quantity 10
amount 3200
Time 10-2013
Geography Toulouse
Product Computer-1

Translating OLAP Operations into SPARQL

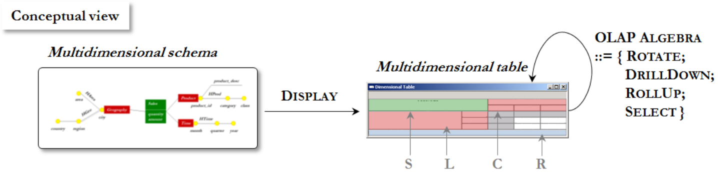
- Mechanism based on a query algebra compliant to the constellation model
- This algebra relies on a graphical multidimensional table (*MT*)

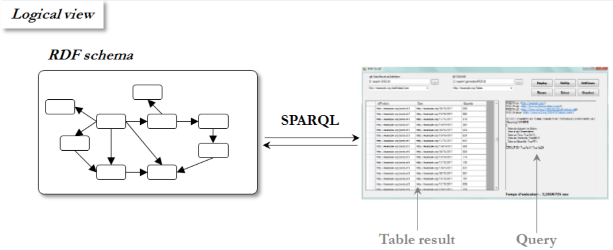


S = represents the analysed subject through facts (aggregations)
L = horizontal analysis axis (dimension)
C = vertical analysis axis (dimension)
R = restrictions of dimensions and fact data (filters)

- Each OLAP operation has an input MT_{SRC} and an output MT_{RES}
- Each MT_{RES} can further be manipulated using operators of the same algebra
- Initial *MT* is built from a constellation *Cs*, using the operator *Display*

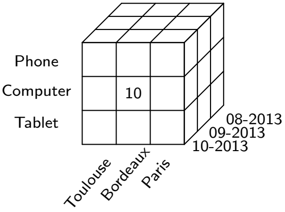
Translating OLAP Operations into SPARQL





Display operation

- Display the root parameters of each hierarchy (lowest level of each dimension hierarchy) :
  1. Identify the fact instances and retrieve root values *PL*1 (horizontal) and *PC*1 (vertical) of the dimensions *DL* and *DC*
  2. Retrieve the value mv_i of each measure m_i and group mv_i by PL_1 and PC_1
  3. Calculate the aggregations by applying to mv_i the aggregation functions Agg_i



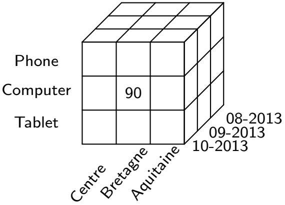
Hierarchy : Country - Region - **City**

SELECT ?PL_1 ?PC_1 ($Agg_i(mv_i)$ AS ?mes_i) WHERE
{
?obs rdf:type qb:Observation.
?obs qb:dataset IRI(F^S).
?obs IRI(DL) ?PL_1.
?obs IRI(DC) ?PC_1.
?obs IRI(m_i) ?mv_i.
}
GROUP BY ?PL_1 ?PC_1

```
SELECT ?prodId ?city (SUM(?qty)
AS qtySales) WHERE
{
?obs rdf:type qb:Observation.
?obs qb:dataset ex:Sales.
?obs ex:Products ?prodId.
?obs ex:Geography ?city.
?obs ex:quantity ?qty.
}
GROUP BY ?prodId ?city
```

Rollup and Drilldown operations

- Provide results by different hierarchical levels
- Use *skos:broader* to navigate between levels
- Recalculates all aggregations from the lowest hierarchical level (no pre-aggregations)



Hierarchy : Country - **Region** - City

```
SELECT ?prodId ?region (SUM(?qty) AS qtySales)
WHERE
{
?obs rdf:type qb:Observation.
?obs qb:dataset ex:Sales.
?obs ex:Product ?prodId.
?obs ex:Geography ?city.
?city skos:broader ?region.
?region skos:inScheme ex:HGeo.
?region rdf:type ex:Region.
?obs ex:quantity ?qty.
}
GROUP BY ?prodId ?region
```

Rollup and Drilldown : non covering hierarchies

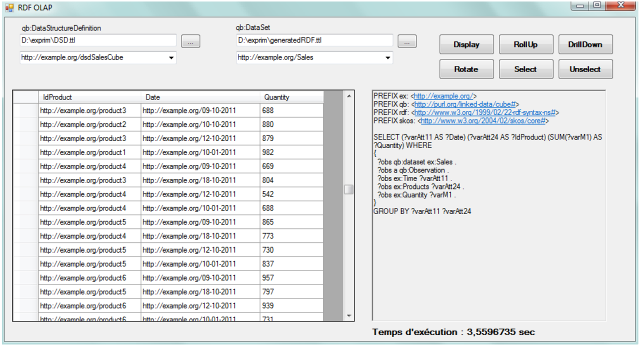
- Not all instances respect the hierarchy
  - Europe ← France ← Toulouse (Europe in level 3)
  - Europe ← Monaco (Europe in level 2)
- May generate wrong aggregation results
- Use UNION operator

```
SELECT ?prodId ?country (SUM(?qty)
AS ?SalesQty)
WHERE
{
?obs rdf:type qb:Observation.
?obs qb:dataset ex:Sales.
?obs ex:quantity ?qty.
?obs ex:Products ?prodId.
?country rdf:type ex:Country.
{
?obs ex:Geography ?geo1.
?geo1 skos:broader ?geo2.
?geo2 skos:inScheme ex:HGeo.
?geo2 skos:broader ?country.
}
UNION
{
?obs ex:Geography ?geo1.
?geo1 skos:broader ?country.
}
UNION
{
?obs ex:Geography ?country.
}
GROUP BY ?prodId ?country
```

Outline

Prototype

- Implemented using Microsoft .NET framework and dotNetRDF API



Outline

Experiments and evaluation

- **Data sets**
  - Real data : Annual producer price of industrial products from CA 1996 Statistical Office of the Republic of Serbia
    - 789 instances of attributes and 156 observations
    - 1 temporal and 1 geographical (about Serbia) dimensions
    - Observations modified for generating non-covering instances data
  - Synthetic data
    - 69888 observations and 1191 instances of attributes
- **(Very) Preliminary evaluation**
  - Limited to manipulations of OLAP operations on these two data sets
  - Adequacy and correctness of results from a sequence of operations
  - No measurements on runtime

Outline

Conclusions and perspectives

- Formalisation of a constellation model in terms of RDF data
- Mechanism for translating OLAP operations into SPARQL
- Handling special cases were hierarchies are not fully covered at instance level
- Perspectives : improve visualisation, expressing advanced OLAP operations, query optimisation and pre-aggregations, exploitation of links (data across data sets).

References

Ravat, F., Teste, O., Tournier, R., and Zurfluh, G. (2008).
Algebraic and graphic languages for olap manipulations.
IJDWM, 4(1):17–46.