Cloud Based Linked Data Management and INSPIRE

Dimitris Kotzinos
FORTH-ICS & University of Cergy - Pontoise
Yannis Roussakis, Kyriakos Kritikos, Athina Kritsotaki, Martin Doerr
FORTH-ICS
Problems to tackle

- Data integration
- Data retrieval in a uniform way
- Data export to other formats (and INSPIRE)
Information Integration

- The **Web of data** becomes a reality by connecting so far isolated islands or data silos such as
  - Enterprise silos (Disparate DBMS Engines & Development Frameworks)
  - Social Media silos (Social Networking Services, Discussion Forums, Blogs, Wikis etc.)
  - Scientific silos (in biology, physics, earth sciences, etc.)

- **Linked Open Data (LOD)** is a way of publishing data on the (Semantic) Web that:
  - Encourages reuse
  - Promotes its (real & potential) inter-connectedness
  - Enables network effects to add value to data

Source: [flickr](http://www.flickr.com/photos/docsearls/5500714140/)
Linked Open Data as Service

- Abstraction layer for data access
  *abstract the applications from the specific setup of the data management service (such as local vs. remote, federation, and distribution)*

- Beyond Data Access
  - Enabling automation of discovery, composition, and use of datasets
  - Data Markets
  - Online Visualization Services
  - Data Publishing Solutions
  - Data Aggregators
  - BI / Analytics as a Service
Data Exports

• Query results viewed as:
  – RDF/XML
  – JSON
  – GML
  – KML
  – Shapefiles

• Transformation services to user provided conceptualizations

• Export to INSPIRE compliant XML formats
Introduction on Available Datasets in the project

• In this project we deal with scientific data described in different formats and notions from diverse areas
  – ground waters, boreholes, chemical analyses, etc.
    • GEUS, EKBAA, BRGM
  – earthquakes, recordings, stations, etc.
    • EPPO
  – Landslides, susceptibility maps, etc.
    • GEO-ZS, EKBAA

• Purpose: integrate, describe and query such heterogeneous data in a uniform way
Overall Approach

- Creation of a Conceptual Model (GSOM) to integrate and cover all the thematic fields
- Map the source data into RDF data compliant to the Conceptual Model
- Provide unique URIs for data
- Rely on a scalable RDF Triple Store to enforce the mappings and enable the storage and query of the RDF data
- Provide an API for the RDF data management
Models, Mappings and Integration

Providers’ Data

- Structured data (spreadsheets)
- Structured data (XML documents)
- Structured data (relational tables)

GSOM mappings

query

results

transform

tsv/csv
xml
json

export

xml
trig
trix
n3
...

formats

GeoJSON
KML
Shapefile
...

Inspire compliant

Inspire Workshop, Aalborg 16 June 2014
Our model:

- needs to be a generic, extendable and interoperable, which relies not only to geographic data
- Integrates different datasets created from sources w.r.t. their semantics
- developed for the documentation of scientific observation – measurement and sampling

An event-centric model that captures complex contexts where time, place and participants are associated

Capable to describe complex relationships – to integrate or connect rich information
Data Mappings to Conceptual Model

- The data providers’ relational data were informally mapped into our Conceptual Model
- Such mappings can be formally expressed using R2RML
  - R2RML is a W3C Recommendation
  - [http://www.w3.org/TR/r2rml/](http://www.w3.org/TR/r2rml/)
  - Enables the automatic creation and synchronization of the Linked Data w.r.t. the Relational Data
• **INSPIRE** (Infrastructure for Spatial Information in Europe) covers 34 Spatial Data Themes. It uses ISO 19100 series standards; oriented to *environmental* and *spatial* data

• The Generic conceptual schema is *not event-oriented* (especially the schema that uses the observation-measurement specification)
Why do we not implement INSPIRE directly?

- INSPIRE is missing some information integration aspects since the different themes are managed independently.
- Querying information on integrated views of actors, activities, places and items of interest (e.g. boreholes) can be very cumbersome at times ...
- Missing modeling in a semantic-enabled language like RDF/S so that we can post meaningful queries directly and link with other datasets.
- INSPIRE becomes very big and difficult to manage at some points.
- Some themes like Earthquakes and Landslides are not rich enough.

Inspire Workshop, Aalborg 16 June 2014
Export Linked Data into INSPIRE

- We created a GSOM-to-INSPIRE mapping mechanism which allows to transform GSOM-based RDF data, acquired from SPARQL queries, into XML-based INSPIRE-compliant data.
- Exported Data are made INSPIRE-compliant.
- However INSPIRE has limitations which do not allow for a complete export of all the GSOM notions (i.e., we cannot express all the GSOM notions in INSPIRE notions).

*Inspire Workshop, Aalborg 16 June 2014*
Approach for INSPIRE-based Exporting (I)

• Separation of our data (from all data providers) into themes
  – chemical analyses, boreholes, landslides, earthquakes etc.

• Investigation of the GSOM notions which could be mapped to inspire notions as only these notions will be eventually exported

• Creation of a set of API methods which export Linked Data into an INSPIRE compliant format according to two main alternative ways
Approach for INSPIRE-based Exporting (II)

• First export way:
  – The user can select one or more themes which interest him in order to export their respective data
  – Next, the user may choose to export all the data which correspond to the selected themes or he can apply SPARQL filter conditions for a selective export
  – Finally, the service `inspire_export` can be called with the above parameters and the INSPIRE compliant data are created
Second export way:

- Instead of selecting specific themes the user can provide the appropriate SPARQL query related to the data he/she desires to obtain to the *inspire_query_export* API method.
- This method, by knowing all GSOM-to-INSPIRE mappings, maps the SPARQL query variables to INSPIRE-mapped GSOM notions and makes the corresponding result transformation to INSPIRE.
- This method is appropriate for users which are familiar with GSOM and of course with SPARQL.
Approach to INSPIRE-based Exporting (IV)

• Second export way:
  – Can also be used to obtain related results which span different themes and different data providers – by allowing the user to provide a SPARQL query we enable the description of any data requirement
  – Due to INSPIRE incompleteness issues raised previously, not all SPARQL query variables are guaranteed to be mapped into INSPIRE-compliant information. On the contrary, the first INSPIRE export method provides a more controlled way to produce the required INSPIRE-compliant information (all SPARQL variables are mapped to INSPIRE information)
LINKED DATA API & IMPLEMENTATION
Virtuoso Universal Server is a hybrid of DB engine & middleware combining the func. of a RDBMS, ODBMS, and a RDFStore

- Supports various Internet & Web protocols:
  - HTTP(s), WebDav, SOAP, UDDI, SPARQL

- Implemented various data access APIs:
  - ODBC, JDBC, OLE DB, ADO . NET

- Provides a Web-based DB Administration UI

- Provides an open-source version

- Can create a clustered-based system of RDFStores + a non-free version available in the Cloud
Linked Data Storage - Virtuoso
• Reasons for selection:
  – Scalable to handle millions of triples + exhibits a very good performance (~1B triples in this project)
  – Offers API for RDF data management
  – Supports SPARQL and SPARUL
  – Supports 2-D features and offers spatial operators
  – Supports R2RML and the respective relational-to-RDF data mapping mechanism
  – Offers a cloud-based version
Linked Data Storage - Virtuoso

• Scalability / Elasticity
  – Planned at least 2 instances of Elastic LD Storage to be available in the platform to handle the query load (cross-provider queries are demanding)
  – The elastic capabilities of Amazon Cloud are exploited to change the number of instances when the load surpasses or goes below a specific threshold
  – Stress tests were conducted to determine these upper and lower thresholds
Linked Data Storage and Querying

• Successfully inserted ~1B triples into Virtuoso
• Created specific queries
  – For each partner’s dataset
  – Combining similar notions from cross provider datasets
• Relative small execution times depending on the number of fetched results + query complexity
• At the moment using only 1 instance in the Amazon Cloud but we have also used an elastic group of instances along with a load balancer
• Created the R2RML mappings for ground water datasets
Conceptual Data Integration & Linking API

• Methods for:
  – Querying data/metadata via SPARQL (both provider-specific & cross-provider queries are supported)
  – Importing data/metadata in various formats (both blocking and non-blocking depending on the size of meta-data to be imported)
    • Direct through respective API method
      – Data provider should be responsible of synchronizing his/her relational and RDF data
    • Indirect through the API method for the definition of relational-to-RDF mappings
      – Apart from the definition of mappings, the data provider does not need to perform any kind of synchronization, as the system will be responsible for performing such a synchronization on his/her behalf
  – Exporting data/metadata in various formats (inline in the response or available in a specific URL)
  – Updating data/metadata via SPARUL
Demo Scenario

• Given a specific data provider, we will use a set of the API’s methods in order to:
  – Transform the original data into Linked Open Data and publish them on the cloud
  – Query the Linked Open Data
  – Export them into INSPIRE
Demo Scenario – Import (I)

- EKBAA provider – Borehole data and Groundwater sample analyses
  - Original data are stored in relational tables
  - Ground waters sample takings taken from the regions of Mygdonia, Thriasio, etc. in Greece
  - Chemical analyses results over sample takings
  - Information about the sample taking date and origin (e.g., boreholes, springs, etc)
### Demo Scenario – Import (II)

#### Chem. Analysis

| CODE | SDATE | pH | COND | Ca | Mg | Na | ... | V | Be | ...
|------|-------|----|------|----|----|----|------|---|----|-----

- **Borehole code**
- **Sampling date**
- **Chemical substances (mgr/lt or μgr/lt)**

#### Borehole

<table>
<thead>
<tr>
<th>CODE</th>
<th>PCODE</th>
<th>TYPE</th>
<th>LATITUDE</th>
<th>LONGITUDE</th>
<th>CODE_WDIST</th>
<th>NAME_WDIST</th>
<th>CODE_AQUIF</th>
<th>NAME_AQUIF</th>
<th>DEPTH</th>
<th>DRILL_DIAM</th>
<th>GEOLOGY</th>
<th>....</th>
</tr>
</thead>
</table>

- **Borehole code**
- **Borehole name**
- **Sampling source**
- **Location**
- **Water district**
- **Aquifer**
- **Soil composite.**

*Inspire Workshop, Aalborg 16 June 2014*
Demo Scenario – Import (III)

Inspire Workshop, Aalborg 16 June 2014
Demo Scenario – Import (IV)

- The informal Relational-to-Linked data mappings have to be formalized using R2RML language
- The final step is to call a specific API method (addR2RMLMappings), to enforce the R2RML mappings and create the corresponding Linked Data
- As a result:
  - the original Relational Data are transformed into Linked Data w.r.t. the R2RML mappings
  - There exists an one-way synchronization from the Relational Data to Linked Data
The transformation of providers’ data into Linked Data enables the evaluation of uniform queries over the underlying triple store (Virtuoso) which can be of the following types:

- Provider-specific queries
- Cross-theme or cross-provider queries

Queries are expressed via the W3C’s SPARQL language (http://www.w3.org/TR/rdf-sparql-query/)
Demo Scenario – Query (II)

• **Query Description:** Select the samples of the compounds with ph>7 and Na>50. Moreover, select the locations and the boreholes (names) which were analyzed.

• This query is either applied over a specific dataset (EKBAA) or over all the datasets (EKBBA, GEUS, BRGM).

• Query is expressed in SPARQL and then *ldquery* API method is called.
Demo Scenario – Query (III)

**Underlying SPARQL query**

```
where {
?st sci:05_removed ?sample;
  sci:04_sampled_at ?bhole.
  filter(regex(?sample, "GrWater/"))
?meas crm:P39Measured ?sample;
  crm:P40_ExpectedDimension ?cdim.
?cadim crm:P2_hasType ?compound;
  crm:P90_hasValue ?cvalue.
?bhole a sci:S16_Borehole;
  sci:07_consists_of ?bcollar;
  crm:P1_isIdentifiedBy ?bname.
  filter(regex(?bname, "Name"))
?bcollar geo:hasGeometry ?point.
?point geo:asWKT ?asWKT.
filter(regex(?asWKT, "/4326"]).

  filter (regex(?compound, "pH") && ?cvalue > 7) ||
  (regex(?compound, "Na") && ?cvalue > 100).
```

Select the samples and chemical analyses of the compounds with \( \text{pH} > 7 \) and \( \text{Na} > 100 \). Moreover select the locations and the boreholes(names) which were analyzed.

Inspire Workshop, Aalborg 16 June 2014
<table>
<thead>
<tr>
<th>sample</th>
<th>compound</th>
<th>value</th>
<th>bname</th>
<th>asWKT</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample</td>
<td>compound</td>
<td>cvalue</td>
<td>bname</td>
<td>asWKT</td>
</tr>
<tr>
<td>-------------------------</td>
<td>----------</td>
<td>--------</td>
<td>----------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------</td>
</tr>
<tr>
<td><a href="http://servicesoads.eaufrance.fr/GrWater/4036641">http://servicesoads.eaufrance.fr/GrWater/4036641</a></td>
<td>pH</td>
<td>7.1</td>
<td><a href="http://servicesoads.eaufrance.fr/Borehole/Name/FORAGE">http://servicesoads.eaufrance.fr/Borehole/Name/FORAGE</a> DU CUF DE BŒUF</td>
<td><a href="http://www.opengis.net:def/crs/EPSG/0/4326">http://www.opengis.net:def/crs/EPSG/0/4326</a> Point(5.08089810166797 45.5181359777231)</td>
</tr>
<tr>
<td><a href="http://servicesoads.eaufrance.fr/GrWater/4037516">http://servicesoads.eaufrance.fr/GrWater/4037516</a></td>
<td>pH</td>
<td>7.2</td>
<td><a href="http://servicesoads.eaufrance.fr/Borehole/Name/FORAGE">http://servicesoads.eaufrance.fr/Borehole/Name/FORAGE</a> DU CUF DE BŒUF</td>
<td><a href="http://www.opengis.net:def/crs/EPSG/0/4326">http://www.opengis.net:def/crs/EPSG/0/4326</a> Point(5.08089810166797 45.5181359777231)</td>
</tr>
<tr>
<td><a href="http://servicesoads.eaufrance.fr/GrWater/4038347">http://servicesoads.eaufrance.fr/GrWater/4038347</a></td>
<td>pH</td>
<td>7.1</td>
<td><a href="http://servicesoads.eaufrance.fr/Borehole/Name/FORAGE">http://servicesoads.eaufrance.fr/Borehole/Name/FORAGE</a> DU CUF DE BŒUF</td>
<td><a href="http://www.opengis.net:def/crs/EPSG/0/4326">http://www.opengis.net:def/crs/EPSG/0/4326</a> Point(5.08089810166797 45.5181359777231)</td>
</tr>
<tr>
<td><a href="http://servicesoads.eaufrance.fr/GrWater/4039197">http://servicesoads.eaufrance.fr/GrWater/4039197</a></td>
<td>pH</td>
<td>7.3</td>
<td><a href="http://servicesoads.eaufrance.fr/Borehole/Name/FORAGE">http://servicesoads.eaufrance.fr/Borehole/Name/FORAGE</a> DU CUF DE BŒUF</td>
<td><a href="http://www.opengis.net:def/crs/EPSG/0/4326">http://www.opengis.net:def/crs/EPSG/0/4326</a> Point(5.08089810166797 45.5181359777231)</td>
</tr>
<tr>
<td><a href="http://servicesoads.eaufrance.fr/GrWater/4039726">http://servicesoads.eaufrance.fr/GrWater/4039726</a></td>
<td>pH</td>
<td>7.1</td>
<td><a href="http://servicesoads.eaufrance.fr/Borehole/Name/FORAGE">http://servicesoads.eaufrance.fr/Borehole/Name/FORAGE</a> DU CUF DE BŒUF</td>
<td><a href="http://www.opengis.net:def/crs/EPSG/0/4326">http://www.opengis.net:def/crs/EPSG/0/4326</a> Point(5.08089810166797 45.5181359777231)</td>
</tr>
</tbody>
</table>
Demo Scenario – Results Visualization (I)

- As the results are provided in a particular machine-processable form, user can request a better visualization by transforming them to a format which can be loaded on various GUI tools/services, such as Google Maps.
- To this end, if the SPARQL results contain spatial information, the API offers the `geoldtransform` method through which they can be transformed into geographical feature collection representations, such as KML, GeoJSON, GML and Shape.
Demo Scenario – Results Visualization (II)

Result No.: 115
bname: ΜΣ25
phVal: 8.12
longitude: 23.059280
latitude: 40.791640
depth: 110.0

Οδηγίες Αναζήτηση στην γύρω περιοχή Αποθήκευση σε... περισσότερα

1 των 4 κοντινών αποτελεσμάτων Επόμενο »
Demo Scenario – Results Visualization (III)
Demo Scenario – INSPIRE Export (I)

- Recall that the dataset refers to Boreholes, Samples and Chemical Analyses
- The relevant themes which can be exported are those of: **boreholes, chemical analyses**
- The exporting scenario considers all theme-related data even from other data providers (the latter data can be filtered by providing filtering constraints)
• Recall the query shown in the previous section
  – Select the samples of the compounds with pH>7 and Na > 100. Moreover select the locations and the boreholes(names) which were analyzed.

• Export INSPIRE data from chemical analyses w.r.t. the above condition parameters
  – **Case 1**: Call first INSPIRE export method with parameters:
    • themes: **Chemical Analysis**
    • constraints: (compound = “pH” and value > 7) or (compound = "Na" and value > 200) or compound = “.”
Case 2: Call second INSPIRE export method with parameters:

- query: the respective query that produces the equivalent results

• Result: a zipped file containing the xml specification of the chemical data is created w.r.t. the appropriate XSD specifications of INSPIRE
Usage and Implementation (I)

- Smart Queries
  - Tools for data providers to add and edit SPARQL queries
  - Tools for data retrieval through a filtering providing interface (using the already inserted queries)
Usage and Implementation (II)

- Linking Data Services
  - Combine external and internal data sources
  - Linking data to each other
  - Retrieve information on demand
Conceptual Data Integration & Linking API

Different datasets

Different formats, models e.g. INSPIRE

Conclusions
- Easy integration
- Easy querying
- Hide complexities
- Easy Linking

- Link your data
- Open your data
- Data for all

Inspire Workshop, Aalborg 16 June 2014
Questions
&
Thank you
Backup Slides
The model in details - Observation, Measurement, Sample Taking

Proposed (New) Concepts
Reused Concepts (CIDOC)
Reused Concepts (DC, RDF)

LEGEND

Inspire Workshop, Aalborg 16 June 2014
GEUS data -> Conceptual Model – Borehole and Location

Inspire Workshop, Aalborg 16 June 2014
GEUS data -> Conceptual Model – Ground Water Sample

Inspire Workshop, Aalborg 16 June 2014
GEUS data -> Conceptual Model – Water Level Measurement

Inspire Workshop, Aalborg 16 June 2014
GEUS data -> Conceptual Model – Ground Water Chemical Analysis

Inspire Workshop, Aalborg 16 June 2014
Conceptual Model and INSPIRE

Inspire Workshop, Aalborg 16 June 2014
Conceptual Model and INSPIRE

• Model Mapped to INSPIRE
Demo Example

• Cross query over datasets
  – Find the boreholes and their locations in all datasets if ground water samples where taken from them after the date 01/01/2005.
Demo Example

```sql
select ?bhole ?date ?latitude ?longitude where {
    ?st sci:O5.removed ?sample;
    crm:P4F.has_time-span ?date;
    sci:O4.sampled_at ?place;
    crm:P2F.has_type "GRWATER_SAMPLING".
?bhole sci:O9.contains_or_confines ?place;
    a sci:S16.Borehole;
?bcollar crm:P87F.is_identified_by ?latitude;
    crm:P87F.is_identified_by ?longitude.
filter (regex(?latitude, "Latitude").)
filter (regex(?longitude, "Longitude").)
filter(?date > "2005-01-01"^^xsd:date).
} limit 100
}
UNION
{
    select ?bhole ?date ?latitude ?longitude where {
        ?anal crm:P2F.has_type "QUALITY_MEASUREMENT";
        crm:P4F.has_time-span ?date;
        sci:O4.sampled_at ?bhole;
?bcollar crm:P87F.is_identified_by ?latitude;
    crm:P87F.is_identified_by ?longitude.
filter (regex(?latitude, "Latitude").)
filter (regex(?longitude, "Longitude").)
filter(?date > "2005-01-01"^^xsd:date).
} limit 100
}
```
<table>
<thead>
<tr>
<th>bhole</th>
<th>date</th>
<th>latitude</th>
<th>longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://services.ades.eaufrance.fr/pointeau/1178ZZ2082/PZ2">http://services.ades.eaufrance.fr/pointeau/1178ZZ2082/PZ2</a></td>
<td>2005-04-04</td>
<td><a href="http://services.ades.eaufrance.fr/Latitude/14.649293295">http://services.ades.eaufrance.fr/Latitude/14.649293295</a></td>
<td><a href="http://services.ades.eaufrance.fr/Longitude/-61.0704197827841">http://services.ades.eaufrance.fr/Longitude/-61.0704197827841</a></td>
</tr>
<tr>
<td><a href="http://services.ades.eaufrance.fr/pointeau/1178ZZ2082/PZ2">http://services.ades.eaufrance.fr/pointeau/1178ZZ2082/PZ2</a></td>
<td>2005-04-04</td>
<td><a href="http://services.ades.eaufrance.fr/Latitude/14.649293295">http://services.ades.eaufrance.fr/Latitude/14.649293295</a></td>
<td><a href="http://services.ades.eaufrance.fr/Longitude/-61.0704197827841">http://services.ades.eaufrance.fr/Longitude/-61.0704197827841</a></td>
</tr>
<tr>
<td><a href="http://services.ades.eaufrance.fr/pointeau/1178ZZ2082/PZ2">http://services.ades.eaufrance.fr/pointeau/1178ZZ2082/PZ2</a></td>
<td>2005-06-23</td>
<td><a href="http://services.ades.eaufrance.fr/Latitude/14.649293295">http://services.ades.eaufrance.fr/Latitude/14.649293295</a></td>
<td><a href="http://services.ades.eaufrance.fr/Longitude/-61.0704197827841">http://services.ades.eaufrance.fr/Longitude/-61.0704197827841</a></td>
</tr>
<tr>
<td><a href="http://services.ades.eaufrance.fr/pointeau/1178ZZ2082/PZ2">http://services.ades.eaufrance.fr/pointeau/1178ZZ2082/PZ2</a></td>
<td>2006-08-23</td>
<td><a href="http://services.ades.eaufrance.fr/Latitude/14.649293295">http://services.ades.eaufrance.fr/Latitude/14.649293295</a></td>
<td><a href="http://services.ades.eaufrance.fr/Longitude/-61.0704197827841">http://services.ades.eaufrance.fr/Longitude/-61.0704197827841</a></td>
</tr>
<tr>
<td><a href="http://services.ades.eaufrance.fr/pointeau/1178ZZ2082/PZ2">http://services.ades.eaufrance.fr/pointeau/1178ZZ2082/PZ2</a></td>
<td>2006-08-23</td>
<td><a href="http://services.ades.eaufrance.fr/Latitude/14.649293295">http://services.ades.eaufrance.fr/Latitude/14.649293295</a></td>
<td><a href="http://services.ades.eaufrance.fr/Longitude/-61.0704197827841">http://services.ades.eaufrance.fr/Longitude/-61.0704197827841</a></td>
</tr>
<tr>
<td><a href="http://services.ades.eaufrance.fr/pointeau/1178ZZ2082/PZ2">http://services.ades.eaufrance.fr/pointeau/1178ZZ2082/PZ2</a></td>
<td>2007-06-18</td>
<td><a href="http://services.ades.eaufrance.fr/Latitude/14.649293295">http://services.ades.eaufrance.fr/Latitude/14.649293295</a></td>
<td><a href="http://services.ades.eaufrance.fr/Longitude/-61.0704197827841">http://services.ades.eaufrance.fr/Longitude/-61.0704197827841</a></td>
</tr>
</tbody>
</table>
Demo Example

Inspire Workshop, Aalborg 16 June 2014