# A Federated Approach for Array and Entity Environmental Linked Data

Shahed Bassam Almobydeen, José R.R.Viqueira, and Manuel Lama Penín

Centro Singular de Investigación en Tecnoloxías da Información (CITIUS)
Universidade de Santiago de Compostela (USC)
Santiago de Compostela, Spain
{shahed.al-mobydeen, jrr.viqueira, manuel.lama}@usc.es

**Abstract.** This paper discusses the main challenges that arise during the design and implementation of a federated solution for entity and array based environmental linked data. The proposed solution enables the integrated querying of geospatial relational databases, large scientific arrays of spatio-temporal dimensions and linked data sources with GeoSPARQL. To achieve this, a query decomposition algorithm and two new operators have to be incorporated in an already existing SPARQL query engine.

**Keywords:** Linked data, Geospatial and environmental data, SPARQL Query processing

## 1 Introduction

Large amounts of environmental data are becoming available over the entire world in the form of open data repositories. Two major types of these geospatial data are available, namely *entity-based* and *field-based* data. The former fits conventional models and it is usually managed within spatially enabled databases. The latter is modeled as large arrays with spatial and temporal dimensions, and it is usually recorded either with specific scientific array formats or in array database technologies.

Most of the current environmental data consumers are experts of different scientific domains who have the required skills to discover, access and analyze them. However, many other applications could benefit from these data if they were appropriately accessible through standard linked data technologies [1]. One such example is Tourism, where already existing linked data repositories like DBPedia may be combined with geospatial *entity-based* data of locations, hotels, restaurants or site seeing and with meteorological predictions modeled as large spatio-temporal arrays (see Fig. 1).

A major challenge related to applications like the above one is the representation of very large spatio-temporal arrays in the RDF data model [3], which is the model for data representation in the linked data paradigm. Obviously, spatial and temporal dimensions, which are not explicitly recorded in scientific

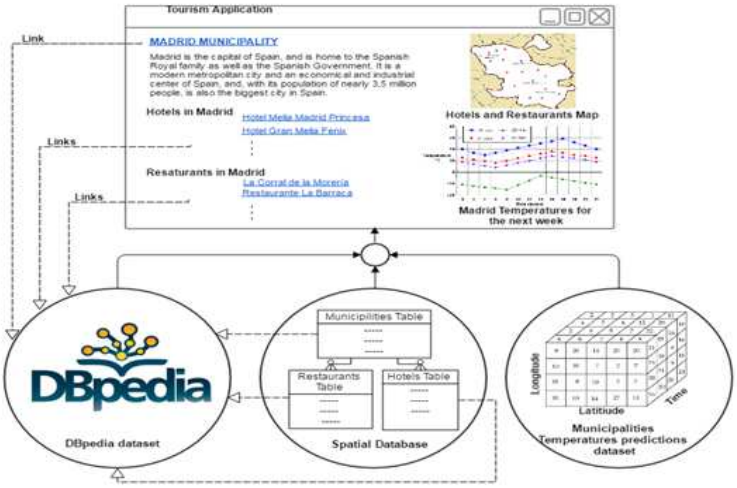


**Fig. 1.** Tourism Application.

array formats, must be explicitly recorded in traditional RDF encodings, leading this way to huge datasets that cannot be efficiently queried.

Many different solutions have been proposed during the last decade for the integrated querying of relational and linked data sources [6]. Some of them already considered datasets of *entity-based* geospatial data [2,4], which may currently be queried with the geographic extension of SPARQL (GeoSPARQL [5]). The above approaches may be classified according to their underlying data integration approach. Data warehouse approaches [4] use Extract Transform and Load (ETL) tasks to import relational spatial data into a spatially enabled RDF data storage technology. On the other hand, federated approaches [2] translate SPARQL queries to SQL to be evaluated directly by spatial relational databases. It is noticed that none of the above solutions has considered the incorporation of *field-based* large multidimensional array datasets.

A data warehouse solution would demand efficient storage formats and data access methods to integrate both entity and array-based RDF datasets. On the other hand, a federated approach, as the proposed in the present paper, requires the decomposition of SPARQL query algebra plans into three parts: one part to be translated to SQL, another one to be translated to some array query language, and a final part to be executed by the SPARQL engine by combining the above parts with native linked data sources.

## 2 Proposed federated approach

The main components of the architecture are (see Fig. 2):

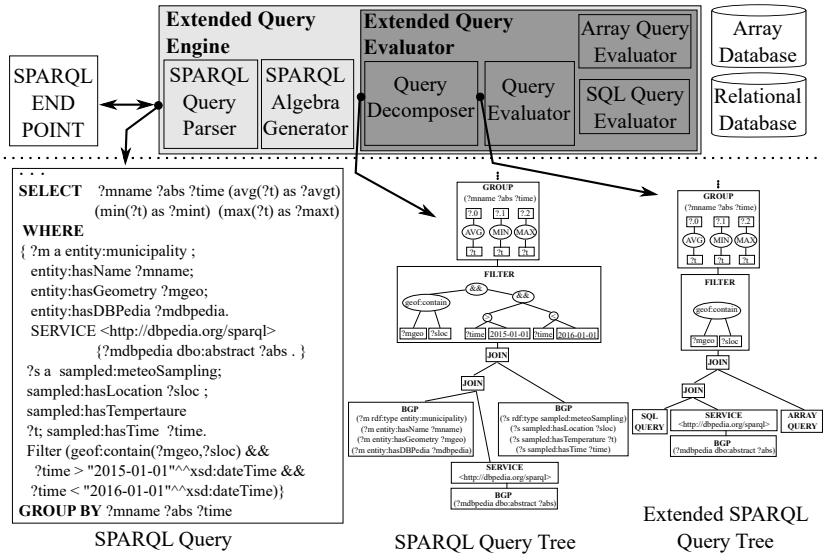


**Fig. 2.** Federated architecture.

- **SPARQL Query Parser**. GeoSPARQL queries submitted to the SPARQL end point are first parsed. The query illustrated in Fig. 2 retrieves the average, minimum and maximum predicted temperature for each municipality of DBPedia, at each time instant and for a specific time period.
- **SPARQL Algebra Generator**. A SPARQL query algebra tree is generated from the input parsed query. State of the art query optimization techniques are next applied. It is expected that filter conditions expressed using the vocabulary of just one data source are pushed down the tree close to the Basic Graph Pattern (BGP) operators that access the relevant data. In the above example the condition that restricts the time range in the array data source is pushed down to lay just above the BGP operator that access to the array data.
- **Query Decomposer**. This is the key component of the proposed solution. The optimized SPARQL query algebra tree is decomposed into three parts. The first one contains the maximum subtree that includes only vocabulary from the relational data source. In the above example, this is just the bottom left BGP operator that retrieves municipalities from the database. The second part contains the maximum subtree that includes only vocabulary from the array data source. In the above example, this is the bottom right BGP operator that retrieves temperature prediction, and the FILTER operator that restricts the search to a specific time period. The last part contains all the remainder nodes of the tree. The relational and array data source subtrees are next transformed to SQL and array query language, respectively. Those expressions are evaluated by two new SPARQL operators that gen-

erate result RDF triples from the retrieved relational and array data. The above operators are combined with the remainder part of the tree to yield a global SPARQL query algebra tree for the global query.

It is noticed that the proposed approach requires the implementation of a query decomposition algorithm and two new operators, one to execute SQL queries and another one to evaluate array queries. Both the query decomposition algorithm and the two operators are currently being implemented in the Apache Jena ARQ SPARQL engine.

## 3 Conclusion and future work

In this paper, we propose a federated architecture for accessing to entity and array environmental databases using linked data technologies. In our federated approach, SPARQL queries are used to access environmental data directly without loading them into RDF data stores. The new federated architecture enables expressing queries like "What is the predicted average of temperature of each municipality of Spain for the next week?". To achieve this, a query decomposition algorithm and two new opera-tors have to be developed. The implementation of the proposed federated architecture is still a work-in-progress. A comparison between the performance of the proposed federated architecture and already available data warehouse solutions will be done. Future work is mainly focused on the global query optimization at the mediator and on the incorporation of new Big Data storage and processing technologies.

## References


1. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. International Journal of Semantic Web Information Systems 5(3), 1–22 (2009)
2. Green, J., Hart, G., Dolbear, C., Engelbrecht, P.C., Goodwin, J.: Creating a semantic integration system using spatial data. In: Proceedings of the ISWC2008 Poster and Demonstration Session (2008)
3. Koubarakis, M.: Linked open earth observation data: The leo project. In: Proceedings of the The Image Information Mining Conference: The Sentinels Era (ESA-EUSC-JRC 2014). pp. 5–7. Bucharest, Romania (2014)
4. Kyzirakos, K., Karpathiotakis, M., Koubarakis, M.: Strabon: A semantic geospatial DBMS. In: The Semantic Web - ISWC 2012 - Proceedings of the 11th International Semantic Web Conference (ISWC2008). pp. 295–311 (2012)
5. Perry, M., Herring, J.: Ogc geosparql - A geographic query language for rdf data (2012)
6. Sahoo, S.S., Halb, W., Hellmann, S., Idehen, K., Thibodeau Jr, T., Auer, S., Sequeda, J., Ezzat, A.: A survey of current approaches for mapping of relational databases to rdf. Tech. rep., W3C RDB2RDF Incubator Group (2009), `http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.pdf`