A Semantic Mediation Architecture for RDF Data Integration

Amine Kerzazi, Othmane Chniber, Ismael Navas-Delgado, José F.Aldana-Montes

E.T.S. Ingeniería Informática
Universidad de Málaga
Campus de Teatinos
29071 Málaga
{kerzazi,chniber,ismael,jfam}@lcc.uma.es

Abstract. The emergence of the Semantic Web has produced several changes in the way in which applications are developed. Besides, the Semantic Web idea has changed itself since its initial appearance. Thus, the most significance change in the most recent proposal of the layer of protocols for the Semantic Web is the use of RDF for data interchange. In this sense, this paper presents a proposal for integrating data sources capable of providing their data as RDF documents. Thus, the use of an ontology-based mediator is able to integrate RDF data taking advantage of a domain ontology.

Keywords: Semantic Directory, ontology, Data Integration, mapping.

1 Introduction

Integration of data from heterogeneous knowledge sources represents consolidation of heterogeneous data aiming to generate new knowledge that can not be derived from single data sources. Researching in that sense has been topical for quite a long time, numerous papers were produced [1], and various strategies approaches were defined. The area of data integration in Semantic Web has gained popularity in recent years; integrated access to multiple distributed and autonomous RDF data source is a key challenge for many Semantic Web applications.

In [2], we proposed an *Ontology-Based Mediator* by making use of an infrastructure for building applications in the Semantic Web called *Semantic Directories* (SD) which is based on the registration of resources' semantics by relating them with ontologies. In this contribution, we present a system which provides access to distributed RDF data sources using Semantic Web technology. The proposed system retrieves data using SPARQL [3] queries (the W3C Recommendation for an RDF query language), which supports querying multiple RDF graphs. The system allows not only integrating ontologies into the Semantic Directory infrastructure, but also rewriting the SPARQL query in term of the integrated ontology making use of maching techniques.

2 Concept and architecture

The study of data integration systems has allowed us to determine the main elements of a data integration system, and thus to extract the pattern for building this kind of system. The system receives a query as input, translates and optimizes this query in several phases into an executable query plan, and executes the plan in order to obtain the results of the query.
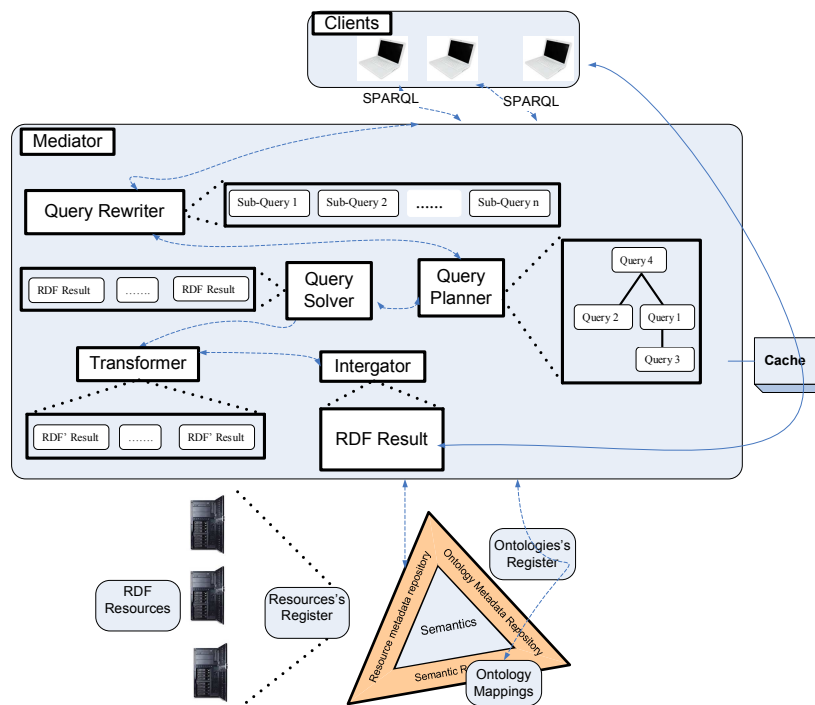

Figure 1: Architecture proposed.

The main purpose our architecture should serve is to provide a semantically unified interface for querying heterogeneous information sources. The architecture proposed (shown in the Figure 1) is composed of three major kinds of elements:

- *Semantic Directory* stores and manages meta-data concerning a number of domain ontologies, as well as the relationships among the ontologies and the data source/databases.

- *Data services* include the access to data sources/databases that could be queried through the Web.

- The *mediator* provides a way of using queries from users to produce integrated results. Its goal is to provide access to the data in the data sources using RDF as a common data model, and SPARQL as a common query language. Our architecture provides a semantically coherent model representation of the

combined data from the wrapped data sources and a transparent access to the combined data from the data sources through queries to the mediating view. The architecture of the proposed mediator is composed of five main components: *Query Rewriter (QR), Query Planner (QP), Query Solver (QS), Query Transformer (QT)* and *the Integrator (QI).*

The QR receives the SPARQL query user, parses and translates it into an *internal representation* that can be easily processed by all the components. The query is rewritten in all local registered ontologies term making use of mappings between them registered in the SD. Then, the *QP* receives a set of SPARQL queries (one for each integrated ontology), and aims to find a query plan. This component has been implemented including the most basic reasoning to take advantage of described semantics (subsumption and classification). The *QS* receives the plan and performs the corresponding call to the RDF data services involved in the sub-queries (SQ1, ...,SQn) to obtain a set of RDF documents. These RDF documents are composed and transformed to documents in term of the user ontology by the *QT*, and the Q*I* combines them to obtain the final result.

3   Conclusions

A solution to the problem of integrating heterogeneous information sources is needed in order to provide a uniform access to data from different sources available through the Web. The proposed integration architecture makes use of Semantic Web technologies. The Semantic Web perspective has been redefined in a new layered tower where on the top of it, two choices for data representation and interchange [4]: RDF and XML. This is interesting because the use of RDF as an interchange format will open new perspectives to the implementation of applications, and makes possible to use Semantic Web query languages to access this data. In this sense, SPARQL is proposed as the language for querying RDF documents.

References

1. J. Hendler, T. Berners-Lee and E. Miller, "Integrating Applications on the Semantic Web", Journal of the Institute of Electrical Engineers of Japan, Vol 122(10), October, 2002, pp. 676-680.
2. A. Kerzazi; Othmane Chniber; Ismael Navas Delgado; Francisco Aldana Montes; Raúl Montañez Martínez; Almudena Pino Angeles; Francisca Maria. Sánchez Jiménez. A Model-Based Mediator System for Biological Data Integration. Journées Scientifiques en Bio-Informatique. JSB'07, 28-29-30 Novembre 2007, EMI, Rabat.
3. http://www.w3.org/TR/rdf-sparql-query/
4. http://www.w3.org/2007/03/layerCake.png