World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:6, No:2, 2012

Multi-level Metadata Integration System: XML, RDF and RuleML

Messaouda Fareh, Omar Boussaid, Rachid Challal

***Abstract*—**Our work is part of the heterogeneous data integration, with the definition of a structural and semantic mediation model. Our aim is to propose architecture for the heterogeneous sources metadata mediation, represented by XML, RDF and RuleML models, providing to the user the metadata transparency. This, by including data structures, of natures fundamentally different, and allowing the decomposition of a query involving multiple sources, to queries specific to these sources, then recompose the result.

***Keywords*—**Mediator, Metadata, Query, RDF, RuleML, XML, Xquery.

I. Introduction

THE term "metadata" is used in general to refer to a structured set of information used to describe a resource. This information are represented and stored in multiple heterogeneously data sources. The basic need, is to able to query these different metadata sources, simultaneously, and give the impression to the user, it queries a single source. To do this, the solution of systems integration has been proposed.

It is to provide a consistent and transparent interface to relevant data via a global schema. The diversity of information distributed sources and their heterogeneity is one of the main difficulties faced by users today. Integrate data sources in order to provide users a uniform access interface is a difficult task. This problem involves three aspects: (1) the data heterogeneity, (2) the sources autonomy, and (3) the sources evolution.

The data heterogeneity concerning both the structure and semantics. The structural heterogeneity is the fact that data sources can have different structures and / or different formats to store their data. Many approaches to solve this type of heterogeneity have been proposed in the context of federated databases and multi-databases.

There are two main approaches for the information sources integration:

The first approach is to consider this integration as the construction of real databases, called Datawarehouses, containing information relevant to the applications considered.

The existing research work on the integration of semi-structured data C-web project[1], [14], [7], Xyleme[2]) or migration from one format to another [11] fall under this approach.

As for the second, it is to federate heterogeneous data from multiple sources into a single view. The data is not stored at the mediator and are accessible at the level of information sources.

In this paper, we focus on the mediator approach applied to heterogeneous metadata are represented by the models XML, RDF and RuleML.

We have organized our paper as follows: Section 2 presents a state of the art in data integration. Section 3 presents our system of mediation of heterogeneous metadata with details of each of its modules. Section 4 presents medical analysis such as fields of experimentation to test our architecture for integrating heterogeneous metadata. Finally, we conclude this paper by presenting the perspectives of our research.

II. State of The Art in Data Integration

A data integration system must provide users a uniform view of data sources it uses, and allow to question in a transparent manner, while the data are distributed across multiple sites. It provides a unified view of data from multiple sources and provides access to this data through an interface, regardless of their structure or their location [16].

A data integration system is a triple I: <G, S, M>, where:

- G is the global schema (defined on an alphabet AG) modeling the integrated schema,
- S is the set of source schemas (defined on an alphabet AS) describing the structure of sources participating in the integration process
- M is a correspondence between G and S that establishes the connection between elements of the global schema and those sources.

The first integration approaches, in the form of federated systems, have emerged in the 1980s, [5]. After federated systems, appeared integration systems based of mediators [8].

The set of manipulated data models is the relational model, it is a structured model and the XML model which is semi-structured. As an example, include the work of [6] which focuses on the integration of relational and XML data through an adaptive mediation system for support the integration of heterogeneous data.

And other work, we find such unstructured data or text: the work of [4] which focuses on relational databases and text type data.

Messaouda Fareh is with the LRDSI Laboratory, Blida University, Soumaa, Algeria (email: farehm@gmail.com).

Omar Boussaid is with the ERIC Laboratory, Lyon 2, University. Campus Porte des Alpes,France (email: omar.Boussaid@univ-lyon2.fr).

Rachid Challal is with the LMCS Laboratory, ESI (ex :INI), Oued-Smar (Algier), Algeria (email: r_chalal@esi.dz).

[1]C-web Project http://cweb.inrea.fr

[2] Xyleme Project http://www.xyleme.com

Most of the existing research work concerned by integration, it process data, whatever of their types, the adopted main data models: the relational model, the object-oriented model, the model XML ...., but for the metadata, there is only, some work, that focus on their integration, its focused by a simple format metadata, for example the XML and RDF model, but, these work ignored the rule format metadata as the RuleML model, which we are interested, there are a multiple of metadata representation language, our choice is focused on XML, RDF and RuleML models.

The most important problems of data mediation, on which research has focused in recent years, are:

- data modeling (how to integrate different schemas for sources), and
- their interrogation (how to effectively respond to queries posed to global schema).

We distinguish three main ways: the overall Global As View approach (GAV), Local As View approach (LAV), and the Both As View approach (BAV).

The GAV approach has been adopted in several systems such as HERMES [13], TSIMMIS [1] and e-XMLMedia [3], it is to build the global schema as views on local schemas.

For cons, the LAV approach is to conduct the reverse, ie to define the schema of data sources to be integrated as views of the global schema.

The LAV approach is very flexible with respect to the addition (or removing) of data sources to integrate: this has no effect on the global schema, only views should be added (or deleted). But it presents difficulties in reformulating queries. This approach was used in PICSEL [12], C-Web [2] Agora [9]....

The BAV approach [10] is data integration by bidirectional transformation rules of schemas. BAV is based on the use of a reversible transformation sequence diagrams.

Most systems that we studied (Sims, Tsimmis, Hermes, Agora, Xylem, Picsel, e-XMLMedia and C-Web) have a common architecture described in [15]), but they differ in

- The languages used to model the global schema, schemas of data sources to be integrated and user query, and
- How correspondence is established between the global schema and the schemas of data sources to integrate.

In the next section we present an approach for integrating heterogeneous metadata presented by the XML, RDF and RuleML models, and we are interested in the field of medical analysis.

III. Mediation System Architecture

In this section we propose an approach for integrating metadata sources XML, RDF and RuleML.

Our solution is characterized by:

1) Centralized architecture that provides transparent access to the location, the source schemas and query languages for metadata sources. We have defined a strategy module for efficient processing of query and improving global system performance.

2) Offer a structural integration of heterogeneous metadata XML, RDF and RuleML,

3) Integration approach down, and an adaptation of GAV mapping rules provide flexibility to change.

4) Processing sub-queries for data sources to support legacy systems.

Our integration system has an architecture with four levels:

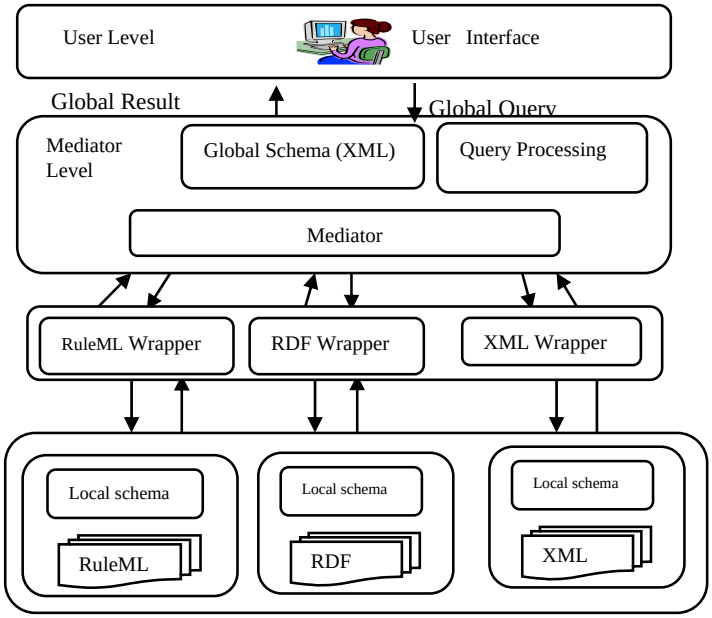

Fig.1 Mediation system Architecture

A. Level 1: metadatas Sources

We are interested in the mediation system by sources of metadata: XML, RDF and RuleML:

- *A RDF source:*

The RDF (Resource Description Framework) is the emerging standard proposed by the W3C for the representation and exchange of metadata on the Web, it has a syntax, and semantics. It is dedicated to the description of Web resources.

The RDF model data item consists of three types of objects:

(*1) Resources*: all things described by RDF expressions are called resources. A resource can be a whole web page or part of a web page. It can also be an object that is not directly accessible from the Web, ex., a printed book.

(2) The Property: A property is an aspect, characteristic, attribute, or relation used to describe a specific resource. Each property has a specific meaning, and

(*3) The statements:* A specific resource associated with a property defined as the value of this property for this resource is an RDF statement. These three individual parts of a statement are called, respectively, subject, predicate, and object

Our RDF metadata source is a metadata source named "medical exam"; it is interested in medical testing. Exam: An abstract class multi-family of exams (bacteriophage Viro

Parasitology, Biology, Dosage of drugs, tumor markers and pathology ...) so-called domain of exam. Each family consists of several tests that are represented by concrete classes (ALSO, ECB, glucose, CPK, theophylline, isoniazid, ACE, Alpha FP, Semen, Fibrinogen, VS, T3 T4 FT3 FT4 ....).

- *A XML source:*

It is a metadata source named "Test". It is interested in tests of each exam, with their Norms. Test: Each exam has a number of tests (eg exam "CPK" which belongs to the family "Biology" contains two tests: "CPK CPK Mb total" and "normal CPK value"). Norm: Class that represents a set of characteristics describing a test. All that is in between a Norm is considered "normal" (For the previous example of "CPK" each test has its own norm such as: CPK CPK Mb total must be less than 30% of normal value CPK should be between 10% is 100).

- *A source RuleML:*

The RuleML[3] aims to establish an "interlanguage" Web classic for the treatment of administrative rules, and that using the XML language, and formal semantics.

RuleML can thus improve to RDF, the rules expressiveness. The rules semantics is not specified in the specification. However, the translation into first order logic has no ambiguity. RuleML maniple graphs. XML tags are introduced for each specific type of rule

Our RuleML source is a document named "Exam-composed-test".

Here is a fragment of this document:

```
<ruleml:rulebase>
<Atom>
        <Rel>composed_of</Rel>
        <Var>Test</Var>
        <Var>ECB</Var>
</Atom>
<Atom>
        <Rel> composed_of </Rel>
        <Var>ECB</Var>
        <Var>Cell</Var>
</Atom>
……
</ruleml:rulebase>
```

B. Level 2: Wrapper

This Level includes wrappers for each source of local metadata. These wrappers provide an interface between heterogeneous sources of local metadata and the mediator. The first step in integrating is to generate XML schema corresponding to each source of local metadata. It is about transforming the source schema in an XML schema.

The use of XML for modeling local schemas provides a uniform representation of different metadata sources and facilitate their subsequent integration. Each wrapper (RDF and RuleML) is responsible for translating the local source schema in an XML schema, and execute the appropriate subquery. The XML wrappers execute the appropriate subquery.

The following tables compare the RDF and RuleML models:

TABLE I
COMPARISON BETWEEN RDF AND RULEML

Criterion	RuleML	RDF
Inference	Yes	No
XML Syntax	Yes	Yes
RDF Syntax	Yes	Yes
Anonymous resources	Yes	Yes
Negation	Yes	No
Rules representation	Yes	No
Metadata representation	Yes	Yes

TABLE II
MAPPING BETWEEN RDF AND RULEML

Criterion	RuleML	RDF
Concepts	Atom	Concept, or relationship Instance
Proprieties	-opr	Role which is represented by an arc
Objects	In OORuleML: Ind wref= the object value	A literal RDF
Predicates	In OORuleML :Ind with wref (web reference) attribute	Predicate RDF.
Annotations	In Object Objet RuleML: <Atom> <oid><Ind wlab=Subject </oid> <slot> <Ind wref= predicate1 <Data>Object1</Data> </slot> <slot> <Ind wref=Predicate2> <Ind wref=Object2> </slot> </Atom>	<rdf:Description About=Subject <predicate1> Object1 </Predicate1> <Predicate2 rdf:resource=Object2/> </rdf:Description>
Triplet	Predicate(subject, object) or rel(ur, ur|ind) <if> <atom> <rel> predicate </rel> <ur> subject </ur> <ind> object </ind> </atom> </if>	Triplet RDF (predicate, subject, object)
Resource type	In OORuleML : is the relationship with its atom <Atom> <oid><Ind wlab= Suject</oid> <opr><Rel wref= SujectType /> </opr> <slot><Ind wref= predicate/> <Data> Object </Data> </slot> . . . </Atom>	Is the propriety rdf: type <rdf:Description about Subject> <rdf:type resource=SujectType/> <predicate>Object</predicate> . . . </rdf:Description>

[3] RuleML Model: http://ruleml.org/

Here is a fragment of RuleML schema translated into XML:

```
<? xml version='1.0' encoding='iso-8859-1'?>
<ECB>
        <composed_of>
                <Cell></Cell>
                <Germ></Germ>
        </ composed_of >
</ECB>
<VS>
        < composed_of >
                <VS_1h></VS_1h>
                <VS_2h></VS_2h>
        </ composed_of >
</VS>
<CPK>
        < composed_of >
                <CPK_Value></CPK_Value>
              <CPK_MB></CPK_MB>
        </ composed_of >
</CPK>
  ……
```

We present, too, a fragment of RDF schema translated into XML:

```
<Exam>
<Bacterio_Viro_Parasitology>
<ECB></ECB>
<ASLO></ASLO></Bacterio_Viro_Parasitologye>
<Biology>
        <Glycemy></Glycemy>
        <CPK></CPK></Biology>
<Dosage_of_drugs>
        <theophylline></theophylline>
        <isoniazid></isoniazid>
</Dosage_of_drugs>
< tumor_Markers>
<ACE></ACE>
<Alpha_FP></Alpha_FP></ tumor_Markers >
<Anatomical_pathology>
        <Sperm></Sperm>
</ Anatomical_pathology >
<Hematology>
…….
```

C. Level 3: Mediator

The heart of our system is in the mediator. It is decomposed into modules connected to each other, which are: (1) Creation global schema module. (2) Query processing module, and (3) results fusion Module. In the next section we detail the different algorithms, for each mediator module:

1) Global schema creation

The presence of a global schema is necessary because it provides a unique vocabulary for expressing user query. This schema unifies the heterogeneous schemas of sources to be integrated, based on a homogeneous, uniform and abstract description of sources content by views. Our global schema is created using the common format (XML), and the GAV approach, that is to say that the global schema is considered to be a view of the sources schemas.

We present a fragment of global schema*:*

```
<?xml version="1.0"?>
< global _ schema >
< view_xml>
        <person> </person>
        <test>
         <g_a_j> </g_a_j>
         <vol_spr> </vol_spr>
        …….
< /view_xml>
< view_rdf>
      <exam>  <bvp>
              <ecb> </ecb>
               <aslo> </aslo>
              </bvp>
               …..
</ view_rdf>
< view_ruleml>
               <ecb></ecb>
               <aslo></aslo>
              <glyc></glyc>
                   ……
</ view_ruleml>
```

2) Query processing

As the global schema is expressed in XML, queries should be based on XML. There are a variety of query languages for XML documents namely XSLT, XPath, XQL, XML-QL, QUILT, XQueryX, Lorel, XQuery1.0 and TexRet. But after studying these languages we find that XQuery is the most appropriate query language to examine our global schema. In fact, XQuery is a powerful query language, allows querying of heterogeneous data sources. It differentiates between set, and to use predefined functions, which does not exist in other query languages.

Example: Consider the following user query:
Global query: for $c in doc ("Global-schema.xml")/ Medicals- Analysis
for $a in $c/bio return { $a }

The latter can be decomposed into three sub queries defined on the global schema, each is designed to a source.

Sub Query 1: for $c in doc ("RDF-source.xml")/ Medicals-Analysis
for $a1 in $c/bio return {$a1}

Sub Query 2: for $c in doc ("RuleML-source.xml")/ Medicals-Analysis
for $a2 in $c/glyc return {$a2}

Sub Query 3: for $c in doc ("XML-source.xml")/ Medicals-Analysis
for $a3 in $c/g-a-j return {$a3}

3) Fusion of results

This module builds the global response to a query, using the results of local queries sent by the wrappers. We present the algorithm for construction of the response of a sub-query Qi.

Algorithm : Fusion of sub-results
Input : a set of sub-response
Output : global response
Begin
For each sub-response *Do*
Insert ion the global response RG, SR

End

D.Level 4: User

It is a simple communication interface allows a user to communicate with the system, it sends query to the mediator and receives responses, and this interface contains fields for selecting, which helps the user to select items in the query.

IV. Experimentation

A prototype has been realized to demonstrate the feasibility of our approach. This prototype contains mainly: a client interface allowing the user to query the system and an administrative interface for configuring the system.

We selected medical analysis metadata, as a field of experimentation. And metadata sources that we used for our system are:

1) A RDF source to describe the exams domain, realized under Altova SemanticWorks 2011.

2) A RuleML source for the description of exam of each domain, realized under the DR-device environment, and

3) A XML source for description of tests for each exam realized under Stylus Studio 2011 XML Enterprise.

The first step of integration is to generate for each source a local schemaXML, treatment is applied to these schemas to create the global schema, applying the algorithm for creating the global schema.

The user can pose his query: by domain, exam, or test. As an example, selection of domain, ie the user selects any domain, so he seeks the values of all tests of each exam of this domain. Once the selection is made, the user can see the query generated in XML format.

The query is decomposed into three sub queries: one for the extraction domain, the second for the extraction of selected domain exam and the third for the extraction of exam tests for chosen domain.

After that, the user can see the result of the query generated by the system following the selected domains.

In case which the user selects the field "*Dosage_of_drugs*" the system precedes as follows:

1. From the correspondence table (which contains a global concept for each local concept), the system extract the local concept corresponds to the global concept "*Dosage_of_drugs*," it is in the source RDF.

2. From the RuleML source, the system proceeds to the extraction of all exams in the domain "*Dosage_of_drugs*" which are *theophylline* and *isoniazid*.

3.1. For the *theophylline* examen, the systems extract its test "*therapeutic- concentrations_of_theophylline* "found in the XML source.

3.2. For the exam *Ionized*, the system extract its test "*therapeutic concentrations of isoniazid*" found in the XML source.

4. For each test, the system extracts its norm.

5**.** Finally, the system merges the results and displays the values of tests with their norms.

V. Conclusion

In this paper, we proposed an approach for integrating heterogeneous metadata represented by XML, RDF and RuleML models. The system we propose is a mediator- based system using the approach GAV. Metadata will be stored at source, and the mediator saves to its level, their descriptions (as virtual views). The global schema of our system will automatically construct. Our experimental field was the medical domain and specifically medical analysis. We have described the general architecture of the system, the transformation schemas sources, the integration of local schemas, the steps of query rewrite and results integration from the different metadata sources.

Finally, we show the experimentations on a set of metadata sources of medical analysis. The results obtained are encouraging. We envisage making our system more extensible by treating the case of the distribution of sources, and query optimization.

References

[1] G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

[2] Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J.K. and Widom, J. The tsimmis projec: Integration of heterogenous information sources. Journal of Intelligent Information Systems, 8(2), 1997.

[3] Christophidesy, V., Community Webs (C-Webs): Technological Assessment and System Architecture. Technical Report, 2000.

[4] Dang-Ngoc, T.T. and Gardarin, G. : Conception et Evaluation de XQuery dans une architecture de médiation Tout-XML, Revue ISI (Intégration de systèmes d'information) : Vol. 8, number 5-6, pp 11-25, 2003.

[5] Diallo, G. : Une Architecture à Base d'Ontologies pour la Gestion Unifiée des Données Structurées et non Structurées. Doctoral thesis, Joseph Fourier university – Grenoble I. 2006.

[6] Hurson, A. R. and Bright, M. W.: Multidatabase systems: An advanced concept in handling distributed data. Advances in Computers, Vol. 32. pp149-200, 1991.

[7] Kostadinov D., V., Peralta, A. Soukane, and X. Xue : Intégration de données hétérogènes basée sur la qualité. Congres INFORSID, Grenoble, France, Pp 471-486, 2005.
[8] Mary F., Fernandez, Daniela F., Alan Y. and Dan S.: Declarative specification of web sites with strudel, VLDB journal, pp 38-55, 2000.
[9] Mena, E., Illarramendi, A., Kashyap, V. and Sheth, A. P., Observer: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. In Int. Journal Distributed and Parallel Databases, Distributed and Parallel Databases, Vol. 8 nb. 2. pp 223-271, 2000.
[10] Manolescu, I., Florescu, D., Kossmann, D., Xhumari, F. and Olteanu, D., Agora: Living with XML and Relational. In Proceedings of the 26th VLDB Conference, Cairo, Egypt, 2000.
[11] McBrien P. and Poulovassilis A. : Data Integration by Bi-Directional Schema Transformation Rules. 19th International Conference on Data Engineering, Bangalore, India, 2003.
[12] Papakonstantinou Y.,Garcia molina H. and Widom J. : Object Exchange Across Heterogeneous Information Sources. In ICDE Conf, on management of data.Maurizio L., Data Integration: A Theoretical Perspective, p2, 2002 http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.91.9907.
[13] Rousset, M-C., Bidault, A., Froidevaux, C., Gagliardi, H., Goasdoué, F., Reynaud, C, and Safar, B. : Construction de médiateurs pour intégrer des sources d'information multiples et hétérogènes. PICSEL project, 2002.
[14] Subrahmanian, V. S., Adali, S., Brink, A., Emery, R., James J.Lu, Rajput, A., Rogers, Timothy J., Ross, R. and Ward, C.s., HERMES: A Heterogeneous reasoning and mediator system. Technical report, University of Maryland, 1995.
[15] Vassilias C., Sopie C. and Térome S.: On wrapping query languages and efficient XML integration. In SIGMOD, Dallas, Texas, 2000
[16] Wiederhold, G.: Mediators in the Architecture of Future Information Systems, Actes IEEE Computer, p.38-49, 1992.
[17] Djema L., Boumghar F.O., Debiane S., L'imagerie Médicale Dans une Base De Données Distribuée Multimédia Sous Oracle 9i, pp3-5, 2007. http://www.setit.rnu.tn/last_edition/setit2007/TI/110.pdf.