An XML Schema-based Semantic Data Integration

Dongkwang Kim[1], Karpjoo Jeong[2], Hyoseop Shin[2] *, Suntae Hwang[3]
[1] *Department of Advanced Technology Fusion, Konkuk University, Seoul, Korea*
[2] *Department of Internet & Multimedia Engineering, Konkuk University, Seoul, Korea*
[3] *Department of Computer Science, Kookmin University, Seoul, Korea*
[1] *walhalla@gcslab.konkuk.ac.kr,* [2]*{jeongk, hsshin}@konkuk.ac.kr,* [3]*sthwang@kookmin.ac.kr*

Abstract

Cyber-infrastructures for scientific and engineering applications require integrating heterogeneous legacy data in different formats and from various domains. Such data integration raises challenging issues: (1) Support for multiple independently-managed schemas, (2) Ease of schema evolution, and (3) Simple schema mappings. In order to address these issues, we propose a novel approach to semantic integration of scientific data which uses XML schemas and RDF-based schema mappings. In this approach, XML schema allows scientists to manage data models intuitively and to use commodity XML DBMS tools. A simple RDF-based ontological representation scheme is used for only structural relations among independently-managed XML schemas from different institutes or domains We present the design and implementation of a prototype system developed for the national cyber-environments for civil engineering research activities in Korea (similar to the NEES project in USA) which is called KOCEDgrid.

1. Introduction

A scientific experimental process is often a long and complicated workflow which includes various tasks and their inter-dependencies. From those tasks, various types of experimental data (e.g., sensor data, images, video recordings) may be generated. In addition, contextual data about those tasks and their data (e.g., experimental parameters and configurations) may also be collected. We consider scientific data to include both experimental data and contextual information.

In the e-science era, the *integration of scientific data from different organizations or areas* is crucial for many advanced research techniques. Such integration allows scientists to retrieve data from different sources in an integrated way, regardless of different schemas. Such data integration is very difficult to support in traditional data management approaches [14].

Challenges for scientific data modeling and integration are:

1) *Ease of Schema Evolution*. Due to the endless pursuit for new problems and better research approaches in scientific communities, scientists continue to change or improve scientific experimental processes. Therefore, it is required to evolve data models (i.e., schemas), accordingly. Such schema evolution must allow scientists to manage data on both new and old schemas.
2) *Support for Multiple Independently-Managed Schemas*. In the case of data integration for e-Science or grid applications which include multiple institutes, it is very difficult to expect the management of those schemas in a centralized manner. Therefore, it is important to support distributed schema management in which schemas can be independently dealt with.
3) *Simple Schema Mapping*. Data integration requires relating multiple schemas (which is called schema mapping) in order to allow scientists to retrieve data of different schemas in an integrated way.

2. Our Approach

2.1. XML Schema-based Data Model

In our work, we take the *XML-based approach* because most scientific communities are these days working or planning on converting their scientific data or its metadata into XML documents. The common XML-based data modeling approach is to use RDF (Resource Description Framework) for data model representation which allows ontological reasoning.

* Author for correspondence: +82-2-2049-6117, hsshin@konkuk.ac.kr

Proceedings of the Fifth International Conference on Grid and Cooperative Computing (GCC'06)
0-7695-2694-2/06 $20.00 © 2006 IEEE

IEEE COMPUTER SOCIETY

But we use *XML schema for data model representation* and *RDF only for schema mapping* because XML schemas are more intuitive than RDF representations and RDF-based reasoning requires serious processing overheads. XML schemas also allow us to use commodity data management software such as XML DMBS [6] and query processing systems such as XQuery [7, 8].

2.2. Schema Management

In our approach, we assume multiple sites to participate in data integration and to *manage their own schemas independently*. In order to support simple queries, we also assume a global schema registration system and query generation system. This global layer is designed to allow the user to formulate queries without understanding those multiple schemas in detail.

As explained previously, *schema evolution* (i.e., experimental process evolution) is crucial for scientific data management. There have been a lot of research work on schema evolution in the database community, but it is still challenging in the traditional relational database system [9]. In our approach, we take a simple approach which treats those schemas as independent. For example, suppose that site A creates a schema, and then site B extends it. We allow both sites to manage those schemas independently, but provide the global query generation system which creates a separate query.

In order to facilitate query generation for multiple schemas, we use *schema mapping mechanisms*. We design the schema registration system to manage XML schemas in RDF in order to support ontological reasoning among multiple schemas. The difference between our approach and conventional RDF-based ontology approach is that we do not represent a domain structure (concepts and their relationships), but the hierarchical structures of XML schemas and simple equivalence relationships among elements among them.

3. System Design

3.1. System Model

<Figure 1> describes the scenario of the proposed data integration system. It consists of two processes. One is defining and registering metadata model of research data and it is establishing integrated ontology. Another process is semantically searching the data sources by using the integrated ontology.

In first process, we define data model of metadata including experiment information (e.g., site name & description, participants) and result data, and register this data model into schema database. Then we establish integrated ontology that describes structural differences between this data models and the pre-defined model; 1) Schema Creation.

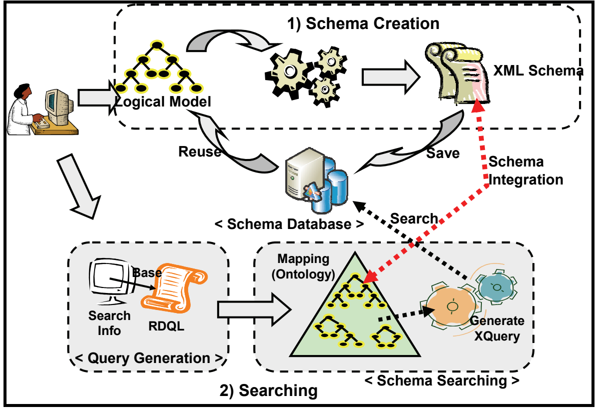

Fig 1 Use Scenarios of Our System

The second process is the integrated search using the defined ontology. Scientists select elements in integrated ontology and inserting search words using UI based Web, and then RDQL [10] queries are generated from inserted parameters. The system makes it possible to know structural information of XML Schema through executing the generated RDQL and recognizing structural relationships between the elements in the ontology. Finally, It make XQuery fitting on these data model and be executing the XQuery and show merging results of these to user; 2) Searching.

3.2. Schema Mapping & Integrated Searching

The goals of our schema mapping are to overcome structural conflicts by representing semantic relationships between elements of xml schema-based data models. We express structural relationships between elements and sub-elements of data model as RDF statement [11], and establish the relationships between elements of other data models. The Global Schema (integrated ontology) is created by including this information.

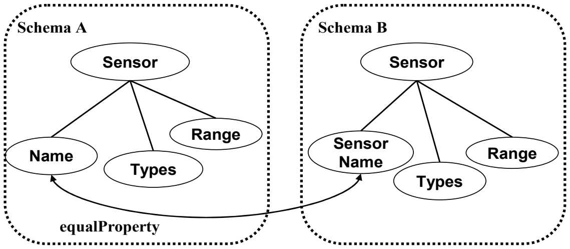

Fig 2 Structural relationships between elements

We define translation roles for representing elements of data model into RDF as follows. It represents all elements of the model into subjects and sub-elements into objects (1). And then, there are linking relationships between element and element of other schema (2). There are three predicates to define above-mentioned relationships of elements and one relationship between element and properties: (1) *hasProperty* , (2) *equalPropery*

Semantic searching is an integrated search between data models regardless of different structures. We can not search diverse XML Schema-based data model with different structures by using XQuery language. By using integrated ontology, we can translate inserted query to XQuery adapted every schema in schema database.

The execution of query is executed as follows:

1) Scientists insert elements and query keyword for searching (e.g., X == 'X310a' and Y <= 823.5).
2) Inserted queries are translated RDQL to search elements linked by equalProperty in integrated ontology(Step.1) and it find relational parent elements linked by hasProperty by executing these query (Step.2)
3) The results of RDQL with structural information of schema are translated XQuery and each XQuery are executed against each schema.
4) Our system collects all the results.

```
while not end(insertQuerys) {
   // Step.1
   equalResult = select ?x where ( insertQuerys[i], equalPropery , ?x );
  // Step.2
   whlie not end(equalResult) {
      equalStructure = select ?y where ( ?y , hasProperty, equalResult[j] );
   }
  // Step.3
   generated_xquerys[i] = RDFtoXQuery( equalStructure, disjoinStructure );
}
// Step.4
execute generated_xquerys[];
```

Fig 3 Integrated Searching Algorithm

4. Implementation

4.1. System Architecture

Our system consists of two components to support managing schema and integrated searching. First component is Schema Management. It defines and managing diverse XML Schema-based data model and establish structural relationships to integrated ontology. Second component is semantic search based on the integrated ontology.

In detail, Schema Management is divided into five services: Schema Register/Viewer/Modifier, Metadata Insert and Ontology Management. Schema Register define visualization tool for data model, Schema Modifier manages and modifies the existing data models. Schema Viewer is GUI-based representation tool that visualizes XML Schema for offering facility to administer. Also Ontology Management establishes integrated ontology about the registered data models. These data models and ontology are stored and managed in various databases of each system through OGSA-DAI [12]. OGSA-DAI are middleware to execute integrated query in heterogeneous database environments (e.g., eXist, Oracle).

Integrated searching service consists of RDF-based Query Generator and Ontology & Schema Searching. First of all, RDF-based Query Generator (RQG) is providing analyzed description of the integrated ontology to researchers. They select these description what they want to seek and insert search keyword, then RQG creates RDQL query based on inserted search information. Ontology & Schema Searching service receives this RDQL query, and analyzes the integrated ontology to execute the query, then this service make XQuery that are suitable for each data model of databases. These XQuery are executing at every database through OGSA-DAI and result data of query are presented to users.

4.2. User Interfaces for web-based semantic data management.

We apply the approach of this paper to KOCED that is grid-infrastructure at civil engineering research area. We implement XML Schema-based metadata management and integrated searching tool that supports to define and manage data model of each experiment for effectively sharing several of research data in KOCED.

Usually, researchers do not know XML Schema language. Then, we support simple data model management tool that modifies registered existing data model by searching and appends new data model. Also we develop automatically generating data model-based insert-form interface to offer researchers opportunities to insert experiments data into database by using registered data model. Our system provides mapping tool for building ontology. Scientists easily establish relationships between new element and the integrated ontology. To support easy searching, our system provides integrated-ontology viewer as search condition. Then, researchers make search queries by selecting search elements among this search conditions and inserting search keys without structure information of data model. This query are translated RDQL and finally it is generated XQuery. There queries are executed in many schema databases over grid system by using OGSA-DAI to integrate executing query.

5. Related Works

NEES and GEONgrid are grid infrastructure for result data management and collaboration with these data [3, 5]. NEES have defined reference data model, and built NEESCentral that is central data storage based on these data model to share the result data among co-workers in distributed research sites. NEESCentral provides a web-interface into the interim metadata model and high-level model based on existing data and practices from equipment sites. Also this system provides hierarchical model with metadata values added to individual directory level nodes. However, NEESCentral can not carefully represent the data model about experiment data, because supported metadata model are only including basic data of experiment, and then researchers can not append new data model for their experiment and modify pre-defined data model into theirs. Geosciences Network (GEON) is building ontology-based grid infrastructure for result data integration. This system provides ontology-based searching of geological images using OWL. This ontology-based searching is developed by using the ontology that is mapping between metadata of geologic maps and categories for describing geological features. Users are only able to search geological data by selecting integrated categories.

The Edutella system and Piazza enable interoperability among other systems that are producing and managing diverse formats [13, 14]. The Edutella system provides query and storage services for RDF, and this project is on translating the RDF data and queries to the underlying storage format and query language. But this project uses canonical mappings to store it in different systems. Piazza offers a language for mediating between data sources to enables interoperability of XML data. The language of mediating are to represent structural relations between the data defined XML, it is like XQuery. And it is not required a global schema, unlike our system, Piazza is modifying and defining XML data to integrate and share

6. Conclusions and Future Work

The data integration in grid system is very important for effective collaboration in scientific research fields. They want easy schema management, multiple schema management and simple schema mapping for effective data management from diverse format and systems. To address this problem, we provide semantic data integration and integrated searching to effectively access data between research institutes and to easily comprehend these data.

In this paper, we propose XML Schema-based semantic data integration for effective sharing heterogeneous data sources in different format and from diverse system. We define data model of metadata based on flexible XML Schema. Also, we develop the integrated searching system to effectively search and manage changeable data model continuously. This approach is making integrated ontology that is maintaining relationships between each element of data models to search various data models for integrated searching.

In the near future, we will develop visualization tool to improve usability for maintaining and searching relations between XML Schema-defined data model and ontology by presenting RDF.

7. References

[1] Mark Ellisman and Steve Peltier: Medical Data Federation: The Biomedical Informatics Research Networks, The Grid 2 Second Edition , Pages: 109-120, 2004

[2] Korea Construction Engineering Development Collaboration (KOCED) , www.koced.net

[3] NEESgrid, http://it.nees.org

[4] Jun Peng , Kincho H. Law: Reference NEESgrid Data Model [TR-2004-40] (2004)

[5] GEONgrid, http://www.geongrid.org

[6] Bellahsene Zohra, Milo Tova, Rys Michael, Suciu Dan, Unland Rainer: Database And XML Technologies , Second International Xml Database Symposium, Xsym 2004, Toronto, Canada, August 29-30, 2004, Proceedings

[7] S.Boag, D.Chamberlin, M.F.Fernandez : XQuery 1.0: An XML query Language, 30 Aprial 2002, http://www.w3.org/TR/xquery

[8] Z. G. Ives, A. Y. Halevy, and D. S. Weld. An XML query engine for network-bound data. VLDB Journal, 11(4):380-402, December 2002.

[9] Avi Silberschatz, Henry F. Korth, S. Sudarshan: Database System Concepts Fifth Edition , ISBN 0-07-295886-3

[10] Andy Seaborne, HP Labs Bristol : RDQL - A Query Language for RDF , 9 January 2004, http://www.w3.org/Submission/RDQL/

[11] Dan Brickley and R. V. Guha.: W3C Resource Description Framework (RDF) Schema Specification, http://www.w3.org/TR/1998/WD-rdf-schema/, March 2000. W3C Candidate Recommendation.

[12] Mario Antonioletti, Malcolm Atkinson, Rob Baxter, Andrew Borley: The design and im-plementation of Grid database services in OGSA-DAI, Concurrency and Computation: Practice & Experience archive Volume 17 , Issue 2-4 , Pages: 357 - 376 (2005)

[13] Wolfgang Nejdl, Boris Wolf, Changtao Qu : EDUTELLA: A P2P Networking Infrastructure Based on RDF (2002), May 7-11, 2002, WWW2002

[14] Zachary G. Ives, Alon Y. Halevy, Peter Mork : Piazza: Mediation and Integration Infra-structure for Semantic Web Data, Journal of Web Semantics manuscript.

IEEE COMPUTER SOCIETY