Ontology Merging and Database Schema Integration: An Approach to Identify Semantic Similarity and Resolve Schematic Heterogeneity in Interoperable GIS Application

Nanna Suryana[1], Shahrin Sahib [1], Ridlwan Habibi[2] ,Norayu Abdul Ghani [3], Zahriah Othman[1],
Ahmad Tajuddin Samsudin [4]

Faculty of Information Technology and Communication (FTMK), The University Technical of Malaysia Melaka (UTeM), PO Box 1752, Durian Tunggal, 76109, Melaka, Malaysia

[1]{nsuryana, shahrinshahib, zahriah}@utem.edu.my, [2]ridlwan.habibi@yahoo.com,
[3]norayugan@gmail.com, [4]tajuddin@tnrnd.com.my

Abstract

Data and information sharing is driven by the need to maintain more accurate and up to date spatial temporal database and at the same time reduce the data acquisition and maintenance costs. This paper discusses the need of a database schema integration and ontology merging to support interoperability GIS applications. This enable translation query from one database schema into another to support finally the development of XML-GML based document. Research results shows the potential use of the approaches to solve problems associated with seamless GIS based information sharing.

1. Introduction

This research has been conducted under the framework of GIS Interoperability Research Project co-financed by UTeM's short grant and e-Science Fund, the Malaysian Ministry of Science Technology and Innovation. The overall objective of the project is intended to develop interoperable GIS architecture, GIS interoperability model and semantic translator engine. Thus it increases the usefulness of GIS into different possible application domains.

Within the last few decades, GIS has been showing its strong capability for assembling, storing, manipulating and displaying spatial relationship [1]. However one of the primary obstacles in GIS application is the heterogeneity of different sources of GIS data. In [3] and [9] the author has identified the heterogeneities can be classified as syntactic, schematic and semantic heterogeneities.

Rapid development of Internet technology support different data and information to be easily available widely and freely. But those data are not always useable for other users due to the differences in data acquisition technique, data definition and their semantic meaning. In other words, problems associated with heterogeneous of databases are generated by differences in the data model being used, in its schema and data definition. This leads to lacking of interoperability and compatibility among different GIS platforms. Editing and acquisition spatial data is usually problematic costly and time consuming [9]. Information sharing between two or more GIS platforms have not been fully implemented. Because of this situation, unnecessary redundant spatial data acquisition conducted by different GIS users becomes unavoidably occurred. Several vendors have introduced export and import conversion machine.

However it has been found to be the main cause of losing too much data and accuracy. Therefore issues related to interoperable GIS have been discussed by millions of GIS users and researchers all around the world. To overcome the short coming of information sharing an alternative solution to provide a uniform data access to different data sources is required. It implies the highly need of interoperable GIS which has been considered to become the best option of the current and future GIS architecture from which is considered to be economically beneficial to all GIS users especially for those who have been planning to reutilize the existing available data.

As stated by [9] technical development supports the distributed data-information sharing using interrelated modules: firstly clearing house module to serve as an active link to several external data servers and data owners. Secondly and thirdly are client and provider modules that include a "Semantic Translator Engine

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.187

(STE)". Under the light of STE development, this paper discusses the use of database schema integration and ontology merging to support firstly a uniform data access to the different data sources, secondly to facilitate the integration of two different database schemas and finally to produce a common interoperable database i.e., the XML based format that can be stored in a Federated Database Schema. Adopting this concept allows the proposed STE to detect automatically semantic similarity and differences from two and more GIS platforms.

2. Conceptual Database Schema Integration and Ontology Merging

Figure 1 below shows the nature of existing information that belongs to different users. **Figure 1.(a)** illustrate pieces of information could have different format, from different source and varying in quality. **Figure 1.(b)** shows also the integration of all these pieces of information should construct compatible format with compatible semantic meaning. However the reality is some point of these pieces is not always compatible as shown in **Figure 1.(c)**.

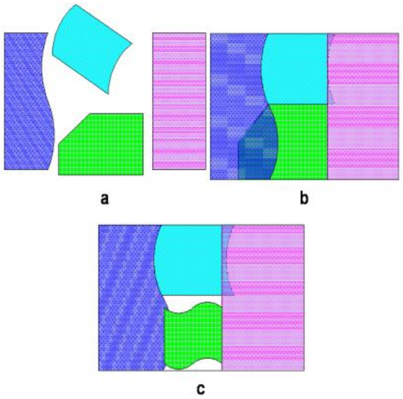


Figure 1. Information Integration (a) Piece of Information (b) The ideal situation (c) The Reality

Based on integration framework as above, **Figure 2** below shows that the ultimate objective of database schema integration is to create one common format of database which allows different users to relocate and reutilize the required geospatial data.

The focus of this particular part of this paper is on how to integrate different schema from different GIS application databases associated with different GIS users and vendors who have defined many application based on different specific data models. However it must be noted that the transformation and translation of source of data involves not only earth-referenced spatial data between source data and targeted data from different GIS platforms, but includes also at the same time migrating their spatial data attribute, geo-referencing, data quality report, data dictionary, and other supporting metadata.

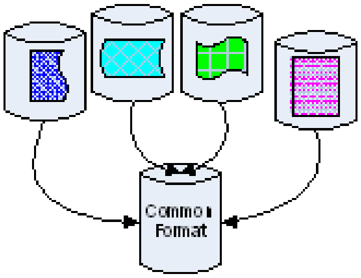


Figure 2. Common Format of Database

An integration requirement between two or more databases requires one global definition to represent two different classes within two different databases schema but represents the same concept. This is possible since the proposed semantic model allows to present data in a very abstract and understandable manner. It has been currently used in designing the conceptual structure of database. When one source of data could be transformed and translated directly into desired targeted data, this could eliminate duplication at the same time avoid problem of multiple updates, and minimize inconsistencies across applications. **Figure 3** bellow shows architecture for integrated database schema.

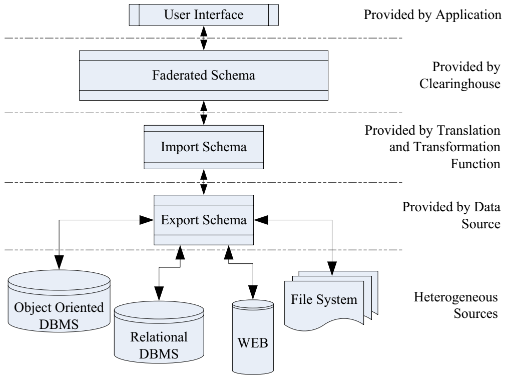


Figure 3. Architecture of Integrated Database Schema

As presented in **Figure 3** and stated by [4], the Federated Database Schema consists of a collection of possible heterogeneous, interoperating but autonomous component databases. It has become the preferred strategy for reconciling the proliferation of private and independent databases, and for increasing the productivity of information technology investment [8] below.

Based on how Federated Database Schema is defined as above, it becomes apparent that traditional searching based on keyword is not sufficient capacity to check semantic meaning of searched objects since it only consider them as character strings [7]. From ongoing research and also supported by [6], it is clear that the key points for a semantic search refinement process depend on the availability of domain ontology and the ability to understand semantic relationships between ontology concepts as explained in **Figure 4**. Schema Ψ of set of information denoted as $\Psi_1,\ldots,\Psi_n$ and set of ontology β for schema of set of information denote as $\beta_1,\ldots,\beta_n$, Merging ontology, β is based on finding similarities or differences among schema of set of information, Ψ.

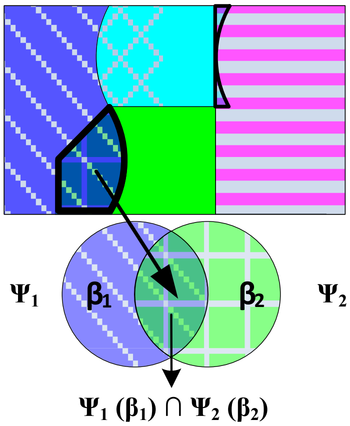


Figure 4. Basic Rules for Merge of Ontology

Giving that two ontology β_1 and β_2 from two different schema Ψ_1 and Ψ_2 respectively, the similarity could be defined as $(\Psi_1 \cap \Psi_2)$ or differences as (Ψ_1/Ψ_2) for these two schemas. Base on the basic rules, the above schemas could have $(\Psi_1(\beta_1) \cap \Psi_2(\beta_2))$ to present the intersection of the ontology of two schemas and $(\Psi_1(\beta_1)/\Psi_2(\beta_2))$ to present the differences of two schemas as shown in the equation **(1)** and **(2)**.

Therefore when two classes in two databases schema definition refers to the same concept, then one global class definition must be created to represent two classes in the component schemas. For an example in **Figure 5** generalized water body is a superclass of water and ocean.

In **Figure 6** the global class definition water body inherits two local specialized class definitions river and lake with their associated attributes namely distance, height and area respectively.

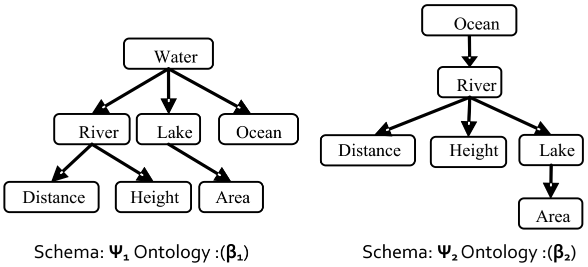


Figure 5. Different Ontology Describing Water Body Entity

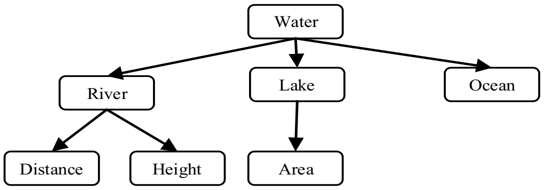


Figure 6. Merging by Finding Equal Concept

Semantic has been defined here as mapping between an object being modeled, represented and/or stored in an information system to represent a real world object(s). This mapping represents the semantic of the object being modeled by describing or identifying the meaning and the user perspectives. Generally, ontology concerns about what kind of things exist – what entity (real things) they are in universal; meanwhile in information technology, ontology can be understood as description of the working model of entities and interaction in some particular domain of knowledge or practices. In general ontology in this work could be understood as the specification and conceptualization used to help program and human to allocate knowledge in order to generate an agree upon vocabulary for exchange of information

As discussed in [1] three main methods could be applied to develop the ontology, and the component of the ontology integration system consists of global ontology, local ontology, and the mapping between local and global ontology. Ontology with typed of multi-valued attributes may connected by binary, symmetric, many to-many roles while attribute and roles can be inherited through inheritance (is-a) links between classes. The semantic mapping could be broken into three phase namely (1) Identifying the concept (2) Brainstorming and (3) Categorization.

An example we identify river which is a part of water system as a concept to be semantically mapped. The next phase is an attempt to explain how we could integrate new information with our existing framework of knowledge.

Categorization is to identify the relationship among the new information to the existing knowledge to form the above schemata map. Two semantic relation play importance role in the specification of ontology called *is-a relation* and *part-whole* relations. *Is-a relationship* is to indicate that one class is a subclass to another class. *Part-whole relationship* indicates that one or more of the object is part of another object

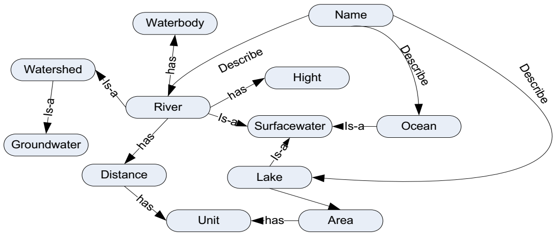


Figure 7. Common Ontology for Water Body

Ontology allows user to convey queries in their own terms according to their own conceptualization without having to known the underlying modeling and representation of data in heterogeneous databases. Concept used by the user in a query can be then compared in order to search not only for what the user has explicitly requested but also for semantically similar terms. These concepts are compared at the ontological level where there is a more complete description of the semantics of terms.

Two entities are similar means nothing, since similarity cannot express categorization unless we understand how the similarity is processed with respect to what property or properties they are similar. Using set theory, similarity [6] *S*, is measured in term of a matching process; this measurement produces a similarity value that is the result of common as well as different characteristic of class.

It is calculated as a function of common and different features as $S(\beta1, \beta2)$. In $S(\beta1, \beta2)$, $\beta1$ and $\beta2$ are two entity classes, $\Psi1$ and $\Psi2$ correspond to the description sets of $\beta1$ and $\beta2$, and $\beta3$ is the first class that subsume $\beta1$ and $\beta2$ by the is-a or whole-of (has) relation.

$$S(\beta1,\beta2) = \frac{(\psi1 \cap \psi2)}{(\psi1 \cap \psi2) + \alpha(\beta1,\beta2)(\ \psi1|\psi2) + (1 - (\beta1 \cap \beta2)|(\psi2|\psi1))} \qquad \textbf{(1)}$$

Where,

$$\alpha(\beta1,\beta2)\begin{cases} \dfrac{distance(\beta1,\beta3)}{distance\ (\beta1,\beta2)} & if\ distance\ (\beta1,\beta3)\ \le distance\ (\beta2,\beta3) \\ 1 - \dfrac{distance(\beta1,\beta3)}{distance(\beta1,\beta2)} & if\ distance\ (\beta1,\beta3)\ > distance\ (\beta2,\beta3) \end{cases} \qquad \textbf{(2)}$$

$0 \le \alpha \le 1$, *β1* and *β2* is term of comparison where *β1* is the target and *β2* is the base. Cardinality is denote as ‖ and α is a function defines the relative importance of non common characteristics.

3. XML – GML Application

As presented in **Figure 8**, the XML and GML technologies offer significant advantages in the data exchange between interoperable systems due to its flexibility and richness in data representation. Translation and transformation enable source data to be converted to target data as desire by the requester with correct semantic meaning. Since GIS spatial data are stored in hierarchy database, Object relational mapping among database is more suitable. This could model XML document as a tree of objects that are specific to the data in the document.

Since XML enable cross platform communication, to achieve interoperability we need such a method to standardizing the service request and response into an XML-based document that will be exchanged among service providers and service requesters will significantly improve the communication and implementation process for interoperable GIS services. As mention there are so many software and hardware systems in market employable for GIS users to represent and model certain types of phenomena they created. Thus we need technique to represent all the schemas exist into one command way that enable everybody shares those data.

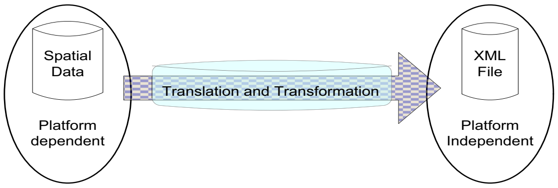
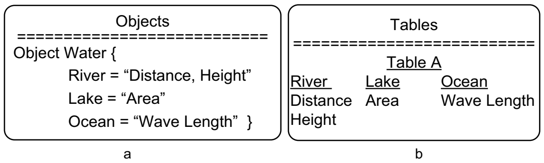


Figure 8. Spatial Data Transformation and Translation to XML File

Mapping is command as a basis for transferring database to XML-GML and from XML-GML to database. Two type of mapping are possible, table-base mapping and object-relational mapping. However object-relational mapping more suitable due to capable handling huge subset of XML-GML document.

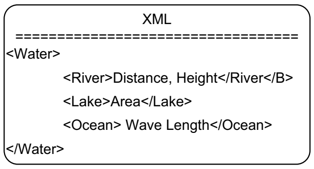
Section bellow provides overview presentation of XML-GML schema, Object Oriented Database (OOD) schema and Relational Database Management System (RDBMS) schema. **Figure 9-Figure 10** and **Figure 11** below shows the schema for object defined in Object Oriented Database and Tables based on merging ontology as in **Figure 7**.



```
Objects
===========================
Object Water {
        River = “Distance, Height”
        Lake = “Area”
        Ocean = “Wave Length”  }
```

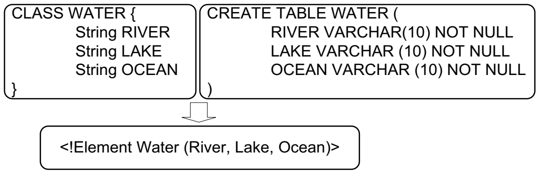
```
Tables
==========================
Table A
River      Lake     Ocean
Distance   Area     Wave Length
Height
```

Figure 9. Schemas for Merging Ontology (a) OOD Schema (b) Table Schema



```
XML
=================================
<Water>
        <River>Distance, Height</River</B>
        <Lake>Area</Lake>
        <Ocean> Wave Length</Ocean>
</Water>
```

Figure 10. Schema of XML for Merging Ontology



```
CLASS WATER {                 CREATE TABLE WATER (
        String RIVER                  RIVER VARCHAR(10) NOT NULL
        String LAKE                   LAKE VARCHAR (10) NOT NULL
        String OCEAN                  OCEAN VARCHAR (10) NOT NULL
}                             )

<!Element Water (River, Lake, Ocean)>
```

Figure 11. Mapping between DTD, Classes and Table Schema

OOD and Table are able mapped to XML document. Correspondingly there is apparent mapping between Document Type Data (DTD), Classes and Table Schema as shown above.

4. Concluding Remarks

Database schema integration has been proposed to realize interoperable GIS application with capability to provide a uniform access to multiple heterogeneous information sources. This research relates to the development of mediator to simplify relocate geospatial data regardless of platform and format used by requester. This implies that database schema integration is the major activities to accomplish this vision. The translation and transformation with XML and GML technologies make possible to cross different GIS platforms and finally to establish seeming less information sharing between different format, database schemas as well as semantic heterogeneity.

5. References

[1]. A. Buccella, A. Cechich, and N. R. Brisaboa, An ontology Approach to Data Integration, *Journal of Computer Science & Technology* , Universidad Nacional De La Plate, vol 3 No 2, (Oct 2003).

[2]. C. C. Pan, P. Mitra, P. Liu, *Semantic Access Control for Information Interoperation,* Proceedings of the eleventh ACM symposium on Access Control Models and Technologies, pp. 237 – 246 (2006).

[3]. I.R. Cruz, H. Xiao, F. Hsu, *An ontology-based framework for XML Schematic Integration*, Proceedings of International Database Engineering and Applications Symposium, 2004. IDEAS '04, (7-9 July 2004), pp. 217 – 226 (2004).

[4]. D. Bonino, F. Corno, L. Farinetti, and A. Bosca, *Ontology Driven Semantic Search*, WSEAS Transaction on Information Science and Application, Issue 6, vol: 1, (December 2004), pp. 1597-1605 (2004)

[5]. M. Jones, and G. Taylor, *Metadata Spatial data handling and integration issues*, School of Computing, University of Glamorgan, Technical Report Series, CS-03-01, (2003)

[6]. M. A. Rodriguez and M. J. Egenhofer, *Determining Semantic Similarity among Entity Classes from Different Ontologies,* IEEE Transactions on Knowledge and data Engineering, vol 15, no 2, March/April (2003)

[7]. M. Gao, C. Liu and F. Chen, *An Ontology Search Engine Based on Semantic Analysis*. Information Technology and Applications, 2005. (Third International Conference, ICITA 2005 July 4-7, 2005, vol: 1, pp. 256- 259 (2005)

[8]. R. Motz, *Problems in the Maintenance of a Federated Database Schema*, Computer Science Society, 2002. SCCC 2002, Proceedings of 22nd International Conference of the Chilean on Nov, 6-8. 2002, pp. 124 – 132 (2002)

[9]. N. Suryana and N. A. Ghani, *Role of Clearinghouse Server Infrastructure for GIS Interoperability, Business Information System*. GI Edition, Lecture Note in Informatics. Abramowicz and Mayr (eds.) Vol P-85. ISBN3-88579-179-X. ISSN 1617-5468. Klagenfurt Austria. June 2006.