A GML SCHEMA MAPPING APPROACH TO OVERCOME SEMANTIC HETEROGENEITY IN GIS

Manoj Paul, S. K. Ghosh

School of Information Technology, Indian Institute of Technology,
Kharagpur 721302, India - (mpaul, skg)@sit.iitkgp.ernet.in

Commission VI, WG IV/2

KEY WORDS: Application Schema Mapping, GIS interoperability, Ontology

ABSTRACT:

With the increasing need of spatial data in various decision support systems, the access and sharing of geospatial data over the Internet has become an important issue. But the underlying heterogeneity in geospatial data syntax and semantics are the major bottleneck towards this direction. The present standardization approach suggests that there should be a core schema for each of the data repositories providing the metadata information of the data sources. The development of the core schema at the organizational level, at the national level has also been thought of. Open Geospatial Consortium (OGC) has addressed the heterogeneity problem and defined several standards for data sharing and accessing. GML has been proposed by OGC to be the standard interoperable data format. This underlying data structure for GML data has been termed as application schema. In this paper we propose an approach for semantic based matching of the application schemas across several data repositories both at the element level and structure level. Through this mapping methodology, data conforming to one schema can be exported to the other schema. The application of ontology has been utilized to generate the mapping rules from one schema to the other. This will subsequently help in converting data in one schema format to the other.

1. INTRODUCTION

With the advent of spatial information in various decision support system and the non-availability of the same in uniform fashion has raised the issue of standard means of sharing and utilizing the spatial information in mutually beneficial manner. The underlying heterogeneity in geospatial data syntax and semantics and lack of standard model for the data repositories are the major bottleneck towards this direction (Kim 1991). The present need is to standardize the structure of the information available in the repository in the form of conceptual schema, which will support interoperability among these data sources by easing the sharing the data. Every individual data providers need to publish this conceptual schema for increasing the accessibility of the data source. Although this can resolve some of the problem, some other problems are pertinent to arise due to heterogeneity. There are many levels of heterogeneity – syntactic heterogeneity, structural heterogeneity and semantic heterogeneity. Although standardized, the heterogeneity in schema of data repositories arises due to various reasons – *naming conflict* arising due to homonyms and synonyms, *scaling conflict* arising due to different measure units. Unless a suitable method for schema mapping is found, interoperability will be far from realization.

As pointed out that the present standardization approach suggests that there should be a core schema for each of the data repositories that can provide the metadata information of the data sources. The development of the core schema at the organizational level, at the national level has also been thought of. When an organization wants to export data form a data repository with the use of the repository schema into the repositories of data of the organizations that happens to have different schema. Since the schema, in addition to the syntactic nature of the data, also holds the semantic information, this issue also raises correspondence of semantic aspects among the data sources. This is a natural problem occurring while integrating distributed data repositories because the schema of the repositories have been developed independently and different organizations may have used different terminology while developing the schema of the same domain. The problem of semantic heterogeneity arises due to the heterogeneity in terminology in the schema of the data.

Open Geospatial Consortium (OGC) has addressed the heterogeneity problem and defined several standards specifications for data sharing and accessing. Geography Mark-up Language (GML) (Cox 2003) has been proposed by OGC to be the standard interoperable data format for sharing and transport of spatial data. The structure of GML data can be formulated using DTD (Document Type definition) or XML Schema. This underlying data structure for GML data has been termed as application schema. GML defines some base schemas like geometry schema, feature schema etc. from which an application schema can be constructed. Application schema declares the actual feature type for a particular domain. The application schema itself could be semantically heterogeneous even if developed in the same organization but by different personnel.

In this paper we propose an approach for semantic based matching of the application schemas across several data repositories. A fundamental operation for schema mapping is *match*, which takes two schemas as input and produces a mapping between elements of the two schemas that correspond semantically to each other (Rahm 2001). This may be termed as mapping rules. The generation of the rules takes into account the use of ontology for the domain. The difference in terminology can be matched based on the use of any semantic thesauri like WordNet. Once these rules can be built up, data export from one schema format to the other becomes just a matter of parsing.

The rest of the paper is organized as follows: a background on the related on work in this area has been given in section 2, section 3 provides the detail of the methodology for GML application schema mapping, and finally a conclusion has been drawn in section 4.

2. RELATED WORK

The mapping of the schema for sharing data among the users is of utmost importance. This is an essential for achieving interoperability among multiple data sources (Rahm 2001). Unless there is some way of mapping the schema, sharing of data can not be realized. Schema mapping has been studied extensively in the context of machine learning, information retrieval, data warehousing, and E-commerce etc. The approaches proposed for schema mapping describe the need and methodology for *Linguistic mapping, similarity measurement, Structural mapping* etc. Support in identifying similar parts between two schemas can be maximised using a combination of such matching approaches. There are several works in the literature emphasising on the techniques and method for achieving interoperability in GIS (Rahm 2001, Bergamaschi 2001, Guan 2003).

The existing approaches for schema mapping are at schema-level and instance-level, element-level and structure-level, and language-based and constraint-based matchers (Bergamaschi 2001, Guan 2003). Currently, schema matching is typically performed manually, perhaps supported by a graphical user interface. Obviously, manually specifying schema matches is a tedious, time consuming, error-prone, and therefore expensive process (Guan 2003). The available present tools for schema mapping e.g. Clio project (Howard 2001), depend heavily on human intervention, as they need lots of manual processing. We propose an approach focusing on the semantics of the schema. The source and target schema can be compared to identify the required parts for translation. Although the approach of (Guan 2003) is close to the one proposed in this paper, they don't consider the structural mapping in the schema hierarchy and restricts the mapping only at the element level. We are much influenced by the element level matching and applied almost the same idea for our purpose also.

This paper is focused on finding a suitable method for GML application schema mapping with a framework for the proposed mapping approach. Schema mapping is a major research issue for the past few decades. It takes two schemas as input and produces a mapping between elements of the two schemas that correspond semantically to each other. Both schema level and instance level mapping can be. To integrate different GML application schemas, we are proposing method for schema level mapping, which only considers schema information, not the data instance of the schema.

3. THE SCHEMA MAPPING APPROCH

The proposed method poses no restriction on the data producer in producing data. Local level data can be produced following the local schema model. But the method proposes to restrict the distribution and sharing of data in standard and commonly agreed form. This is to say, a method is necessary for enabling the mapping of the schema of the data from local level to a global level of agreed upon standards. Even a common schema is followed for local level data generation it is pertinent to change due to some modifications on the data requirement. Thus schema mapping is the utmost requirement for interoperability. The framework of the system for schema matching is shown in figure 1. The *Schema Comparator*, with the help of Ontology description, generates mapping rules. These rules are then used by the *GML Converter* for converting the source GML document into the target GML document.

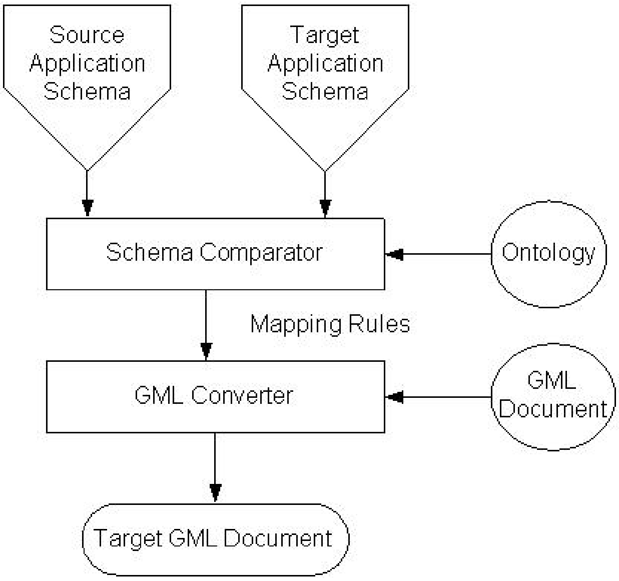


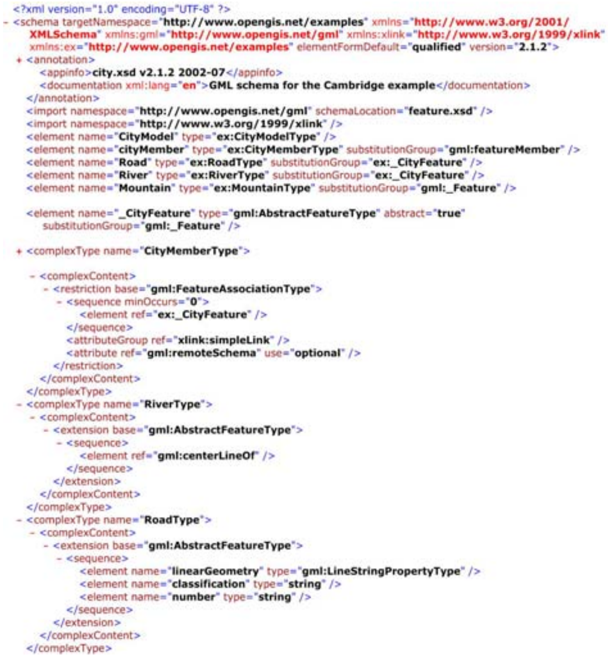
Figure 1. The Schema mapping approach

3.1 GML Application Schema

GML is a mark-up language, which means that GML documents have to follow certain rules in order to create valid GML documents. These set of rules are defined in a schema document. The documents should conform to the requirements in the GML specification. The analysis of application schema can be thought of as the pre-processing step for finding a possible mapping in the relational schema. GML version 1.0 uses the Document Type Descriptors (DTDs) for defining the elements and their associated attributes. GML version 2.0 and 3.0 use GML schemas instead of DTDs.

A GML schema is written in XML. The elements and their attributes used in a GML document must be defined in the related GML schema for the document to be valid. A GML schema provides a set of type definitions and element declarations that can be used to check the validity of well-formed GML documents (Cox 2003). An example schema is shown in figure 2. The UML[1] model for the schema is also shown in figure 3. GML defines various entities such as features, geometries and topologies through a hierarchy of GML objects. GML specification provides a series of schema for describing geographic data. These include feature, geometry, topology, value, coordinate reference system, and style-descriptor etc (Bohannon 2002). The use of base schemas provides the flexibility in using GML to represent diverse types of spatial objects. Most applications make use of only a subset of the schemas that have been defined in the GML specification

[1] Unified Modeling Language: www.uml.org



```
  <?xml version="1.0" encoding="UTF-8" ?>
- <schema targetNamespace="http://www.opengis.net/examples" xmlns="http://www.w3.org/2001/
    XMLSchema" xmlns:gml="http://www.opengis.net/gml" xmlns:xlink="http://www.w3.org/1999/xlink"
    xmlns:ex="http://www.opengis.net/examples" elementFormDefault="qualified" version="2.1.2">
  + <annotation>
      <appinfo>city.xsd v2.1.2 2002-07</appinfo>
      <documentation xml:lang="en">GML schema for the Cambridge example</documentation>
    </annotation>
    <import namespace="http://www.opengis.net/gml" schemaLocation="feature.xsd" />
    <import namespace="http://www.w3.org/1999/xlink" />
    <element name="CityModel" type="ex:CityModelType" />
    <element name="cityMember" type="ex:CityMemberType" substitutionGroup="gml:featureMember" />
    <element name="Road" type="ex:RoadType" substitutionGroup="ex:_CityFeature" />
    <element name="River" type="ex:RiverType" substitutionGroup="ex:_CityFeature" />
    <element name="Mountain" type="ex:MountainType" substitutionGroup="gml:_Feature" />

    <element name="_CityFeature" type="gml:AbstractFeatureType" abstract="true"
      substitutionGroup="gml:_Feature" />

  + <complexType name="CityMemberType">

    - <complexContent>
      - <restriction base="gml:FeatureAssociationType">
        - <sequence minOccurs="0">
            <element ref="ex:_CityFeature" />
          </sequence>
          <attributeGroup ref="xlink:simpleLink" />
          <attribute ref="gml:remoteSchema" use="optional" />
        </restriction>
      </complexContent>
    </complexType>
  - <complexType name="RiverType">
    - <complexContent>
      - <extension base="gml:AbstractFeatureType">
        - <sequence>
            <element ref="gml:centerLineOf" />
          </sequence>
        </extension>
      </complexContent>
    </complexType>
  - <complexType name="RoadType">
    - <complexContent>
      - <extension base="gml:AbstractFeatureType">
        - <sequence>
            <element name="linearGeometry" type="gml:LineStringPropertyType" />
            <element name="classification" type="string" />
            <element name="number" type="string" />
          </sequence>
        </extension>
      </complexContent>
    </complexType>
```

Figure 2. An example application schema

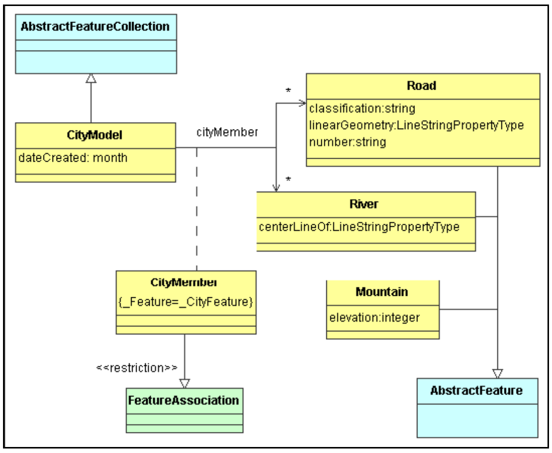


Figure 3. UML model for the application schema of city

3.2 Ontology and Schema Mapping

Ontology has been introduced in AI as an explicit specification of a conceptualization; therefore it can be used to describe the semantics of the information sources and to make the content explicit with respect to the integration tasks (Wache 2001). An ontology is an explicit, formal specification of a shared conceptualization of a domain of interest (Fensel 2001). It can be defined in many ways that suits the need of its purpose. For our purpose we can define it as consisting of schema and metadata.

$$\mathbf{O: = (C, H_C, R_C, H_R)}$$

An ontology O is a tuple consisting of the following. The concepts C of the schema are arranged in a subsumption hierarchy H_C. Relations R_C exist between single concepts. Relations can also be arranged in a hierarchy H_R.

Although the structural mapping of GML application schema to relation schema can be a solution for geospatial data storage, the semantic heterogeneity in the schema itself can only be resolved by the utilization of a shared vocabulary for the domain in the form of ontology. There are several causes for semantic heterogeneity arising due to confounding conflicts and naming conflicts. Confounding conflicts occur when information items seem to have the same meaning, but differ in reality. Naming conflicts occur when naming schemes of information differ significantly.

We examine the applicability of ontology for the schema matching purpose. In general, there are three different possible ways of how ontologies are employed; single ontology approaches, multiple ontologies approaches and hybrid approaches (Wache 2001). Single ontology approach uses a global ontology to give a shared vocabulary for semantics' specifications. Multiple ontologies approaches use separate local ontology for each information source, which can simplify the integration task and supports the change, but increase the difficulties to compare different source ontologies. We adopt Hybrid ontology approaches, which is the combination of single and multiple ontology approaches, in which the semantics of each source is described by its own ontology and a global shared vocabulary is also built on the local ontologies. An ontology structure for the domain *University* is shown in figure 4. The concepts provided model entities of interest in the domain. They are typically organized into a taxonomy tree where each node represents a concept and each concept is a specialization of its parent. It is a data descriptive language, which means that the data is stored in a self-descriptive manner.

Many formal languages to specify ontologies have been proposed for the Semantic Web, such as DAML+OIL, RDF (Brickley 2000). Though these languages differ in their terminologies and expressiveness, the ontologies that they model essentially share the same features we described above.

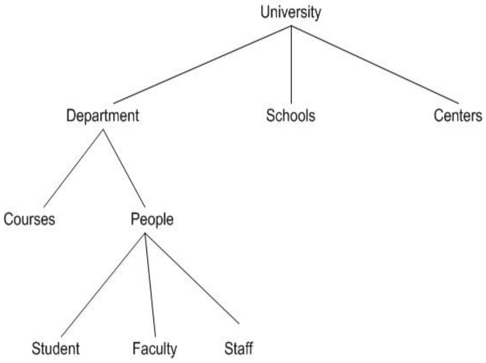


Figure 4. The Ontological Hierarchy for feature matching

3.3 Mapping Rules

Ontology has found much use in recent times for the conceptual description of a system. We investigate the use of ontology for the efficient integration of schema. These schema-mapping approaches depend on the mapping at the structural level as well at the element level. A conceptual model can also be used for finding the similarities in the schema. In this context ontology-based conceptual model can be used effectively. Ontologies provide a possibility to manage heterogeneous schemas, because they are generic in respect to applications within a certain domain and enable more automation in schema mapping by providing rigor similarity measurements. As schemas continuously evolve and new data sets are introduced due to change of user requirements, the effort of developing ontologies can turn out to be beneficial in the long term.

The proposed approach for schema mapping is concerned on the mapping of GML application schemas, which is done at two levels: element level and structural level. We will explain these two procedures in the next two subsection. The need of auxiliary information sources such as dictionaries, thesauri for reusing any previous mappings has also been discussed in the paper. Reuse-oriented approaches are promising, since we expect that many schemas need to be matched and that schemas often are very similar to each other and to previously matched schemas. On the other hand, structural matching is concerned with the matching combinations of elements as a hierarchy structure using the equivalence pattern form the ontologies. A library of equivalence pattern is used for this type of matching.

3.3.1 Element Matching: In this method each element of the schema is mapped against the meaning of the terms and mapping rules are generated. Certain mapping rules are defined according to the information available in ontologies for element level mapping. The mapping procedure basically considers the similarity of the terms on the basis of synonyms, hyponyms etc. As an example the term *Road* is similar to the term *Street.* Standard thesauri like WordNet can be used for this purpose. The abbreviated terms like *Cust_id* is expanded to *customer identity* and subsequent rules are generated to indicate that these elements bear the same meaning. The composite terms are also broken into tokens and those are used for mapping purpose. As an example *AirportWeatherReport* is broken into tokens *Airport, Weather and Report,* which are used for the mapping purpose.

The granularity of the mapping can be at atomic level elements e.g. elementary attributes like place name, address, geometry or at higher level elements like complex times in XML schema. We also consider the cardinality of matching while performing the element level mapping (Guan 2003). While 1:1 mapping gives the terms in schemas which are different but semantically similar, 1: n or n: 1 mapping gives rules which corresponds to multiple elements from a schema to a single element in the other schema and vice versa.

3.3.2 Structural Matching: The *Feature Mapping* step finds mapping of the GML schema features into corresponding to one schema to that of the other. The *Feature Mapping* process matches the structural similarity of a feature in terms of its sub-feature to the structure of the relational schema. Thus we need to match the sub-concepts under the feature concepts as well. The mapping procedure exploits both semantic and syntactic mapping for this purpose. The structure matching can be at three different levels.

- At the *Direct matching,* structure of one feature in the schema is mapped to one in the other schema. As an example the structure of the features *road, street, path, highway, lane* could be matched directly for finding similarities by applying the element matching procedure.
- In the *Sub-Super class matching,* a feature of one schema is mapped to the any of the subclasses of the other schema in the schema hierarchy. As an example *Department* in one feature in one schema can be mapped to the subclass of the schema feature *University* in the other.
- *Super-Sub Class matching* corresponds to just the reverse mapping of the *Sub-Super class matching.* A *Employee* in one schema can be matched with a super class of the feature *Faculty* in the other.

The matching procedure directly resembles the use of ontology for the matching purpose. We have used ontologies as basis when developing application schemas as to support automated schema mapping. The hierarchical matching is efficient using ontology as this corresponds directly to the hierarchical conceptual structure of the ontology. This enables to infer implicit taxonomic relationships and equalities automatically and this will be used to match schema elements. The overall procedure stands to be *Structure Matching* followed by *Element Matching* of the matched structure. The usability of domain dependent ontology is achieved for the structural level matching while domain independent ontology has been used for the element level matching.

4. CONCLUSION

With the growing demand of sharing and standardizing data repositories, the need of standardizing application schema itself has been put on great concern. Each of the data repositories should publish the application specific schemas for enhancing the sharing of information to the third party. Since the application schema itself is developed independently, the schemas are bound to be semantically heterogeneous.

In this paper we have proposed an approach for mapping the various schemas at the semantic level and generate some rules out of the schemas with application of ontology. Once these mapping rules are derived, the conversion of data in conformance to one schema to that of the other can be done on the basis of these rules. We have derived these rules at the element level as well as at the structural level. The schema mapping approach throws light on the interoperability aspects at the schema level.

5. REFERENCES

Kim, W., Sea, J., 1991. Classifying Schematic and. Data Heterogeneity in Multidatabase Systems. *IEEE Computer*, 24(12):12-18, 1991.

Open Geospatial Consortium (OGC). http://www.opengeospatial.org/.

Guan, J., Zhou, S., Chen, J., Chen, X. et al., 2003. Ontology-based GML schema matching for spatial information integration. *Proceedings of the Second International Conference on Machine Learning and Cybernetics,* Xian.

Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G. Et al., 2001. Ontology-Based Integration of

Information A Survey of Existing Approaches. *IJCAI Workshop: Ontologies and Information Sharing.*

Fensel. D., 2001. Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce. Springer-Verlag publication.

Rahm E, Bernstein P A., 2001. A Survey of approaches to automatic schema matching. *The VLDB Joumal*, 10 334-350

Bergamaschi, S., Castano, S., Vincini, M., Beneventano, D., 2001. Semantic integration of heterogeneous information sources. Data Knowledge Engineering, 36(3):215-249

Cox, S., Daisey, P., Lake, R., Portele, C. and Whiteside, A., 2003. OGC Geography Markup Language (GML 3.0) Implementation Specification, OGC Specifications http://www.Open gis.org/specs/?page=specs.

Brickley, D., and Guha, R., 2000. Resource Description Framework Schema Specification 1.0

Howard, C. T., Ronald, H., Popa, Fagin L. et al., 2001. The Clio Project: Managing Heterogeneity, in ACM SIGMOD Record 30, 1, pp. 78-83.

Bohannon, P., Freire, J., Roy, P., and Simeon, J., 2002. From XML Schema to Relations: A Cost-Based Approach to XML Storage. In the Proceedings of International Conference on Data Engineering (ICDE).