Extending XML Web Querying to Heterogeneous Geospatial Information

Nancy Wiegand, Naijun Zhou and Stephen Ventura
Land Information and Computer Graphics Facility
University of Wisconsin, Madison, WI 53706
wiegand@cs.wisc.edu, nzhou@wisc.edu,
sventura@wisc.edu
www.lic.wisc.edu/DG_Project/DGhomepage.html

Isabel F. Cruz
Computer Science Department
University of Illinois at Chicago
Chicago, Illinois 60607
ifc@cs.uic.edu
http://www.cs.uic.edu/~ifc/grants/DG/

Abstract

This paper describes a Web-based query system for semantically heterogeneous geospatial data. Our goal is to provide DBMS type query capabilities to a proposed statewide land information system. One of the main problems in querying distributed local data sources is the difference in semantics describing the characteristics (attributes) of spatial objects between various jurisdictions. To address this problem, we developed a mapping technique and tool to resolve semantics at the value level. Semantic resolution is incorporated into an XML Web-based DBMS. Our method works for any heterogeneous set of values, but we use land use codes from multiple classification systems as an example.

1. Introduction

Our goal is to support full-fledged Database Management System (DBMS) querying over heterogeneous distributed Web data sources. We are working in the context of a proposed Wisconsin Land Information System (WLIS, 1999). WLIS will be a statewide system with Web-based access to distributed data sets residing on servers under local control. It will contain geospatial data in GIS formats and nonspatial data such as documents. We are extending the initial clearinghouse vision of the WLIS working group to provide a system with full query support over WLIS data. However, one of the main problems for comprehensive querying across jurisdictions is that locally produced data sets tend to be highly heterogeneous.

This paper presents our Web-based query system that resolves semantic heterogeneities between data sets. Section 2 discusses the semantic heterogeneity problem at the value level using land use codes as an example. Section 3 presents our existing working system. A summary is given in Section 4.

2. Semantic heterogeneity at the value level

Syntactic, *schematic, and semantic* DBMS heterogeneities have been addressed at the schema level (e.g., Bishr 1998; Bouguettaya et al., 1998). Semantic heterogeneity concerns discrepancies in the meaning, interpretation, and intended use of the same or related data (Sheth and Larson, 1990). Semantic heterogeneity can also exist at the value level with three types of conflicts, according to (Bouguettaya et al., 1998). These are: differences in expression (e.g., 4.0 vs. A), differences in units (e.g., miles vs. kilometers), and differences in precision (cardinality differences, e.g., low, medium, high vs. a range with 5 choices). Our land use application exemplifies these conflicts. Differences in expression occur (e.g., Agriculture codes beginning with “A” vs. beginning with “9”). Also, units and precisions vary (e.g., hectares vs. acres and 6 subcategories for Agriculture vs. 11).

In addition, we found another type of data level heterogeneity: differences in *categorization.* For example, a coding scheme for the Commercial category that is divided into “Commercial Sales” and “Commercial Services” cannot easily be compared to another code scheme divided into “Commercial Intensive” and “Commercial Nonintensive”.

Historically, a standard coding system was never imposed, and individual communities preferred to develop land use codes that more closely represented the particular land uses in their area.

A multi-dimensional coding system called the Land Based Classification System (Everett and Ngo, 1999) was developed by the American Planning Association to help provide a standard. However, to date, it has not been widely adopted.

Table 1 shows example land use codes used in Wisconsin. A query to find all cropland over a watershed that spans several counties is problematic because the meanings of codes vary in each jurisdiction. For example, the 8110 code of the City of Madison makes no distinction between cropland and farm buildings, whereas the Dane County Regional Planning Commission has a separate code for farm buildings (93). Eau Claire County's most specific code that would include cropland is at the general agriculture level which also includes various other subcategories such as dairying. A system to automatically make comparisons between diverse code systems is extremely valuable for comprehensive planning (Faella, 2002) because, currently, efforts to resolve codes have to be done by hand.

Planning Authority	Attribute Identifier	Land Use Code	Code Description
Dane County	Lucode	91	Cropland/Pasture
Racine County	Tag	811 815	Cropland Pasture & Other Agriculture
Eau Claire County	Lu1	AA	General Agriculture
City of Madison	Lu_4_4	8110	Farms

Table 1. Heterogeneity in land use codes

3. Method

The following sections describe our system. Our method can be generalized to capture semantic differences among values for any attribute.

3.1 Web-based XML DBMS

The Niagara XML Internet DBMS (Naughton et al., 2001) forms a base for our geospatial query system. Niagara provides DBMS type querying over distributed XML data on the Web. The Niagara Java query engine processes queries in XML-QL (Deutsch et al., 1998).

However, current Web-based query systems, such as Niagara, do not have semantic integration facilities. We added such facilities as shown in our system architecture (Figure 1).

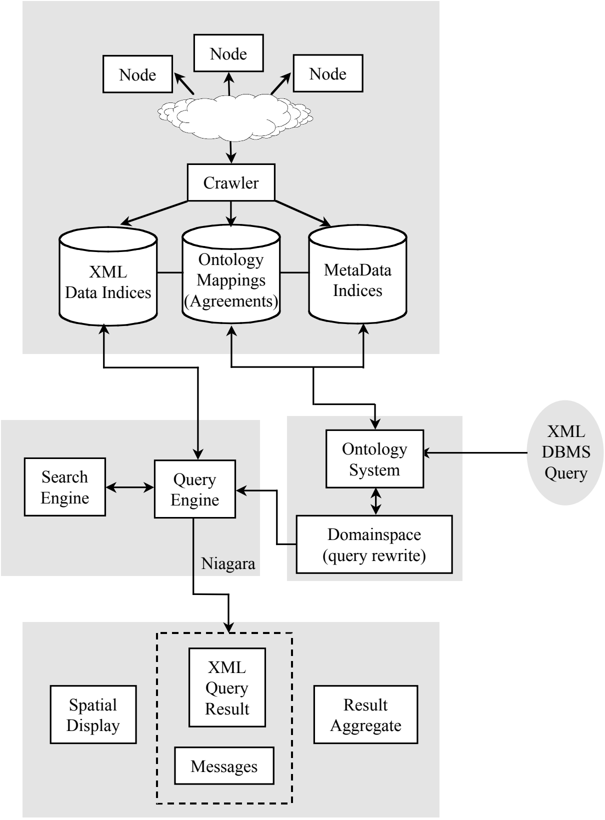

Figure 1. System architecture

3.2 Ontology approach

To achieve semantic integration, we use an ontology approach. Ontologies have been proposed as a solution for semantic integration (e.g., Fensel 2001). Ontology driven GIS has been proposed for geographic information integration, especially between GIS and remote sensing systems (Fonseca et al., 2002). We assume an ontology of land use codes developed by domain experts to satisfy the query needs of users.

The Document Type Definition (DTD) for our ontology database is shown in Figure 2. For now, we focus on an ontology for land use data and, in particular, on the values for the land use field.

```
<!ELEMENT database     (table+)>
<!ELEMENT table         (tuple+)>
<!ELEMENT tuple         (attrname+)>
<!ELEMENT attrname     (attrvalue*)>
<!ELEMENT attrvalue    (attrvalue*)>
<!ATTLIST database  id CDATA #REQUIRED>
```

```
<!ATTLIST table        id CDATA #REQUIRED>
<!ATTLIST tuple        id CDATA #REQUIRED>
<!ATTLIST attrname   id CDATA #REQUIRED>
<!ATTLIST attrvalue  id CDATA #REQUIRED>
```

Figure 2. Ontology DTD

3.3 Land use code ontology

We express the ontology for land use codes using attribute value elements (attrvalue) for the land use code attribute (Figure 3). As shown for the commercial code, various divisions of subcategories can be included in the ontology to help with the problem of different categorizations. In fact, our ontology method can handle any level of precision.

```
<attrname id= “land_use_code”>
 <attrvalue id= “Commercial”>
    <attrvalue id= “Commercial-Scale>
     <attrvalue id= “Commercial-Scale-Intense”/>
     <attrvalue id= “Commercial-Scale-NonIntense”/>
    </attrvalue>
    <attrvalue id= “Commercial-Function>
      <attrvalue id= “Commercial-Function-Sales”/>
      <attrvalue id= “Commercial-Function-Service”/>
    </attrvalue>
        …
    <attrvalue id= “Other”/>
 </attrvalue>
 <attrvalue code= “Residential”>
        …
 </attrvalue>
 …
</attrname>
```

Figure 3. Ontology for land use codes

3.4 XML agreement files

The ontology needs to be mapped to each land use coding system that is part of WLIS. Because of the semantic difficulty in automatically resolving codes, we developed a tool with which a local domain expert establishes the correspondence between the master codes and local codes.

The mapping tool (Cruz et al., 2002) captures the cardinality of the mappings and automatically generates an XML *agreement* file. Mapping types include 1:1, 1:N, N:1, and 1:NULL. XML tags and attributes are used to record the semantics of the mappings. In the agreement file shown in Figure 4, the 1:N agriculture example lists the included local values, and the multi-family example resolves a 1:NULL mapping by specifying a more general level. The information from the agreement files is used to generate subqueries sent into Niagara.

```
<ontology_value code = “Agriculture”
  mapping_to_localcode = “one-to-many” >
   <localvalue> 91 </localvalue>
   <localvalue> 92 </localvalue>
      …
</ontology_value>

<ontology_value code= “Multi-Family”
  mapping_to_localcode= “one-to-null”
  level_up= “Residential”>
</ontology_value>
```

Figure 4. XML agreement file

3.5 DomainSpace

This section explains our query re-writing techniques. Our example query, “*Find all cropland over a watershed that spans several counties*”, is different from a typical DBMS query because more than one data source is identified, but there is no join. Niagara’s “IN*” is not appropriate here because the user restricts the jurisdictions for a query. That is, in our user interface, the user chooses an area over which a query will range. The user also selects a land use code predicate. Our demo is described in (Wiegand et al., 2003).

A formal mechanism is needed to represent the type of DBMS query in which the same predicate is applied to multiple data sets. For this, we developed a DomainSpace concept. We added a DOMAINSPACE statement to the XML-QL query language (Figure 5). We introduce a variable, e.g., “*Area*”, to hold the list of URLs for the data sources needed in the query. The variable is then used in the body of the query as a qualifier for the generic ontology terms.

```
DOMAINSPACE Area = “www.co.wi.us/Dane.xml,

www.co.wi.us/EauClaire.xml”
WHERE <$*>
            <Area:LandUseCode> “cropland” </>
             </>  ELEMENT_AS $a
CONSTRUCT $a
```

Figure 5. DomainSpace in an XML-QL query

To send this query into the XML query engine, we first rewrite it into multiple subqueries expressed in native terms. For example, the subquery pertaining to Eau Claire County is shown in Figure 6.

```
WHERE <$*>
          <lu1> “AA” </lu1>
          </>  ELEMENT_AS $a
IN  www.co.wi.us/EauClaire.xml
CONSTRUCT $a
```

Figure 6. A generated subquery

4. Summary

One of the most difficult aspects of providing query support over distributed data is semantic heterogeneity. We developed an ontology and query rewrite system on top of an XML Web DBMS to handle semantic heterogeneity. We focused our efforts on resolving differences at the value level.

5. Acknowledgements

This work was supported by the Digital Government Program of NSF, Grant No. 091489.

6. References

Y. Bishr, “Overcoming the Semantic and Other Barriers to GIS Interoperability”, International Journal of Geographical Information Science, Vol. 12, No. 4, 1998, pp. 299-314.

A. Bouguettaya, B. Benatallah, and A. Elmagarmid, *Interconnecting Heterogeneous Information Systems*, Kluwer Academic Publishers, 1998.

I.F. Cruz, A. Rajendran, W. Sunna, and N. Wiegand, “Handling Semantic Heterogeneities Using Declarative Agreements”, In Proceedings of ACM GIS, Nov. 2002, pp. 168-174.

A. Deutsch, M. Fernandez, D. Florescu, A. Levy, and D. Suciu, “XML-QL: A Query Language for XML”, 1998, http://www.w3.org/TR/NOTE-xml-ql/.

J. Everett and C. Ngo, “Land-Based Classification Standards-Federal Role.” In Proceedings of APA National Planning Conference, 1999, http://www.asu.edu/caed/proceedings99/LBCS/EVERETT.HTM

T. Faella, Information Technology Manager, East Central Wisconsin Regional Planning Commission, Menasha, WI, Consultation, April, 2002.

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

F. Fonseca, M. Egenhofer, P. Agouris, and G. Camara, “Using Ontologies for Integrated Geographic Information Systems”, Transactions in GIS, 6(3), 2002.

J. Naughton, D. DeWitt, D. Maier, and others. “The Niagara Internet Query System”, IEEE Data Engineering Bulletin, Vol. 24, No. 2, 2001, pp. 27-33.

A.P. Sheth and J.A. Larson, “Federated Database Systems and Managing Distributed, Heterogeneous, and Autonomous Databases”, ACM Computing Surveys, 1990, 22(3):183-226.

WLIS, “Wisconsin Land Information System Technical Report”, Wisconsin Land Council Technical Working Group, 1999.

N. Wiegand, N. Zhou, S. Ventura, I.F. Cruz, and W. Sunna, “Resolving Schema and Value Heterogeneities for XML Web Querying”, Demo**,** In Proceedings National Conference on Digital Government Research, dg.o2003.