Resolving Schema and Value Heterogeneities for XML Web Querying

Nancy Wiegand and Naijun Zhou
University of Wisconsin
550 Babcock Drive
Madison, WI 53706
wiegand@cs.wisc.edu, nzhou@wisc.edu

Isabel F. Cruz and William Sunna
Computer Science Department
University of Illinois at Chicago
Chicago, Illinois 60607
ifc@cs.uic.edu, wsunna@cs.uic.edu

Abstract

To query XML data over the Web, query engines need to be able to resolve semantic differences between heterogeneous attributes that are conceptually similar. This demo presents a mapping tool and method to resolve semantic heterogeneity at the schema and value levels for data sets that are part of a Web-based information system. The mapping tool automatically produces agreement files. We enhanced a base prototype XML Web query system to include an ontology subsystem that generates subqueries using the agreement information. Other contributions include the use of minimal metadata to locate data sets, a formal language construct to support query re-write called a GeoSpace, and post-query aggregate statistics and spatial display.

1. Introduction

Semantic interoperability is necessary for querying distributed data over the Web. Our work is motivated by a proposed Wisconsin statewide land information system that will be a Web-based resource for local, regional, and state data (WLIS) [14]. We extend the clearinghouse vision of the original WLIS working group by incorporating DBMS-type querying over the distributed and highly heterogeneous data sets.

We illustrate our work by integrating and querying data containing land use codes. Land use data is an important component of WLIS because of its value for comprehensive planning decisions. However, land use codes are extremely heterogeneous because there is no standard code system and jurisdictions adapt code systems to emphasize their predominant types of land use.

Although we use land use data in this demo, our method is not limited to that theme. Our framework of semantic mapping and query rewrite can resolve any schema and value level differences. We particularly address the problem of values from heterogeneous domains that cannot be resolved in a straightforward manner. For example, although values in different units of measure can be easily converted, land use values cannot be resolved using a formula.

Related work has resolved heterogeneous schemas at the attribute level, e.g., [1] but has not addressed more complex value level differences. In our work, we demonstrate a method that captures mapping cardinalities and nuances of meaning at the value level.

2. The Semantic Problem

As stated, in addition to schema level mapping, we focus on a type of semantic problem in which formulas or algorithms cannot be used to resolve value level differences between conceptually related attributes in different data sets. We use land use coding systems as an example value domain [11, 13]. Land use coding systems vary by almost every jurisdiction that produces land use data. Example differences in levels of detail and semantics for residential codes between two counties are illustrated in Table 1. As can be seen, categories do not match between code systems.

Table 1. Example Coding Systems

Dane County	Racine County
111 Single Family	111 Single-Family
113 Two Family	120 Two-Family
115 Multiple Family	141 Multi-Family Low Rise (1-3 stories)
129 Group Quarters	142 Multi-Family High Rise (4 or more stories)
140 Mobile Home	150 Mobile Homes
142 Mobile Home Park	199 Residential Land Under Development
116 Farm Unit	
190 Seasonal Residence	

3. Method

The following subsections explain our ontology-enhanced XML query system shown in Figure 1.

3.1 Internet XML DBMS

To provide DBMS-type querying over distributed WLIS data, we use the Niagara Internet XML DBMS [9] as a base for our system. Niagara satisfies the need for general purpose querying over distributed XML data on the Web. However, Niagara does not have semantic integration facilities. To incorporate semantic integration, we modified the Niagara Java source code by adding an ontology subsystem to intercept queries (Figure 1). The ontology subsystem consults metadata indexes and ontology mappings to produce subqueries in local terms.

Our application has a type of query not found in conventional database usage. That is, to accommodate comprehensive or regional decision-making, a typical type of query has a common predicate applied over multiple geographic areas or jurisdictions. An example query for comprehensive land use planning is "*Find all the agricultural lands in Dane and Racine counties*." We call this type of query a GeoQuery because it covers a geographic area.

Niagara's "IN *" capability to range over all elements satisfying a predicate cannot be used here, even if the entire geographic area were specified, because of the heterogeneity of land use data created by independent agencies. Instead, this type of query must be intercepted by our subsystem which generates subqueries for each appropriate data set using semantic mappings.

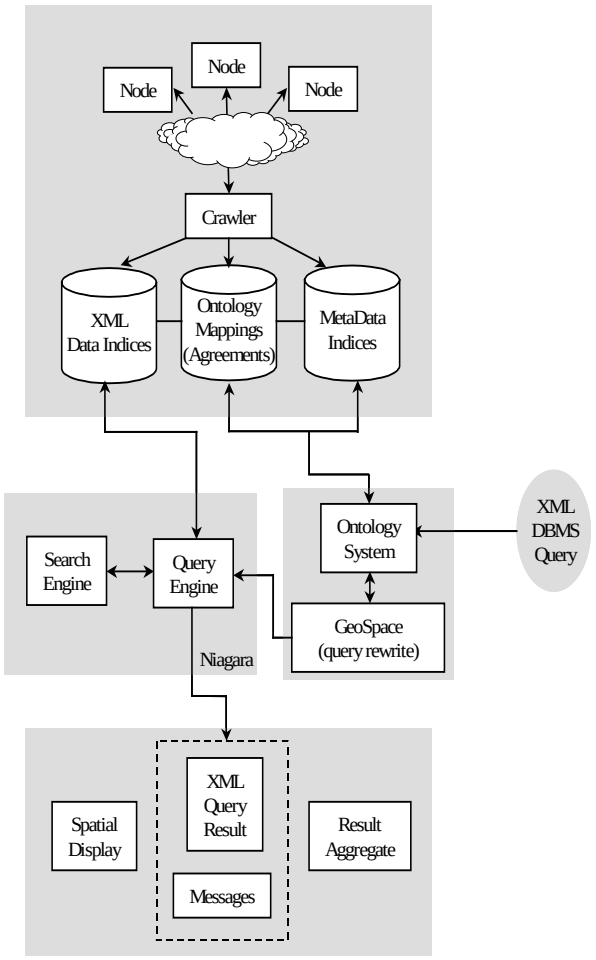

Figure 1. System Architecture

3.2 Ontology Mapping

To solve the heterogeneity problem, we developed an ontology of attributes in the land use theme and a subontology of values for the land use attribute, in particular. The ontologies can be considered to be master sets of terms from which a user can pose a land use query. We developed a tool in which a domain expert indicates schema and value level mappings between the master ontologies and each local data set (Figure 2). At the value level, our method captures the cardinality of the mapping between the ontology value and the local code. The domain expert can specify one to one, one to many, many to one, or one to null mappings. An example of each type of mapping is shown in Table 2.

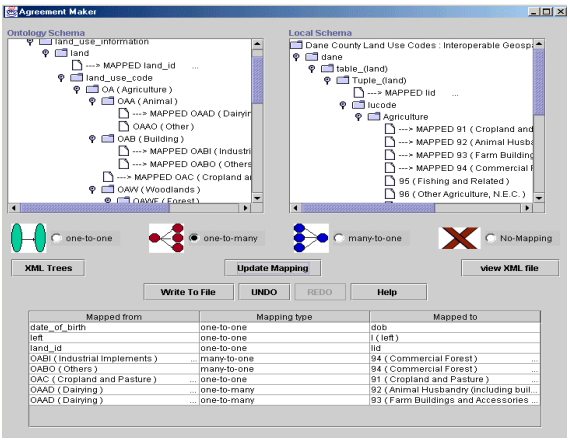

Figure 2. Tool to Create an Agreement File

The mapping tool automatically generates an XML agreement file (Figure 3). As can be seen, semantic information describing the mapping is expressed using the extensibility provided in XML tags. Furthermore, as an option, one to null mappings can be resolved. For example, Table 2 and Figure 3 show a detailed ontology code (multi-family) being resolved to a code at the next level up (residential) for a particular local code set. A complete description of agreement files is given in [2, 3].

Table 2. Value Level Semantic Mappings from the Ontology Codes to Local Codes

Mapping Cardinality	Ontology	Local Code
1 : 1	Cropland	Cropland
1 : N	Agriculture	Cropland, Pasture, etc.
N : 1	Plastics Rubber	Plastics & Rubber
1 : Null (*up a level*)	Multi-family	Residential (*resolved*)
1 : Null (*broader code at same level*)	Commercial Forest	Forest-Other (*resolved*)

```
<Ontology_to_localcode  value = "Agriculture">
   <mapping> one-to-many </mapping>
      <part> cropland </part>
      <part> pasture </part>
       …
</ Ontology_to_localcode >

<Ontology_to_localcode  value = "Multi-Family">
   <mapping> one-to-null </mapping>
      <level_up> Residential </level_up>
</ Ontology_to_localcode >
```

Figure 3. Part of an XML Agreement File

3.3 GeoSpace

To formally represent a GeoQuery, we developed a GeoSpace statement [12] for the XML-QL [4] query language (Figure 4). The GeoSpace statement has a variable that holds the list of URLs needed in the query. The variable also serves as a qualifier for the generic ontology terms in the formal expression of the query.

```
GEOSPACE Area = "www.co.wi.us/EauClaire.xml,
                 www.co.wi.us/Racine.xml"
WHERE <$*>
        <Area:LandUseCode> "agriculture" </>
        </> ELEMENT_AS $a
CONSTRUCT $a
```

Figure 4. GeoSpace Added to XML-QL

To send this query into the XML query engine, the ontology subsystem consults the agreement files to rewrite the formal query into multiple subqueries expressed in native terms. For example, the subquery pertaining to Eau Claire County is shown in Figure 5.

```
WHERE <$*>
        <Lu1> "AA" </Lu1>
        </> ELEMENT_AS $a
IN  www.co.wi.us/EauClaire.xml
CONSTRUCT $a
```

Figure 5. A Generated Subquery

3.4 Other Enhancements

We made further changes to the base XML query system to accommodate heterogeneous geospatial data.

3.4.1 Metadata Indexes

In an information system such as WLIS, users tend to select data sets for queries based on theme and either jurisdiction or spatial extent**.** To identify data sources, we created metadata files for each data set. Our minimal criteria include theme (e.g., land use), jurisdiction type (e.g., city), and jurisdiction name. This information is indexed in separate metadata indexes (Figure 1).

3.4.2 User Interface

Our user interface, shown in Figure 6, is designed to capture the minimal metadata needed

to locate data sources. We also include a spatial ability such that the user can click on a county on a map.

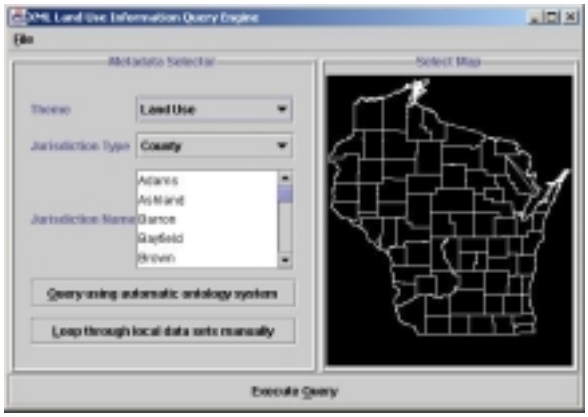

Figure 6. User Interface

3.4.2 User Output--Maps and Messages

Our test data is derived from ArcView [5] files which contain spatial coordinates in addition to nonspatial attribute tables. For each ArcView data set, we combined the spatial and nonspatial information into the same XML file using a feature-based approach. From the spatial data in the query results, MapObjects [5] was used to create spatial displays (Figure 7).

For each data set, we also output semantic information between the ontology code selected in the query and the mapped local code(s) so the user is informed of superset, subset, or resolved null mappings being returned.

Finally, because our potential users almost always asked for displays involving various summary statistics, we processed the client-side results to produce summary information, including sorted results, averages, and counts. For example, the total and average areas coded as agriculture within each jurisdiction are displayed.

Figure 7. Spatial Display

4. Related Work

Ontologies are now being used as a solution for semantic integration [6]. However, most work on ontologies has focused on the schema level and not the value level. Automatically creating ontologies is being explored, for example, in [7]. A use of ontologies in query processing can be found in [10] in which semantic similarities are obtained from a WordNet graph. They also introduced a similarity operator into an XML language. In our application, however, the land use code mappings cannot be found in a collection such as WordNet. Also, we need to hold precise semantic nuance information instead of retrieval relevance rankings. As a result, we needed to develop an automatic or semi-automatic ontology mapping method. Clio [8] represents related work in mapping but is a tool for mapping at the schema level.

5. Demo Description Summary

We are demonstrating a semantic integration query system for heterogeneous data that is built by enhancing an XML Internet DBMS. Our demo has two parts. One part is a tool used to create mappings between ontologies and local data sets (Figure 2). This tool also automatically creates XML agreement files.

The other part of the demo is the overall enhanced XML query system that uses Niagara [9] as a base. Our enhanced interface allows a user to select minimal metadata to determine relevant data sets and themes. The user then uses the appropriate ontology values to pose a query.

The type of query we address is one with a common predicate ranging over multiple data sets. This is typical for a comprehensive planning query that covers a geographic area. Our automated ontology subsystem resolves this type of query (GeoQuery) by generating specific local subqueries using lookups on the agreement mappings and metadata indexes. We formalized a representation of a GeoQuery by introducing a GeoSpace statement into an XML query language. Finally, we process client-side results to create aggregate statistics and spatial displays.

6. Acknowledgement

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

7. References

[1] Bouguettaya, A.; Benatallah, B.; and Elmagarmid, A. Interconnecting Heterogeneous Information Systems. Kluwer Academic Publishers, 1998.

[2] Cruz, I. F. and Rajendran, A. "Semantic Data Integration in Hierarchical Domains", IEEE Intelligent Systems, Vol. 18, No. 2, March-April 2003, pp. 66-73.

[3] Cruz, I.F.; Rajendran, A.; Sunna, W.; and Wiegand, N. "Handling Semantic Heterogeneities Using Declarative Agreements", In Proceedings of ACM GIS, November 2002, pp. 168-174.

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

[5] Environmental Systems Research Institute (ESRI), http://www.esri.com.

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

[7] Malyankar, R. "Vocabulary Development for Markup Languages – A Case Study with Maritime Information", WWW2002, Honolulu, Hawaii, May 2002, pp. 674-685.

[8] Miller, R.; Hernández, M.A.; Haas, L.M.; Yan, L.; Ho, C.T.H.; Fagin, F.; and Popa, L. "The Clio Project: Managing Heterogeneity", ACM SIGMOD Record, 2001, pp. 78-83.

[9] Naughton, J.; DeWitt, D.; Maier, D.; and others, "The Niagara Internet Query System", IEEE Data Engineering Bulletin, 24(2), 2001, pp. 27-33.

[10] Theobald, A. and Weikum, G. "The Index-Based XXL Search Engine for Querying XML Data with Relevance Ranking", In Proceedings of EDBT 2000, Jensen, C.S. (ed.), pp. 477-495.

[11] Wiegand, N.; Zhou, N.; Cruz, I.F.; and Rajendran, A. "Querying Heterogeneous GIS Land Use Data Over the Web", In Proceedings of GIScience 2002 Abstracts, Egenhofer, M. and Mark D. (eds.), Boulder CO, September 2002, pp. 207-210.

[12] Wiegand, N.; Zhou, N.; Patterson, E.D.; and Ventura, S. "A Domainspace Concept for Semantic Integration in a Web Land Information System", Demo, In Proceedings National Conference on Digital Government Research, dg.o2002, 2002, pp. 443-446.

[13] Wiegand, N.; Zhou, N.; and Cruz, I.F. "A Web Query System for Heterogeneous Geospatial Data", Scientific and Statistical Database Management, SSDBM, July 2003, pp. 262-265.

[14] Wisconsin Land Council Technical Working Group. Wisconsin Land Information System Technical Report, 1999.