Integrating Metadata Development, XML, and DBMS Search and Query Techniques in a State of Wisconsin Land Information System

Nancy Wiegand
University of Wisconsin-Madison
wiegand@cs.wisc.edu

Project Overview. In addition to Nancy Wiegand, the PIs on this project are Isabel Cruz at the University of Illinois at Chicago and Stephen Ventura at the University of Wisconsin-Madison. The goal of the project is to provide full Database Management System type querying over government-produced data in a distributed Web-based information system. Current research focuses on the major issue of integrating and querying heterogeneous statewide data. We are using a semantic approach to achieve interoperability.

Internet Querying. We are working in the context of a proposed statewide Web-based information system in Wisconsin [WLIS99]. The Wisconsin Land Information System (WLIS) will allow central access to distributed data sets that remain under local control and reside on local servers across the State. The GIS community was active in promoting the WLIS concept, and much of the data will be geospatial data. Typically, however, geospatial Web sites are limited to searching for geospatial data sets for download. Our first contribution is in promoting to the geospatial community the idea and technology for integrating a query facility over distributed data on the Web (without having the data be resident in a DBMS). That is, we follow and enhance the Niagara Internet XML DBMS approach [NDM+01].

Value Level Data Integration. We are also working on the data integration problem. That is, data sets produced by local units of government tend to be highly heterogeneous. This presents a difficulty when data needs to be compared across jurisdictions as happens, for example, in land use planning efforts. Wisconsin, similar to other states, has passed *Smart Growth* legislation requiring local communities to develop land use plans. One of the most important themes in land use planning is land use. However, land use coding systems vary by almost every jurisdiction. To do comprehensive planning, these systems need to be integrated. Interestingly, resolving land use codes involves working on semantic heterogeneity at the *value level*. That is, within land use attributes, domain values and their meanings differ significantly, as can be seen in Table 1.

Table 1. Heterogeneity in Land Use Codes

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

Ontology Method. Given the semantic differences between the values of land use code systems, it is difficult to use fully automatic methods to resolve discrepancies to the level of precision needed for land use decision-making. Instead, to provide precise answers, we follow the approach of using ontologies as a solution for semantic integration. An ontology is a shared and machine-executable conceptual model in a specific domain of interest [BFM02]. We express ontologies and their mappings in XML. An interesting feature of our approach is that we allow *attribute values* to be part of an ontology in addition to attributes themselves. In this way, we record and map ontology values for problematic attributes such as land use codes. Furthermore, we allow an attribute value element to have nested attribute values to be able to describe any level of subcategories for the hierarchical land use domain. We also allow for multiple

conflicting categorizations to accommodate totally different types of divisions (e.g., Commercial subdivided by intensity in one system and by sales/service in another) and to allow completely different categories (e.g., Permeable and Impermeable in addition to Agriculture, Residential, Commercial, etc.)

We developed a tool that uses an LAV approach to map a global ontology to each local data source. The tool first performs semi-automatic alignment over ontology trees using a deduction algorithm [CR03]. For mappings that cannot be resolved, a local domain expert uses the tool to choose a mapping type between an ontology value and a local value such as exact or approximate. The tool automatically generates an *agreement* file in XML for each source. The agreement file is consulted by the query re-write system to generate a subquery. Another interesting feature of our work is that the semantics of the mappings are given to the user along with query results. For example, if a user requests parcels coded as cropland and the mapped Dane County code is cropland/pasture, the user is informed that the results are a superset.

Web Query System for Heterogeneous Geospatial Data. To provide full DBMS-type querying over distributed WLIS data, we modified Niagara to add an ontology subsystem. The ontology system is used to mediate between the user query and the local data. We also added metadata indexes that contain minimal metadata needed to find data sets involved in a query and ontology indexes for agreement files.

We particularly address a typical type of query that occurs, for example, when land use managers pose a comprehensive query across multiple jurisdictions such as "*Find all cropland over a watershed that spans several counties*". Our user interface captures the selection of multiple geographies and a predicate specifying a land use code. We refer to this type of DBMS query, with the same predicate being applied to multiple data sets, as a G*eoQuery*. To formalize the expression of this query, we developed a GeoSpace concept, which is somewhat similar to the idea of an XML Namespace, and added a GeoSpace statement to XML-QL. A full description of our work is given in [WZC+04].

Broad Impacts. Our research contributes to solving the major problem of transparently accessing heterogeneous information. Our specific work to automate resolution for diverse land use coding systems will enable decision-making that is valuable for land use planning and Smart Growth initiatives. Our full Internet system, that enables querying over distributed diverse data sources, will ultimately allow government officials and citizens to be better informed and able to explore new hypotheses. This Digital Government work has been rewarding, and results are practical.

Websites. http://www.lic.wisc.edu/DG_Project/DGhomepage.html (University of Wisconsin-Madison)
and http://www.cs.uic.edu/~ifc/grants/DG/ (University of Illinois at Chicago).

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

References

[BFM02] Bussler, C.; Fensel, D.; and Maedche, A. 2002. "A Conceptual Architecture for Semantic Web Enabled Web Services", SIGMOD Record, Vol. 31, No. 4, pp. 24-29.

[CR03] Cruz, I.F. and Rajendran, A. 2003. "Exploring a New Approach to the Alignment of Ontologies", Semantic Web Technologies for Searching and Retrieving Scientific Data Workshop, International Semantic Web Conference, Sanibel Island, Florida, pp. 7-12.

[NDM+01] Naughton, J.; DeWitt, D.; Maier, D. and others. 2001. "The Niagara Internet Query System", IEEE Data Engineering Bulletin, Vol. 24, No. 2, pp. 27-33.

[WLIS99] Wisconsin Land Council Technical Working Group. 1999. Wisconsin Land Information System Technical Report.

[WZC+04] Wiegand, N.; Zhou, N.; Cruz, I.; and Sunna, W. 2004. "A Web Query System for Heterogeneous Government Data", In Proceedings of The National Conference on Digital Government Research, dg.o2004.