

Querying Heterogeneous Land Use Data: Problems and Potential

N. Wiegand¹, E. D. Patterson¹, N. Zhou¹, S. Ventura¹, I. F. Cruz²

¹Land Information and Computer Graphics Facility, University of Wisconsin—Madison
wiegand@cs.wisc.edu, http://www.lic.wisc.edu/DG_Project/DGhomepage.html

²Computer Science Department, University of Illinois at Chicago
ifc@cs.uic.edu, <http://www.cs.uic.edu/~ifc/grants/DG/>

Abstract

The proposed Wisconsin Land Information System (WLIS) will be a statewide access mechanism for land related data. It will be a distributed Web-based system with heterogeneous data residing on local and state servers. One of the most important functions for WLIS is to integrate land use data across jurisdictions to enable decision-making for comprehensive land use planning. Here, we describe the semantic problems of integrating land use codes. We then present the current state of our enhancements to an XML query engine to provide query processing over heterogeneous data sets. We also briefly describe the integration of XML databases based on a declaratively specified mapping from each database to a central and expert-defined authority.

1. Introduction

The concept of a statewide information system for land related data has been actively under development for several years by working groups that include various levels of government, the University of Wisconsin-Madison, and private sector (GIS-related) representatives. The Wisconsin Land Information System (WLIS) will be a distributed Web-based system with heterogeneous data residing on local and state servers. The system will include access to multiple forms of data including spatial data (e.g., zoning maps), data tables (e.g., tax assessment listings), text-based documents (e.g., local ordinances), and images (e.g., scanned land titles) maintained by public agencies from local to state levels. It is designed to meet information needs of a broad range of audiences, ranging from land use planning professionals to interested citizens. One of the most important functions for WLIS is to integrate land use data across jurisdictions to enable decision-making for comprehensive land use planning.

The challenge in the WLIS concept is the highly heterogeneous collection of record types, thematic content, level and type of documentation, and computing environments from which data will be drawn. Because data are highly dynamic and because of the politics of information (e.g., local/state government relations), the approach of a massive centralized data repository has been rejected. Instead, the system is expected to draw from as many as 2,000 local, regional, state, tribal, and federal agencies, all of which have made autonomous decisions about records organization and management. Pulling these data together is necessary for making decisions in resource management, environmental protection, transportation planning and management, land use planning, social service delivery, and so forth. The types of data involved in a land information system are extensive and include “... *any physical, legal, economic, or environmental information, or characteristics concerning land, water, ground-water, subsurface resources, or air ...*” (Technical Report 1999).

The problem with integrating data even within one topic (a.k.a. theme) such as land use is the need for conversions in structure, format, and semantics. Land information data developed by local, regional, state, and federal government and the private sector are not homogeneous, and data schemas for commonly used geographic data layers cannot be applied comprehensively across jurisdictional boundaries. As a result, it is difficult to relate data for a given theme from adjacent jurisdictions. For example, a land

valuation application developed for one county will not work with another county’s data even if the same GIS software is used. Although spatial integration can be accomplished by translating the different coordinate systems, attribute data are not compatible. As one moves across jurisdictional boundaries, data tables and attributes vary, and the definitions and values change significantly.

To start our project, we gathered representative land use data and identified specific integration problems. We are adding XML tags to the data so it can be processed across a distributed system by emerging XML search and query engines. However, heterogeneity in the format and meaning of data from various sources requires semantic integration before a goal of full query support can be met. We are enhancing an XML query engine to enable the processing of heterogeneous data and also working on mechanisms to integrate data that will retain as much precision as possible.

2. Land Use Data

We are beginning our project focusing on land use data because of its importance in land use planning and because of the special semantic difficulties involved. To promote regional and statewide land use planning, Wisconsin has recently passed “Smart Growth” legislation in which all communities are required to have a comprehensive land use plan by 2010. One of the goals of WLIS is to make integrated land use data available to users such as local communities working to develop Smart Growth plans.

Land use data generally use parcel databases as a starting spatial framework. Currently, each jurisdiction uses its own land use coding system. To be able to plan across jurisdictional boundaries, data distributed over a wide spatial area and diverse in organization and composition need to be integrated.

2.1 Background

Historically, land use classification systems evolved out of the need to describe certain observable conditions in the landscape. They offer a systematic way to codify and categorize these conditions. Two common examples of land use classification schemes are **exhaustive lists** and **hierarchical models** (Table 1, American Planning Association 1994). Structurally these two systems are quite different, with the hierarchical model being more highly organized than the exhaustive list.

Exhaustive List	Hierarchical
009 Shopping Center	1 Urban and Developed Land
010 Open Water	1.01 Residential
111 Single Family	1.01.01 Single Family Detached or Duplex
113 Two Family	1.01.02 Mobile Homes Not in Parks
115 Multiple Family	1.01.03 Multi-family Dwellings
116 Farm Unit	1.01.03.01 Three Unit Multi-family
129 Group Quarters	1.01.03.02 Four Unit Multi-family
140 Mobile Home	1.01.03.03 Five or More Multi-family
142 Mobile Home Park	1.01.04 Mobile Home Parks

Table 1: Classification schemes

The genealogy of land use classification systems in this country is rich and still evolving. The Standard Industrial Classification (SIC) system could be considered the cornerstone from which subsequent land use classification systems were derived. The SIC was borne of the need to standardize and codify the industrial sector in the United States (Pearce 1957). A Technical Committee, established under the direction of the Central Statistical Board, had, by 1940, published two volumes of codes and lists for *Manufacturing Industries* and *Nonmanufacturing Industries*. The use of domain experts was critical to the success of this project (ibid). They brought with them the expertise needed to decide where questionable activities lay within the classification system; this same reliance on domain experts is still common today.

Because the SIC was designed primarily to classify industry, it is an imperfect surrogate for land use classification. In response to the SIC shortcomings in this regard, the Federal Highway Administration and Department of Housing, in 1965, published the Standard Land Use Coding Manual (SLUCM). While based upon the SIC framework, “*the SLUCM coding was to provide an exhaustive set of land uses...and a limited set of attribute data to further define some of the land-use categories*” (Everett and Ngo 1999). The manual, which was not a mandated standard, proved popular throughout the nation for over a decade before it was abandoned for failing to account for the rapidly evolving nature of land use planning (ibid).

In 1999, the SIC codes were deemed outdated and outmoded and were subsequently replaced by the North American Industrial Classification System (NAICS). The NAICS codes are far more detailed than the SIC codes and account for new types of industry that did not exist sixty years ago. The codes are also used beyond the borders of the United States, in both Mexico and Canada (Jeer 1997). The NAICS codes do not, however, correct the shortcomings associated with the SIC regarding suitability for land use planning; the codes are oriented towards industry and are similarly an imperfect system for land use planning purposes. As a result, despite these efforts to advance new classification systems, jurisdictions continue to rely heavily on site-specific systems and the use of domain experts to help classify land uses.

Currently, there are myriad types of classification systems in use across the nation. Some of these may be based around the SIC or SLUCM framework, and others may be entirely new and unique systems. While these many systems allow for greater customization, they also create difficulties in the integration of land use information across geographic areas. Critics of this individualism, and the host of problems this creates, feel another standard for land use classification is needed. In response to this, in 1994 the American Planning Association (APA), with the support of a variety of federal sponsors, spearheaded an effort to develop a new land use classification system that would not only update the SLUCM codes but also go beyond them to address the changed landscape of land use planning. As a result, the APA has created a multi-dimensional system, the Land Based Classification System (LBCS), which codifies land uses in a hierarchical system for each dimension. The five dimensions are activity, function, ownership, site, and structure. While it is not expected that complete information on all five dimensions will be available or even needed in all jurisdictions, the standards provide a database structure and coding system to accommodate each dimension. The multi-dimensional aspect of the LBCS allows for a greater precision of land use information to be captured such as for natural resources, the existing built environment, and ownership/development rights (Jeer 1999). Adoption of the LBCS has been slow; while most planners recognize the need for a unified system, the new standard is complex and conversion from existing systems is time-consuming.

2.2 Land use data in Wisconsin

With seventy-two counties, nine regional planning commissions, and almost 1800 cities, villages, and towns in the state of Wisconsin, the number of classification systems that exist is potentially quite large, presenting serious challenges to the integration, at all levels, of land use information. Typically, classification systems are used within a discreet political jurisdiction. However, some data sets do not fit neatly within a political boundary. A national forest provides a good example where the border could potentially span parts of several counties.

From the data we have collected thus far, there does not seem to be a common ancestor (e.g., SIC or SLUCM) from which systems arose. Instead, because Wisconsin does not have a mandated land use standard to which codes must adhere, classification systems and their associated codes vary in form just as much as locales vary from place to place. This results in systems attuned to local conditions. For example, one jurisdiction could have five separate codes to describe *Agricultural* land uses while another could only have two. It is likely that the first jurisdiction is a more rural and farm-oriented community.

Differences in the level of detail in codes will affect the specificity of queries. For example, there are differences in codes relating to commercial lands for Dane and Racine counties. If a user were to query the Dane County Regional Planning Commission (RPC) data source for all the commercial lands, there

would be twenty-two possible codes that would satisfy the query. The user could either take all codes returned or possibly refine the query to hone in on a specific type of commercial use, such as Financial Institutions (which has its own unique land use code). However, the Southeastern Wisconsin Regional Planning Commission (SEWRPC), which covers Racine County, only has three codes for commercial lands. These codes are very general and not split into the same basic categories as Dane County. The user would not be able to refine the query to find Financial Institutions because there is no unique code to describe that use; it is lumped together with many other types of commercial uses. However, there is a descriptive document that accompanies the SEWRPC code set. A search capability within WLIS could allow the user to scan for words resembling Financial Institutions (such as Bank) to determine the broad category that would contain Financial Institutions (here, *210 Retail Sales & Service—Intensive*). However, all the returned parcels containing the code 210 would be a large super set of what the user wanted. In general, many code sets do not have written descriptions. In fact, refinements of categories may not be recorded anywhere and are only known to a few specialists in the jurisdiction.

A typical query in the effort to promote regional and statewide land use planning consists of a predicate applied over multiple jurisdictions. An example is, *Where are all the row crop fields in Dane, Racine, and Eau Claire Counties?* A query of this kind is relatively straightforward when using one data set but more difficult when posed over a larger geographic area. Table 2 illustrates the heterogeneity of attribute identifiers and codes that would satisfy the query's criteria over these areas.

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

Table 2: Semantic & structural data heterogeneity

The synonyms (*Lucode*, *Tag*, *Lu1*, and *Lu_4_4*) for the attribute identifier for land use in the database schema are more easily resolved than determining whether the code descriptions share common definitions. Unfortunately, often these definitions are not exact matches; each description slightly varies from one another. For example, the 8110 code from the city of Madison makes no distinction between cropland and farm buildings, whereas the Dane County RPC has a separate code for farm buildings. Eau Claire County's AA code includes dairying and other activities in addition to cropland.

Another problem in answering a query concerning land use is knowing which data set to use. There may be multiple data sets covering all or parts of a geographic area, arising from overlapping jurisdictions. For example, regional planning commissions may overlap county data, and cities are nested within counties, as seen in Table 2 with Dane County and the city of Madison. At the other extreme, holes may exist in data sets such as Eau Claire County data that excludes the city of Eau Claire. Therefore, a choice of, or a combination of, data sets is needed to cover a geographic area.

More complex queries may involve multiple data sets for different purposes or from different agencies covering the same jurisdiction, each containing a specific type of land use information, such as tax assessment, zoning, transportation, or natural resources. Dane County, for example, includes several authorities that have a land use classification system. At the county level, Dane has two separate institutions that have created their own individual land use classification systems based on their respective needs. The Dane County RPC is oriented toward planning, while the Dane County Land Information

Office (LIO) supports property taxation. Although both use land use data, their codes reflect these institutional differences. The city of Madison, which is part of Dane County, has its own land use classification system that is unique from the RPC. In fact, the city has four distinct classification systems: a 4-digit, two 3-digit, and a 2-digit classification system. The 4-digit is based on the SLUCM codes but has been altered to suit the needs of the city planners. The other codes are specific to the city and were created out of the need to aggregate and simplify the 4-digit classification system. The differences between the two 3-digit codes are slight: one classifies government buildings as a government land use, the other classifies government buildings in a combined commercial/institutional category of structures.

A task-based approach may be necessary in choosing data sets when multiple data sets cover the same jurisdiction. Consider another example, *Where are all the lands in conservation uses for Racine County?* Potential source data sets have different definitions for conservation. Racine County has more than one jurisdictional authority with land use classification systems: the RPC and a federal agency operating in the area, such as the Natural Resources Conservation Service (NRCS) or the Federal Farm Service Agency (FSA). Conservation, as defined by one of these authorities, may not mean the same thing to the others operating in the same political boundary. For the RPC, there is not a specific code that describes “conservation.” In lieu of an explicit code, they assign separate codes for wetlands, woodlands, unused pastures, and fallow agricultural lands—all of which embody their concept of “conservation.” In contrast to this, the NRCS’s National Resources Inventory (NRI) has specific land use codes for Conservation Reserve Program (CRP) lands, a program available to farmers, run under the FSA, for the purpose of conserving marginal agricultural lands. It is likely that if both data sets were queried for conservation lands, there would be discrepancies between areas satisfying the query criteria. For instance, wetlands were excluded from the NRCS codes. Ideally, WLIS would inform the user of the potential for multiple data sets with dissimilar definitions for a given land use query. Then, the user would either choose the appropriate data set for a specific area of interest or accept the results along with system-supplied information regarding its accuracy. Again, the above example suggests a task or purpose-based approach should be part of a query system to help guide the choice of data sets for a given query.

3. Current Work

The previous section identified many types of problems associated with integrating land use data sets for query purposes. We are incrementally approaching these problems and, as a start, solving the problem of resolving land use codes for a type of query that is posed in selected master terms and applied over multiple jurisdictions, such as, *Where are all the agricultural parcels in a multi-county area?*

As described more thoroughly in (Wiegand *et al.* 2002), we are enhancing the Niagara XML query engine (Naughton *et al.* 2000) to process structurally and semantically heterogeneous data. As a summary, to represent a query that will range over multiple data sets, we designed and implemented an extension to XML-QL (Deutsch *et al.* 1998) called a Domain Space, which is similar in concept to an XML namespace. However, a Domain Space holds the URIs of the data sets covered by the query. The Domain Space formalism is needed to be able to represent the general query expressed in the GUI posed in master terms and ranging over multiple data sets. To process such a query, we made changes to Niagara’s parser and to the control flow of the query engine to generate query re-writes within an execution loop. We are also designing a query interface appropriate for multi-jurisdictional data sets, multiple themes, and the selection of diverse codes within a theme.

Because of the diversity in attribute identifiers and meanings of code values found in land use data, we introduce an authority (master list) of codes. An authority allows the user to easily pose a query in master terms over multiple data sets. We developed an initial authority and correspondences between data set codes. Our system also gives the user the option of choosing actual local codes from each data set, a method that retains the most precision especially for knowledgeable users. We are also working on automatic code resolutions for more precise subcategories than master terms.

The concept of an authority has been proposed as a method for the integration of heterogeneous XML databases over the Web (Calnan and Cruz 2001). In this approach, declaratively-specified mappings from each XML database to an expert-defined authority, allow for all the integrated databases to be queried uniformly. This method successfully deals with XML databases that have Document Type Definitions (DTDs) containing elements that are similarly named to the elements of the DTD of the authority. Also, the DTD of each XML database must be such that a graph query, as defined in (Calnan and Cruz 2001), can be defined between the authority's DTD and the DTD of the data. This approach works well in certain data sets, like those of the 2001 presidential results that we obtained from the web, where the names of the XML elements are common across all different states (e.g., county, municipality) (Cruz *et al.* 2002). When this is not the case, as in the land use example described above in detail, we are working on an ontology-based approach in which we capture definitions, synonyms, and mappings that can be inspected by a user when posing a query. We also intend to allow a user to construct mappings that are different from the default for unique purposes.

4. Summary and Conclusions

Our goal is to provide full DBMS-type querying for distributed land use data. To start our project, we collected representative data from local jurisdictions and regional planning commissions and identified numerous problems involved in querying across these data. This paper elaborates on the integration problems associated with land use data. Problems include the geographic coverage of data and the diversity of coding systems developed for a variety of purposes within and between geographic areas and jurisdictions. Our implementation solution, discussed more fully in (Wiegand *et al.* 2002), focuses on the problems of structural and semantic heterogeneity between data sets. We are adding a component system to the Niagara Web-based XML search and query engines and introduce a Domain Space concept to the XML-QL query language to formalize a query expressed in master terms over a multi-jurisdictional area. Using master terms or an authority offers a good solution to the problem of resolving diverse land use codes and has been advocated by land use professionals. However, we are also working on automated mechanisms to retain additional precision between code sets.

Acknowledgement

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

References

- American Planning Association. 1994. "Toward a Standardized Land-Use Coding Standard." Working Paper, Research Department for the Federal Highway Administration, U.S. Department of Transportation, 30 March 1994. <http://www.planning.org/lbcs/2PUBS/scopingproject/index.html>
- Calnan, Paul W. and Cruz, Isabel F. 2001. "Object Interoperability for Geospatial Applications." *International Semantic Web Working Symposium*, Stanford, CA, pp. 229-243.
- Cruz, Isabel F., Rajendran, Afsheen, and Sunna, William. 2002. "XML Database Integration for Visualizing US Election Results." Demo, *Proceedings of the National Conference on Digital Government Research*, dg.o2002.
- Deutsch, Alin; Fernandez, Mary; Florescu, Daniela; Levy, Alon; and Suciu, Dan. 1998. "XML-QL: A Query Language for XML." <http://www.w3.org/TR/NOTE-xml-ql/>.
- Everett, Jerry and Ngo, Chimai. 1999. "Land-Based Classification Standards—Federal Role." *Proc. APA National Planning Conference*. <http://www.asu.edu/caed/proceedings99/LBCS/EVERETT.HTM>
- Jeer, Sanjay. 1999. "Land-Based Classification Standards—Session Outline." *Proceedings APA National Planning Conference*. <http://www.asu.edu/caed/proceedings99/LBCS/JEER.HTM>

Jeer, Sanjay. 1997. "Land-Based Classification Standards Project." LBCS Discussion Issues, American Planning Association Research Department. Revised 20 February 1997.
<http://c1.planning.org/lbcs/publications/LBCSWorkShop/lbcsissuespaper.pdf>

Naughton, Jeffrey; DeWitt, David; Maier, David; and others. 2000. "The Niagara Internet Query System." <http://www.cs.wisc.edu/niagara/Publications.html>.

Pearce, Esther. 1957. "History of the Standard Industrial Classification." <http://www.census.gov/epcd/www/sichist.htm>

Technical Report. 1999. Wisconsin Land Council Technical Working Group. Wisconsin Land Information System Report.

Wiegand, Nancy; Zhou, Naijun; Patterson, E. Daniel; and Ventura, Stephen. 2002. "A Domain Space Concept for Semantic Integration in a Web Land Information System." Demo, *Proceedings National Conference on Digital Government Research*, dg.o2002.