

Nancy Wiegand
Position Paper for the Specialist Meeting on Spatial Webs

Internet DBMSs and Semantic Integration

The perspective presented here is based on Internet Database Management Systems (DBMSs) and semantic interoperability relevant to querying and locating geospatial data. This work is a combination of geospatial and DBMS technologies. A new vision is presented to query geospatial data distributed across the Internet. That is, DBMS technologies are being advanced to enable full DBMS-type querying over the Web without data being resident in a DBMS on a particular server. This emerging technology can be considered to be a great extension compared to HTML search engines, which provide searching over the entire Web but only allow keyword searches and are restricted to returning URLs and not actual answers. This new technology can also extend the functionality currently found in geospatial clearinghouses/portals, which restrict users' keywords to pre-defined categories and only return URLs of data sources.

Contrary to this, an Internet DBMS allows full query expressions and the return of answers. An example is to find the *names* of land owners whose property is classified as cropland and who have property larger than the average size property. To obtain this information without an Internet DBMS, a user would have to locate the appropriate data set, download it, and put it into a local GIS or DBMS. However, Internet DBMSs will allow the user to easily pose many ad hoc queries and perform analyses that are now cumbersome. Over the long term, the technology has promise to promote the dissemination of information and derived knowledge to society. Niagara [NDM+01] is an example prototype Internet DBMS. It has search and query engines to locate data sources and query over them. It indexes data marked up in XML, and data are queried using an XML query language. For example, XQuery is being developed by the W3C as a standard XML query language.

Current Prototype System

We have been studying the use of Internet DBMS technology for geospatial data and, in particular, its use for a Wisconsin Land Information System. We are also extending our work to apply to other geospatial portals such as Geospatial One-Stop. We developed a prototype system that extends the Niagara architecture to illustrate our approach [WZ04, WZC+03, WZC+02]. For example, to find data sources relevant to a query that covers a geographic area of more than one jurisdiction (e.g., find all cropland in a watershed), we extract minimal metadata from the user through our geospatial interface. We also extract query terms based on an ontology for a theme. Our global ontologies are supersets of values to avoid contentious issues in defining a standard. Our ontology subsystem performs look-ups to rewrite generic *GeoSpace* queries in local terms for each data set. We are working on the most heterogeneous type of data, which is to resolve differences in land use coding systems. To do this, we extended work on schema and semantic integration to the *value* level of attributes.

Semantic interoperability is a challenge, and we experimented with several methods. Our first and, ultimately, most accurate method was to have a domain expert perform mappings from a global ontology to each local ontology. Although we started with a manual approach, this task was greatly enhanced by the use of a tool [CRS+02]. The global to local mappings are stored in look-up files (agreement files) in XML format that are stored locally but indexed centrally. To alleviate the manual decision-making, we also experimented with adjustments to the Naives Bayesian classifier [Zho03] and with Formal Concept Analysis to try to automatically match global to local ontologies [ZW]. We hope to do more work with these methods to fully realize their potential and limitations. Our partner in our NSF Digital Government project is also working on other methods for semi-automatic deductions [CSC04].

Our ongoing work will combine and continue to explore methods that more closely approach an automatic resolution of query terms. Our work can also be moved into the new technologies of RDF and OWL that were not available when we started. However, we found that our semantic expressions cannot be stated fully enough in OWL, and we hope to make contributions there. We are also working on more efficient ways to extend Internet DBMS architectures to accommodate an ontology subsystem with look-up files, including investigating P2P architectures. In addition, we are considering more extensible indexing methods for efficient access to large amounts of semantic look-up data.

Locating Geospatial Data Sources Automatically

In addition to the above work, we have proposed the use of Internet DBMS technology to help solve the problem of *locating* geospatial data [Wie04]. That is, current methods often involve searching using HTML search engines, the effectiveness of which we already noted could be greatly enhanced with DBMS technology. Other methods involve the data publisher registering with one to many geospatial Web sites and the user visiting those sites to find data. However, many sites exist (FGDC, ADL, the Geography Network, state and local sites, etc.) making it difficult to know which site to publish in or visit. Although Geospatial One-Stop may alleviate this problem, it could be some time before all data are available in one place. Also, other types of methods have been proposed for handling the dissemination of geospatial data [e.g., OCC+04].

The proposed solution here is to encourage geospatial data producers to publish their FGDC or other metadata files in XML over the Web along with, but separate from, the source data files. Then, Internet DBMS technology would allow querying of the metadata files to specifically locate a required metadata file, which would contain the URL of the source file. In this manner, source files can be precisely located. Furthermore, because many geospatial applications involve the same types of data sets with the only variant being the particular geographic area or jurisdiction, a Web application could be developed to automatically locate data sources. That is, a Web application or service for land use planning (or emergency response) would contain query templates to locate various types of data sources (land use, wetlands, roads, etc.). Template criteria would range over various metadata fields. The Web service would automatically substitute variable information (e.g., jurisdiction name) into a query template and send it to an Internet query engine, which would process the query and return the URL. Because metadata formats are not all the same, semantic technology is needed to do look-ups between possible metadata attributes and values and rewrite queries until the appropriate data source can be found. Such query rewriting techniques would be similar to our existing prototype geospatial Internet query system.

- [CSC04] Cruz, I.F., Sunna, W., and Chaudhry, A. 2004. "Semi-Automatic Ontology Alignment for Geospatial Data Integration", In M. Egenhofer, C. Freksa, & H. Miller (Eds.) *Proceedings GIScience*, Springer, pp. 51-66.
- [CRS+02] Cruz, I. F., Rajendran, A., Sunna, W., & Wiegand, N. 2002. "Handling Semantic Heterogeneities Using Declarative Agreements", In A. Voisard, & S. Chen (Eds.), *Proceedings of ACM GIS*, pp. 168-174.
- [NDM+01] Naughton, J., DeWitt, D., Maier, D., & others. 2001. "The Niagara Internet Query System", *IEEE Data Engineering Bulletin*, 24(2), 27-33.
- [OCC+04] Onsrud, H, Camara, G., Campbell, J., Chakravarthy, N. 2004. "Public Commons of Geographic Data: Research and Development Challenges", In proceedings GIScience, Egenhofer, M., Freksa, C., & Miller, H. (Eds.), Springer, pp. 223-238.
- [Wie04] Wiegand, N. 2004. "Querying Metadata Over the Web to Locate Geospatial Data", In Proceedings of GIScience 2004 Extended Abstracts and Poster Summaries, October, pp. 224-226.
- [WZ04] Wiegand, N. and Zhou, N. "Ontology-Based Geospatial Web Query System", To appear as a chapter in a book resulting from the Next Generation Geospatial Information (NG2I) Workshop, Boston, MA, October 2003, Taylor & Francis.
- [WZC03] Wiegand, N.; Zhou, N.; and Cruz, I.F. 2003. "A Web Query System for Heterogeneous Geospatial Data", In Proceedings Scientific and Statistical Database Management (SSDBM), July, 2003, pp. 262-265.
- [WZC+02] Wiegand, N.; Zhou, N.; Cruz, I.; and Rajendran, A. 2002. "Querying Heterogeneous Land Use Data Over the Web", GIScience 2002 Abstracts, M. Egenhofer and D. Mark (Eds.), Sept. 2002, pp. 207-210.
- [WZC+03] Wiegand, N.; Zhou, N.; Cruz, I. and Sunna, W. 2003. "Resolving Schema and Value Heterogeneities For XML Web Querying", Semantic Integration Workshop at the International Semantic Web Conference (ISWC), Sanibel Island, Florida, October 20, 2003, Demo paper, pp. 149-153.
- [Zho03] Zhou, N. 2003. "Automatic Ontology Mapping of Categorical Information", *Proceedings of the National Conference on Digital Government Research*, Boston, MA, pp. 401-404.
- [ZW] Zhou, N., & Wiegand, N. 2004. "Formal Concept Analysis for Semantic Integration", Unpublished.