Geographic Data Integration to Support Web GIS Development *

André Rocha Coimbra
Laboratoire d'InfoRmatique en Image et Systèmes d'information (LIRIS)
20, avenue Albert Einstein
69621 VILLEURBANNE CEDEX
andre.rocha-coimbra@liris.cnrs.fr

ABSTRACT

Although research in data integration has become one of the main issues in [6], geographic data integration is still a laborious and complicated problem, even for specialists. Geographic data is of great importance as it describes the surface of the Earth and its interactions with mankind. It has its own specific issues as it can be spatially indexed. Exchanging data in real-time through the internet has changed Geographic Information System (GIS) paradigm from closed desktop applications to distributed ones. This research aims at investigating the problems surrounding geographic data integration in order to aid the development of optimized Web GIS.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Spatial Databases and GIS; H.2.5 [**Heterogeneous Databases**]: Data Translation

General Terms

Design, Management

Keywords

Web GIS, data integration, CASE tool

1. INTRODUCTION

During the past few years, geographic data has become more present in people day-life. Technological advances have made it more accessible through, for example, web maps, handheld routing devices, etc. Geographic data is of great importance as it describes the surface of the Earth and its interactions with mankind. Consequently it can be used in various fields of knowledge such as environmental studies, marketing or urban planning. Geographic data has its own specific issues as it can be spatially indexed. This indexing enables spatial operations such as distance calculation or topological operators [5]. Even though acquiring geographic data is an expensive process, the amount of data is increasing at a high rate [10].

The exchange of data in real-time through the internet has changed Geographic Information System (GIS) paradigm from closed desktop applications to distributed ones [6]. In [13], the author defines Web GIS, applications that use as medium the World Wide Web (WWW). These are usually built in a 3-tier architecture (Fig.1). At the bottom tier there is a data storage system accessible by the application layer, usually, through SQL-like languages. At the top tier, there is the user interface (embedded in the web browser) interacting with the application by HTTP calls.

To contribute to develop these applications, [11] has proposed a CASE tool. This CASE tool offers an interface where the application designer can manipulate several data sources, local or remote, and define application logics and presentation rules. After the design phase, the CASE tool automatically generates the application.

The aim of this work is to propose building an abstract data layer to integrate heterogeneous distributed geographic data. This layer would be part of Web GIS new generation architecture. We seek to support the design, development and maintenance of these applications by providing a transparent access to distributed sources. To do so, we have to support the manipulation of the data during the design phase of the applications, and then optimize the queries they execute.

As for the creation of Web GIS, we can identify two actors: the geomatician and the user. The first one is in charge of designing the applications by using the CASE tool. He is a specialist of the field and of the data. The term *geomatician* is used to refer to professionals that are responsible for the design of Web GIS. They are specialists gather, process, present and deliver geographic information to users [11]. The user is the professional that will work with these Web GIS. It is important to define a high level language to manipulate the data in order to facilitate the design phase in the development of these applications.

This work is divided as follows. In Section 2 we discuss the challenges of geographic data integration and Web GIS. Section 3 presents a motivating example. Section 4 discusses

*This research is supported by Business Geografic and the National Association of Research and Technology (ANRT)

MEDES 2009, October 27-30, 2009, Lyon, France.

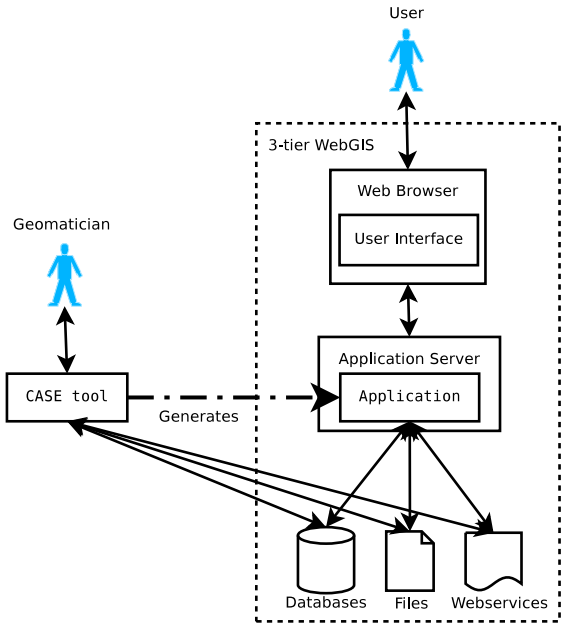

Figure 1: A CASE tool to develop 3-tier Web GIS.

the proposals for a data abstraction layer to support the development and deployment of Web GIS using a CASE tool. In Section 5 we conclude and present further work proposals.

2. GEOGRAPHIC DATA INTEGRATION CHALLENGES

When designing Web GIS, the geomatician needs to manipulate objects from distributed data sources in order to create layers. We use the term *layer* as a view over objects with at least one geographic component and its associated presentation rules. A layer is composed of objects with similar characteristics, for example, roads and bridges may compose a transportation network layer. In order to perform these tasks the geomatician needs tools to manipulate distributed geographic data. In this section we discuss the problems of geographic data integration regarding data types, models and access.

2.1 Data Types

A geographic information system manages mainly two kinds of data: alphanumerical and geographical. GIS power is strongly related to its capacity to access and manipulate data. Geographic data has its own issues as it can be spatially indexed. This indexing enables spatial operations such as distance calculation or topological operators [5]. GIS packages have been designed neither to exchange nor use third party data. For the past decades, geographic information has been locked into closed software packages using specific storage formats and data models [1, 14]. Geographic data can be found in plain files, relational databases, object-oriented databases, webservices (as semi-structured documents), etc.

2.2 Data Models

Data modeling depends on how the designer perceives the real world and how he formalizes real phenomena in a way to best solve a problem. Geographic data is an incomplete abstraction of the real world. One of the main questions regarding geographic data modeling is whether the world is interpreted as a set of discrete objects with defined attributes or as overlapping continuous fields mapped in discrete intervals [5]. There exist a variety of geographic data models and we can argue that there is no best model, each model might be more appropriate to a specific problem.

Geographic data is highly heterogeneous and may differ in data model, schema, data types and interpretation. In [1] the author classifies heterogeneity being one of the three forms:

Syntactic heterogeneity can be seen as different ways to encode information such as different file formats or access protocols.

Structural heterogeneity arises from differences in data models and how they represent the real world. Objects may differ in their data type, relationship types and attributes even if they represent the same real object [2]. Schematic differences appear as a specific type of structural heterogeneity.

Semantic heterogeneity designates differences in the interpretation of data due to its context and the users understanding of the world. For example, a *parcel* may identify a piece of land in a cadastral GIS or a package in a post office application.

Solutions tend to adopt a common data model in order to integrate distributed sources. These approaches provide the user a reference (integrated) schema. One of the problems regarding integration is how to define and manage the reference schema. In order to integrate heterogeneous sources we have to find sets of objects in the sources and reference schemata that represent the same real world phenomena (data mapping). Then, among these sets of objects , solve structural conflicts [3].

2.3 Data Access

Another problem is how to access the data to compose answers when querying the reference schema. During the last two decades, geographic data integration has been bringing the attention of GIS community and approaches followed one of the two: materialized and virtual integration. The former aims at materializing an integrated database through extracting and transforming the data from the sources. It raises problems of synchronizing the sources and the materialized integrated database in order to provide updated data. The latter has to define how to use the sources description and mappings in order to gather data from the sources and compose answers to queries against a virtual integrated schema. Solutions to these tend to one of the following approaches [8]:

Data warehouse solutions follow the first approach which materializes the integrated database. By using wrappers it transforms the data composing a materialized integrated database.

Database Federation is a solution where a virtual integrated schema is presented to the user. The mapping between the virtual integrated schema and the source schemata is based on database views.

Mediation is an approach where a virtual integrated schema is presented to the user through intelligent agents. Queries against this schema are decomposed into sub-queries posed to the sources. The mediator then assembles the answers in order to present the user with the answer to the global query. A mediation system is based on a three level architecture [9]:

- *The application layer* provides the user a single query interface and a query language. The access to the distributed sources is transparent.
- *The mediation layer* is charged of making the link between the global schema and the source schemata. This is usually done by mapping rules. The mediator is also responsible for analyzing the global query and selecting the sources that might contain data to compose an answer. It decomposes the original query into sub-queries and sends them to the sources. It is also the mediator that assembles the sources answers in order to provide the user with a complete answer.
- *The foundation layer* is responsible for the transformations in the flow of data between the source and the mediator. The wrapper converts the data to a common data model in order to provide the mediator an homogeneous view of the sources.

Storage systems containing geographic data differ not only in data types but also in the set of supported spatial operators. For example, sources as shapefiles (ESRI file format) do not support any spatial operator. Furthermore, spatial databases as Oracle Spatial, PostGIS or ArcSDE may differ in the set of supported operators, in operators syntax and implementation. As for this reason, a solution in geographic data integration must integrate not only the data but also an extensible set of supported operators.

2.4 Optimized Web GIS Challenges

Although research in data integration has become one of the main issues in GIS [6], geographic data integration is still a laborious and complicated problem, even for specialists. To better support the development of Web GIS, problems regarding data heterogeneity, data access and spatial operators support must be overcome.

Defining standards can be seen as a step to integration of distributed data sources. Organizations such as International Organization for Standardization (ISO), World Wide Web Consortium (W3C) and Open Geospatial Consortium (OGC) aim at proposing open standards, defining methodologies and specifications in order to promote interoperability in GIS [4]. Standards such as GML, OGCs Web Service (OWS) and the ISO19100 series represent an effort to define how to document, manage and distribute geographic data.

In [12] the authors, inspired by the proposals of ISO and OGC, propose building an architecture for generic GIS. The system would be divided into the presentation, application logics and data tiers, similarly to the 3-tier architecture presented in Section 1.

> "In order to enable reusability and flexibility of the system architecture, the functionality of these tiers must be implemented independently of any particular application " [12]

In the next section we present an application as a scenario in which data integration is needed. We also discuss problems faced when designing and developing the proposed application.

3. A MOTIVATING SCENARIO

Think, for instance, of an application aiming at soil and groundwater pollution control that access two distinct data sources. The first source contains data describing areas where contamination has been measured due to some accident or misuse of herbicides and pesticides. The second source has a description of agricultural properties in the region. In a situation of contamination in a certain region, the purpose of the application is to warn all land owners whose land is within contaminated area. In the following paragraphs we present some of the problems to overcome in order to develop the application described above.

At first, it must be decided what is considered a contaminated area. We can think that several sensors are spread over a region with the aim of measuring the quantity of poisonous substances such as pesticides, herbicides or heavy metals. It might be appropriate to model this phenomenon as a continuous field where the quantity of the substance is described at every point. Although we could also think that this substance can only be considered toxic if found at certain quantities. Consequently a contaminated area could be modeled as a polygon where all points inside it present a poisonous quantity of the substance. This can be seen as a modeling problem of geographic data.

We can consider, for example, that the description of contaminated areas and agricultural properties are modeled as objects in space. To represent heterogeneity in data types, access and the set of supported spatial operators, we could think that the data describing contaminated areas are encoded as an ESRI shapefile and the agricultural properties are stored in a relational database as PostGIS. The data stored in the relational database (PostGIS) can be modeled in several ways. For instance, the information related to rural properties such as geometry and attributes can be stored apart from the information on their related owners, name and address. In this example the sources have different data types, query languages and set of supported spatial operators. In fact, a plain file does not support any operator nor query language.

In a situation of contamination in a certain area the application would execute the following query:

> Which are the names and addresses of all proprietaries whose land are partially or completely located within all contaminated areas?

Answering this query would need joining information concearning the land owners with their respective rural properties upon some common attribute. This operation is usually named geocoding and is used to relate data to geographic coordinates. It is also necessary to select only the rural properties that are intersected by contaminated areas. This is done by applying a spatial operator on the geometry from both objects. In order to execute this operation, the application (if developed as shown in Fig. 1), would need to query each of the sources and then implement the spatial operator using as parameters objects from distinct sources. In this case, the application itself would need to solve the problem of integrating the data. In the context of designing Web GIS with the use of a CASE tool it is interesting to integrate the data at the data access level. To offer a transparent access to the distributed sources would make Web GIS modular. We would have a CASE tool to develop generic applications and the integrated database could be re used by them. In Section 4 we discuss an approach for developing optimized Web GIS and in Section 5 we conclude this article with further research proposals.

4. OUR APPROACH

The geomatician manipulates the data in order to create views over it. These views will represent layers in a map metaphor. Our approach consists in proposing a high level language so that the geomatician can define and manipulate abstract geographic objects in order to compose the layers used by the applications. This language must be flexible enough to support heterogeneous data and the mix of operators. The abstract objects are to be instantiated at execution time. The choice of sources to query depends on their availability which is known by monitoring their metadata. This high level language purpose is to enable the creation and management of an abstract object catalogue.

Object catalogues appear as a solution for publishing and sharing data on the web. They aim at organizing and describing data on remote sources. Through catalogues we are able to model abstract objects describing the concepts of a domain in a flexible way so that these objects can be reused by several applications. Reusing these objects is a good design practice as it prevents from modeling the data used by the application from scratch. These abstract objects allow us to generate abstract query plans. They are a generic description of concepts adapted to some application domain. For example, different applications could use the catalogue that defines the cadastre domain. This would be possible thanks to abstract objects that represent the concepts of the domain such as proprietary and parcel. The catalogue also contains the mapping between the abstract objects used by the applications and the real source data. This would add a supplementary data abstraction layer between application logics and the data management systems. Once the objects in the catalogue are defined and the application is deployed, the application will execute a specific set of queries. Knowing the type of queries we would like to support for an optimized Web GIS, we are able to propose abstract query plans for them. These query plans are to be based on objects from the catalogue and on the knowledge of the sources availability at execution time via metadata monitoring.

Through metadata it is possible to understand the data and how to use it. Metadata is to describe aspects of the data such as: quality, completeness, lineage (methods used in acquiring and modeling), coordinate reference system, legal issues, etc [7]. It is a key element in sharing and reusing data as they help to understand the unique aspects of each dataset that cannot be described by the schemata and data instances [14]. This metadata is accessible and can be collected in execution time for example, through web services interfaces GetCapabilities and DescribeFeatureType. We can identify this metadata as:

- Static (user defined)
  - Abstract Objects Catalogue is a collection of object types that represents the concepts of a domain.
  - Data Mappings between abstract objects in thecatalogue and their materialized representation in the sources.
  - Sources location such as IP addresses or file paths.
- Dynamic (monitored)
  - Sources schemata, the logical representation of the data stored in the sources.
  - Operators Support is a list of operators a source is capable of executing such as relational joins, string manipulations and topological predicates as intersects.
  - Data reference system in which the geographic coordinates were measured.
  - Sources Bandwidth as the time measured for a source to answer a query.
  - Data quantity/size of the sets of objects a source stores.

As the data sources are autonomous and dynamic, they may change their structure from time to time. The geomatician must be supplied with proper tools to manage the evolution of the sources and reference schemata. Additionally, some of the sources may not be accessible at query execution time due to some failure. We aim at developing an optimized data integrator able to choose the sources used in the query processing not only according to the relevance of the data they contain and the operations they support, but also according to aspects such as reliability and performance. In fact, our data integrator should monitor dynamic metadata so as to choose an optimal set of sources to query against.

The metadata acquired through monitoring informs us the current state and evolution of the sources. This information describes, for example, the nature of the sources, usage statistics, availability and bandwidth. It can be used in diagnosing problems at a data source and to optimize query execution. For instance, by having information on the nature of the source it is possible to improve its quality by applying methods such as tunning or index creation.

The goal of this solution is to support a transparent access for the manipulation of heterogeneous distributed data.

We believe that through monitoring dynamic metadata we are able to instantiate query plans during execution time based on sources availability. This would optimize querying autonomous distributed sources. We emphasize that the deployment of the abstraction layer needs to be simple enough to be done by the geomatician. The definition and management of data and mappings cannot add a complexity level higher than the benefits the abstraction layer delivers.

5. CONCLUSION

Geographic data is important so is its dissemination. Web GIS appears as a means to bring geographic data, information and knowledge to a broader audience. This article proposes a generic approach to support the design, development and management of an optimized Web GIS using a CASE tool. It introduces a data abstraction layer in order to integrate geographic heterogeneous distributed data.

This layer should have the following characteristics:

- Use a data model flexible enough to support the diversity and dynamicity of the sources.
- Support a high level language in order to offer a non specialist a transparent access to the data sources.
- This language should enable the geomatician to manipulate the integrated data during the design phase of the applications and to optimize the queries executed by them.
- Support an extensible set of operators by implementing at the mediator level those operators that are not supported by the sources.
- Integrate dynamic metadata in order to facilitate the evolution of the schemata and to optimize queries.
- Monitor abstract query plans based on the availability of the sources.

As mentioned above, GIS power relies on its capacity to access and manipulate data. The data abstraction layer presented would enable the development of new Web GIS with better access to heterogeneous distributed data. These applications would be able to handle changes in the sources structure by monitoring dynamic metadata and to evolve its own structure. A high level language can give more flexibility to the geomatician so that he can construct more specific domain related application schemata composed by abstract objects. Applications would be developed in a more generic way with the aid of a CASE tool so as to decrease the need of Information Technology professional intervention.

6. ACKNOWLEDGMENTS

The author thanks Maryvonne Miquel, Nicolas Lumineau, Olivier Sotin and Frédérique Lévêque for the discussions and comments on previous versions of this work.

7. REFERENCES

[1] Y. Bishr. Overcoming the semantic and other barriers to gis interoperability. *International Journal of Geographical Information Science*, 12:314, 299, 1998.

[2] T. Devogele and C. Parent. On spatial database integration. *International Journal of Geographic Information Systems*, 12(4):335–352, Jan. 1998.

[3] M. Essid, O. Boucelma, F. Colonna, and Y. Lassoued. Query processing in a geographic mediation system. In *Proceedings of the 12th annual ACM international workshop on Geographic information systems*, pages 101–108, Washington DC, USA, 2004. ACM.

[4] F. Fonseca. *Ontology-Driven Geographic Information Systems*. PhD, University of Maine, 2001.

[5] M. F. Goodchild. Geographical information science. In *Int. J. Geographical Information Systems*, volume 6, n. 1, pages 31–45, 1992.

[6] M. F. Goodchild. Geographical information science: fifteen years later. In P. F. Fisher, editor, *Classics from IJGIS: Twenty years of the International Journal of Geographical Information Science and Systems*, pages 199–204. CRC Press, Boca Raton, 2006.

[7] M. F. Goodchild. Epilog : Putting research into pratice. In *Quality Aspects of Spatial Data Mining*, page 345–356. CRC Press, Boca Raton, 2009.

[8] A. Gupta, R. Marciano, I. Zaslavsky, and C. Baru. Integrating GIS and imagery through XML-Based information mediation. In *ISD '99: Selected Papers from the International Workshop on Integrated Spatial Databases, Digital Images and GIS*, pages 211–234. Springer-Verlag, 1999.

[9] A. G. Gupta, I. Zaslavsky, and R. Marciano. Generating query evaluation plans within a spatial mediation framework. 2000.

[10] R. Laurini. Sharing geographic information in distributed databases. In *URISA '94*, 1994.

[11] A. Lbath. *AIGLE : Un Environnement Visuel pour la Conception et la Génération Automatique d'Applications Géomatiques*. PhD thesis, INSA Lyon, november 1997.

[12] M. R. Luaces, N. R. Brisaboa, J. R. Paramá, and J. R. Viqueira. A generic framework for GIS applications. In *Web and Wireless Geographical Information Systems*, pages 94–109. 2005.

[13] Z. R. Peng. An assessment framework for the development of internet GIS. *Environment and Planning B: Planning and Design*, 1999.

[14] G. Percivall, C. Reed, L. Leinenweber, C. Tucker, and T. Cary. *OGC Reference Model*. Open Geospatial Consortium, Nov. 2008.