Developing a Geospatial Search Tool Using a Relational Database Implementation of the FGDC CSDGM Model

Kit Na Goh, Idaho State University, GIS Training and Research Center, 921 S. 8th Ave., Stop 8104, Pocatello, Idaho 83209-8104

Keith Weber GISP, Idaho State University, GIS Training and Research Center, 921 S. 8th Ave., Stop 8104, Pocatello, Idaho 83209-8104 (webekeit@isu.edu)

Daniel P. Ames PhD, PE Department of Geosciences, Idaho State University, Pocatello, Idaho 83209

ABSTRACT

The GIS Training and Research Center (GIS TReC) at Idaho State University provides the public with free access to approximately 26,000 GIS datasets stored within its spatial library. As is the case with many data libraries throughout the world, the GIS TReC data library stores data in file archives (using "zip" compression) to reduce data storage requirements. However, this approach is not conducive to live web-based data queries. Rather, visitors are required to browse a massive directory of folders and sub-folders to find the needed data. In order to facilitate better discovery and delivery of geospatial data, we are developing and deploying a relational database containing geospatial metadata documentation for all datasets within the spatial library and an intelligent web interface to help website visitors more efficiently find required geospatial data. This paper describes the relational database design and the process followed to import existing XML-based metadata documents into the new fully searchable metadata database.

KEYWORDS: metadata, data discovery, public outreach

INTRODUCTION

The Idaho State University (ISU) Geographic Information System Training and Research Center (GIS TReC) is a university-wide facility administered by the Office of Research serving all ISU colleges and departments and the GIS community of southeastern Idaho and is a member of the ESRI-managed Geography Network (geographynetwork.com). The mission of the GIS TReC is to facilitate decision-making through the use and application of state-of-the-art geospatial technologies. The GIS TReC maintains over 26,000 GIS datasets within its spatial library and allows remote users to freely access both raster and vector GIS data via a simple web-based directory structure. The datasets included in this library primarily describe features within the GIS TReC Area of Concern (AOC – figure 1) including for example digital elevation model data, digital orthophoto quads, and vegetation. The library also includes data from outside the AOC such as regional, national, and world data administrative boundaries and large scale environmental data.

Figure 1. Area of Concern

The Current System

The current data delivery mechanism of the GIS TReC spatial library requires clients to browse the spatial library directory of folders and files to find the data they need and then download the data to a local desktop computer to perform GIS mapping and analysis tasks and operations. This approach, though perhaps the most simple spatial library implementation, has several drawbacks including: 1) users must know the name of the required zip file and containing folders; 2) users can not easily search the data to discover new or otherwise useful data; and 3) users must download the entire zip archive for a particular data set before they can open it and browse the contents (including the associated metadata). Figures 2 and 3 illustrate the process followed by a client using the spatial library to find geospatial data. An inherent problem with this approach is that users are not certain if the data they have chosen is correct for their application until they have downloaded, extracted, and previewed it in an application like ArcCatalog.

The existing GIS TReC geospatial library directory structure includes non-archived HTML-based metadata files stored within the same folder as the dataset it describes so that they can be indexed by Google and other search engines. This allows a user unfamiliar with the spatial library to do a Google site search (available from the GIS TReC's website) on these files to gain some idea of the available data. We recognize that the current approach is not optimal for all of these reasons,

[To Parent Directory]

Tuesday, November 28, 2006 12:42 PM	<dir>	AOC basic
Tuesday, November 28, 2006 12:42 PM	<dir>	Idaho
Tuesday, November 28, 2006 12:42 PM	<dir>	Montana
Tuesday, November 28, 2006 12:42 PM	<dir>	Nevada
Tuesday, November 28, 2006 12:42 PM	<dir>	Utah
Tuesday, November 28, 2006 12:42 PM	<dir>	Wyoming

Figure 2: Current spatial data search

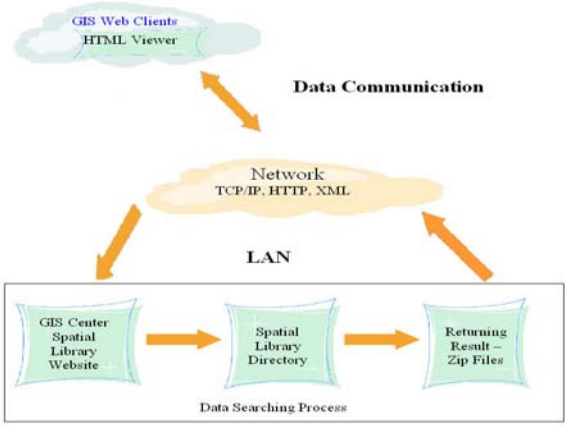

Figure 3. Schematic of how a client currently finds the data

and suspect that other geospatial librarians may have similar problems. Hence, the focus of this research is to develop a geospatial data search tool using a relational database implementation of the FGDC CSDGM model. The goal of this tool is to allow one to easily query all of the metadata describing datasets in the geospatial library through a simple, fast, and powerful search tool.

To determine the types of users of our spatial library, we generated several statistical reports using the software, Flashstats (http://www.maximized.com/). From these reports, it was determined that 26.3% of clients accessed the GIS TReC spatial data library from within the Idaho State University internet domain (isu.edu). An additional 0.43% of the clients were from local state agencies, and 2.4% of clients searched the spatial library from workstations within the GIS TReC itself. These data indicate that about 28.7% of all of our users are students, faculty members, and GIS professionals. Such users are likely to a) know what data they need from our library; and b) have the ability to easily locate it using the current directory browsing access method. However the remaining users (71.3%) who come from outside the university (e.g. from the Geography Network) and the local professional community are surely not as familiar with the GIS servers and especially the organization of the spatial library. Hence we expect that an improved search interface will better serve this large segment of the user community and indeed should expand the user base for this particular regional geospatial data library. It is very likely that other research centers and data archives are used similarly and could also benefit from the development of tools for rapidly searching and accessing geospatial metadata and raw data.

An FGDC Based Metadata Database Approach

To manage and support sophisticated geospatial data searches, we developed a relational database to act as the backbone against which the search interface executes. Our relational database design is intended to improve data discovery and delivery by including all information captured within

geospatial metadata documents. The Federal Geographic Data Committee (FGDC) geospatial metadata format was chosen for the base metadata standard for this project.

The FGDC metadata standard (FGDC 2000) describes GIS datasets using seven major sections and three supporting sections. The seven major sections are: 1) identification, 2) data quality, 3) spatial data organization, 4) spatial reference, 5) entity and attribute, 6) distribution, 7) metadata reference information. The three supporting sections are divided into 1) citation, 2) time period, 3) contact information. In addition to this information, the relational database developed for this project also contains the name and URL of the geospatial dataset, along with descriptive keywords, to better facilitate searching and download. Using a custom search interface developed with Active Server Page (ASP) technology, all geospatial metadata information for the entire library (and potentially outside the library) becomes immediately viewable without the user having to download and extract the dataset.

1.3 Anticipated Benefits

In addition to the improvement in search and query functionality for end-users, an additional benefit of this approach will be the reduction in overall disk space usage and the removal of redundant stand-alone HTML-based metadata files. We anticipate this will also help improve long-term maintenance of the archive (Figure 4).

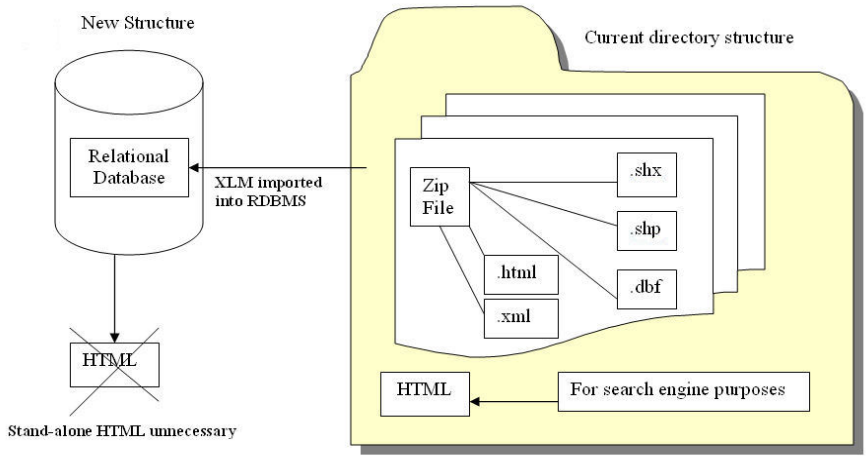

Figure 4. Reduction in redundant documentation

PROJECT DETAILS

Objectives

The objectives of this project include:

1) Improve data discovery: To improve the current search engine of the GIS TReC to help clients more efficiently find required geospatial data and understand the structure of metadata.

2) Develop a robust relational database design using UML CASE visual modeling: design a relational database (Rational Rose software was used to create the UML model) to contain the information within FGDC geospatial metadata documents (extracted from ArcXML documents). The database will include the name and the path to each geospatial dataset, along with descriptive keywords, to better facilitate searching. The database was designed following a top-down approach and normalized through third form normal. The

UML design is available at http://giscenter.isu.edu/research/techpg/nasa_tlcc/template.htm

3) Automate database population: Create an "XML Parser" to parse geospatial metadata (stored within ArcXML documents) and automatically populate the relational database described above.

4) Develop an intelligent web interface to facilitate effective use of the database: Develop an intelligent web interface (that has the capability to query the spatial data with topology keywords, such as, intersect, contain, and adjacent and view the metadata in XML style sheet before downloading) to assist clients searching for data available in the GIS TReC spatial library. This will be accomplished using ASP, HTML, and Javascript.

5) Make the entire process open to the GIS Community: Organize and documents all materials describing database design, XML parser design, and Active Server Page design to assist others in the GIS community.

Database Design Using a UML CASE Class Diagram

At the outset, a database schema or design did not exist that would support geospatial metadata documents. For this reason, the first step was a careful database design normalized to third form normal. Unified Modeling Language (UML) is a standard diagramming language used to design database schemas among other things. One software tool used to create UML models is Rational Rose (Wendy and Michael, 1999). To manage and support sophisticated searches, a relational database with tables that describe GIS datasets based on the seven major sections of the FGDC metadata standard (FGDC 2000) was created using Rational Rose. This database design includes all information captured within geospatial metadata documents, the name and file path to the geospatial dataset, along with descriptive keywords.

The advantage of using a UML model is its ability to transition into a physical database using the UML Class diagram model. Changes can be made to the model and exported using the XML (Extensible Markup Language) file (consisting of UML XMI (XML Metadata Interchange)) which facilitates transfer of UML diagrams to other software applications (such as ArcCatalog). "When a change occurs to the model, Rose can modify the code to incorporate the change" (Wendy and Michael, 1999), this helps to ensure that the model is resilient and flexible.

After exporting the XMI file from Rational Rose, a Microsoft Access relational database (though another RDBMS could just as easily be used) was created using the CASE Schema tool from ArcCatalog. The UML design is available at http://giscenter-sl.isu.edu/umlmodel/index.htm

Development of ArcXML Parser Software

To populate the relational database, all information stored in the HTML-based metadata needed to be imported. Parsing such metadata files can be quite difficult because of the variations of formatting one might encounter in different HTML files. However, when the metadata is stored within the standard ArcXML file format, the process of parsing and importing metadata becomes greatly simplified. An initial "XML File Loader" application was written using Visual Basic 2005 to perform this function and is currently undergoing update and improvement, although it is already functional.

Using the XMLTextReader function in Visual Basic 2005, the XML parser application imports ArcXML formatted metadata to the relational database. The goal of building the XML parser

software is to read an ArcXML metadata document, parse it, extract the values (fields and attributes) and write the data into the appropriate relational database tables.

A Graphical User Interface (GUI) has been designed (Figure 5) for the ArcXML parser program. The GUI includes basic functions (such as Open, Save, Add, Remove Files and Open a folder) (Figure 7) within two areas on the page: a file screen and a preview screen. The file screen displays the name of all XML files contained in the current workspace. The preview screen previews selected XML files along with their keys and values. Keys are objects that are used for selection during data retrievals.

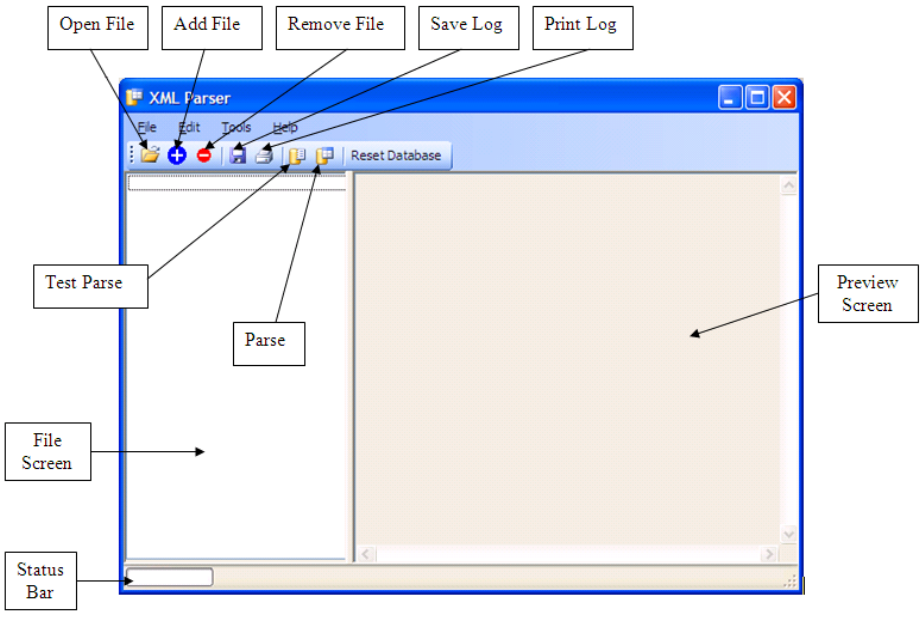

Figure 5. Screenshot of XML Parser software developed for this project.

Development of an Intelligent Web Search Interface

The GIS TReC currently offers two search options to their web clients. The first is a manual search and the second is a simple search powered by Google. The current search engines have limitations and problems which cannot fully facilitate geospatial data discovery and delivery needs, namely a connection between search HTML metadata files and download of the related dataset. Microsoft products or indexing services are also not sufficient as these do not reveal files stored within ZIP files. Because of the size of geospatial data, much of it has been bundled and compressed in ZIP format. This unfortunately, hides much of the real data from clients. In order to locate and retrieve data, an intelligent web interface was required enabling clients to choose or enter search criteria and preview the metadata (all done within the database) without having to download the dataset or try to memorize paths to the required data files.

Search engine tools are becoming common and most search through the meta-databases against meta-descriptions of their geospatial data range.

> "A natural approach is to add advanced features to search engines that allow users to express constraints or preferences in an intuitive manner, resulting in the desired information to be returned among the first results. In fact, search engines have added a variety of such features, often under a special advanced search interface, though mostly

limited to fairly simple conditions on domain, link structure, or last modifications date" (Markowetz et al, 2005).

The proposed method of data retrieval will contain features that allow clients to enter keywords or search phrases and also preferences or constraints such as, building topological relationships (intersect, contain, and adjacent) to locate the correct dataset. This feature will add more capability than the typcial geospatial data search engine (Figure 6).

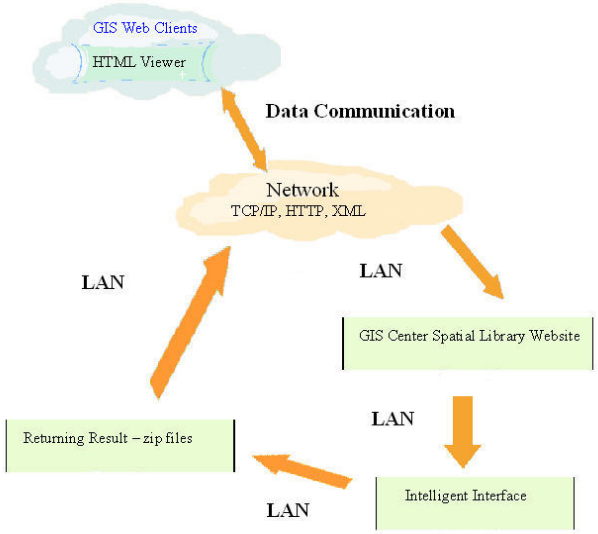

Figure 6. Schematic of new streamlined fashion of finding data

SUMMARY

To facilitate better discovery and delivery of geospatial data for GIS TReC clients, a robust relational database containing FDGC-compliant geospatial metadata was developed. To accomplish this, Rational Rose CASE software was used to create a UML class diagram of the relational database. The model was transitioned (using ArcCatalog's CASE schema tool) to a physical database currently using MS Access (though plans include a transition to IBM DB2 9.1 in the future). Currently, the relational database is being populated using ArcXML metadata documents parsed using XML Parser software developed by the author. Upon completion of this phase of the project, the database will be coupled to an intelligent web interface that searchers for data using geospatial operators.

Initially the metadata within this database will describe geospatial data within the Idaho State University GIS TReC spatial library only. However, once initial implementation is completed, the database will be opened and members of the GIS community will be able to post metadata describing geospatial data from around the world. This distributed design (potentially mirrored on other servers) where only the geospatial metadata is stored within the database (along with a URL path to the data itself) will help facilitate data discovery, sharing, and the development of the GeoWeb.

ACKNOWLEDGEMENTS:

This study was made possible by a grant from the National Aeronautics and Space Administration Goddard Space Flight Center which was made possible through efforts of the Idaho congressional delegation.

LITERATURE CITED

Boggs, M. and W. Boggs, 1999. A Tour of Rose. M*astering UML with Rational Rose (pp.34).* California: SYBEX.

Bradley, J. and A. Millspaugh, 2003. *Programming in Visual Basic .Net.* New York, NY: McGraw-Hill Higher Education.

Date, C. J. 2003. Relational Systems and Others. *An Introduction to Database Systems (8th ed.).* Addison Wesley.

ESRI, 2006. *ESRI Profile of the Content Standard for Digital Geospatial Metadata.* , March 2003. Retrieved 27 September, 2006 from http://www.esri.com/metadata/esriprof80.html

FGDC 2000. Content Standard for Digital Geospatial Metadata Workbook. Federal Geographic Data committee.

Markowetz, A., Y. Y. Chen, S. Torsten, X. Long, and B. Seeger, 2005. *Design and Implementation of a Geographic Search Engine.* Eighth International Workshop on the Web and Databases (WebDb 2005), Baltimore, Maryland, June 16-17, 2005.

Schneider, D. 2005. *An Introduction to Programming Using Visual Basic 2005, sixth edition.* Upper Saddle River, NJ: Pearson Prentice Hall.