Technical Memorandum No. 2

DATA SHARING AND SPATIAL QUERY

Raghavan Srinivasan

Spatial Sciences Laboratory
Texas Agricultural Experiment Station, Texas A&M University

Submitted to
El Paso Water Utilities

Through
Paso del Norte Watershed Council

February 2005

Table of Contents

Data Sharing and Spatial Query

Raghavan Srinivasan

Summary

With multiple data collectors and multiple data sources, efficient data sharing is essential for development of the coordinated database. This technical memorandum explores the options for data transfer and sharing, such as FTP, HTTP, WGET and ArcSDE, ArcIMS with DDE extension, which will eventually be implemented in the Paso del Norte Watershed Council's (PdNWC) Coordinated Water Resources Database and GIS (Geographical Information System) Project. With current PdNWC ArcIMS, a simple one-layer query is already included. For future applications, multi-layer query is desired by regional water resources planners and managers, and recommended for implementation.

Data Transfer and Sharing

There are many ways to advance data transfer and sharing.   Researchers usually refer to FTP, HTTP and WGET for spatial data sharing. The ESRI ArcIMS Data Delivery extension enables GIS data publishing for ArcIMS. It allows ArcIMS user to make their data available in a wide variety of standard spatial formats. The definitions of each follow:

FTP: File Transfer Protocol, or FTP, is a protocol used to upload files from a workstation to a FTP server or download files from a FTP server to a workstation. FTP was one of the first efforts to create a standard means of exchanging files over a TCP/IP network, so the FTP has been around since the 1970's. The key functions of FTP are:

- to promote sharing of files (computer programs and/or data),
- to encourage indirect or implicit (via programs) use of remote computers,
- to shield a user from variations in file storage systems among hosts, and
- to transfer data reliably and efficiently. FTP, though usable directly by a user at a terminal, is designed mainly for use by programs.

FTP enables the user to transfer files between a client host system and a remote server host system. On the client system, a file transfer program provides a user interface to the FTP; on the server, the requests are handled by the FTP daemon, ftpd. FTP control frames are TELNET exchanges and can contain TELNET commands and option negotiation. However, most FTP control frames are simple ASCII text and can be classified as FTP commands or FTP messages.

HTTP: The Hypertext Transfer Protocol (HTTP) has been in use by the World-Wide Web (WWW) global information initiative since 1990. The HTTP protocol is based on a request/response paradigm. The communication generally takes place over a TCP/IP connection on the Internet. The HTTP is an application-level protocol with the lightness and speed necessary for distributed, collaborative, hyper media information systems. It is a generic, stateless, object-oriented protocol which can be used for many tasks, such as name servers and distributed object management systems, through extension of its request methods (commands). A feature of HTTP is the typing and negotiation of data representation, allowing systems to be built independently of the data being transferred.

The innovations that Berners-Lee added to the Internet to create the World Wide Web had two fundamental dimensions: connectivity and interface. Using HTTP, a computer that asked for a file from another computer would know when it received the file, if it was a picture, a movie, or a spoken word. With this feature of HTTP, the Internet began to reflect an important truth - retrieving a file's data is almost useless unless the users know what kind of data it is.

Difference between FTP and HTTP: HTTP, is a protocol used to transfer files from a Web server onto a browser in order to view a Web page that is on the Internet. Unlike FTP, where entire files are transferred from one device to another and copied into memory, HTTP only transfers the contents of a web page into a browser for viewing. FTP is a two-way system as files are transferred back and forth between server and workstation. HTTP is a one-way system as files are transported only from the server onto the workstation's browser. When *http* appears in a URL, it means that the user is connecting to a Web server and not a file server. The files are transferred but not downloaded, therefore not copied into the memory of the receiving device. FTP and HTTP are both essential tools for the Internet, however, FTP's upload ability is still one of the most wanted.

GNU WGET: WGET is a network utility used to retrieve files from the Web using http and ftp, the two most widely used Internet protocols. It works non-interactively, so it will work in the background, after having logged off. The program supports recursive retrieval of web-authoring pages as well as ftp sites. The users can use WGET to make mirrors of archives and home pages or to travel the Web like a WWW robot.

WGET supports a full-featured recursion mechanism, through which the users can retrieve large parts of the web, creating local copies of remote directory hierarchies. Of course, maximum level of recursion and other parameters can be specified. Infinite recursion loops are always avoided by hashing the retrieved data. All of this works for both HTTP and FTP.

The WGET has the following features to make download easier:

- WGET is non-interactive, meaning that it can work in the background, while the user is not logged on.

- WGET can follow links in pages and create local versions of remote web sites, fully recreating the directory structure of the original site.
- File name wildcard matching and recursive mirroring of directories are available when retrieving via FTP. This ability can be extremely useful for automated updating of real time data that may or may not have changed since the previous time a given data site was searched and accessed.
- WGET has been designed for robustness over slow or unstable network connections; if a download fails due to a network problem, it will keep retrying until the whole file has been retrieved.
- WGET supports proxy servers, which can lighten the network load, speed up retrieval and provide access behind firewalls.
- Builtin features offer mechanisms to tune which links the user wishes to follow.
- Most of the features are fully configurable, either through command line options, or via the initialization file.

ArcIMS and DDE: ArcIMS (Internet Mapping Server) is a server-based product that provides a scalable framework for distributing GIS services and data over the Web. The ESRI ArcIMS Data Delivery Extension (DDE) enables GIS data publishing for ArcIMS. It allows ArcIMS users to make their data available in a wide variety of standard spatial formats. End users can translate the data being viewed to a desired spatial output format and projection, and then download the translation results to their desktop. The results include feature attribute data supported by the chosen output format.

The Data Delivery Extension uses Web-based technologies to provide this real-time distribution of spatial data. It performs translation requests using proven Feature Manipulation Engine (FME) technology from Safe Software. The Data Delivery Extension is an extension product to ArcIMS. It is architecturally independent of ArcIMS but can be configured to work with ArcIMS. The integration requires minor modifications to the ArcIMS HTML Viewer.

The DDE consists of four components: Translation Servlet, QServer, FME Server and process Monitor.

The DDE support variable downloadable formats : Adobe Illustrator (Encapsulated PostScript [EPS]) , Autodesk AutoCAD DWG/DXF, Autodesk MapGuide SDL, Design Files (DGN), Bentley/Intergraph), Design Files (V8), EPS, ESRI ArcInfo Coverage, ESRI ArcInfo Export (E00), ESRI ArcInfo Generate, ESRI GML, ESRI Shapefile, Geographix CDF, GML 2, MapInfo MID/MIF, MapInfo TAB, PenMetrics GRD, Raster Image (PNG/GIF),   Scalable Vector Graphics (SVG), VML, VRML.

The DDE also support these data source: ESRI ArcSDE 8.x/Spatial Database Engine (SDE) 3.x, ESRI Personal Geodatabase, ESRI Multiuser Geodatabase, ESRI Shapefile.

The restriction is that the imagery can not be extracted so far, only support the vector data. Using DDE, we can be benefit from: Scalability, Real-Time Translation, Customizable, Data Independence, and Data Security.

Spatial Query

When we work on ArcIMS, we have to work with query functions. Querying a dataset can be in one of three ways:

- A tabular query based on the value of attributes.
- A spatial query based on features selected on a map.
- A combination of a tabular and spatial query.

QUERY, SPATIALQUERY and STOREDQUERY are used in ArcIMS, but it is recommended to use SPATIALQUERY for all query statements with the exception of stored queries.

- QUERY handles only attributes, while SPATIALQUERY handles both attribute and spatial queries.
- QUERY must be used when setting up a stored query.
- Joined tables cannot be used with a QUERY in STOREDQUERY.
- SPATIALQUERY is also used to join DBF files to shapefiles and to relate tables in ArcSDE.
- The main element used around one or more STOREDQUERY expressions in a stored query EXTENSION, for example: extract extension.

Without SPATIALFILTER, SPATIALQUERY works exactly the same as QUERY when querying database attributes. A spatial filter defines the envelope for a spatial query. The envelope can be a rectangle, point, line, polygon, or buffer.

Although ArcIMS only provides querying one-layer at a time, but the user can develop a code algorithm to do multi-layer query, pre-processing of spatial data queries and set up interface for accessing to these spatial data queries. This is not implemented in the current ArcIMS configuration. The users need to develop algorithms and write programs to achieve the same.

In addition to the current implementation of ArcIMS, the one-layer query can be implemented in simple html interface or in a more advanced java client interface. There are add-on packages available to perform the multi-layer drill down operation. That is, if a user clicks on a map in the ArcIMS client interface, the program queries all the layers of interest and returns the information for one location from each layer. However, if an overlay operation between two layers or a conditional query between two or more layers need to be performed, it is not possible in the current implementation of ArcIMS. These need more advanced programming.

Besides the spatial query, ArcSDE is based on standards for managing spatial data in a SQL database that have been defined by both international standards bodies, and industry organizations, a good database management system with ArcSDE will also greatly improve the query efficiencies.

Discussion and Recommendations

It will be the best to automate the transfer of datasets between the various data collectors, such as El Paso Water Utilities, irrigation districts, USGS, IBWC and others. So that once the data is collected, a small script can be developed to automatically upload it to the main central server to host the data in a real time (hourly or within a day). There are various ways to accomplish that. Few of them were listed in this document.

As far as ArcIMS, lots of customization can be made to assure that it is easy to understand its use, including a new feature that allows users to download data. A site can also be developed to distribute metadata along with each data layer. With add-on packages and additional advanced programming, multi-layer queries can be implemented to allow users to perform the multi-layer drill down operation, interlayer operations, or conditional queries between two or more layers.

Glossary of Terms Used

API

Abbreviation of application program interface, a set of routines, protocols, and tools for building software applications. An API is defined at source code level and provides a level of abstraction between the application and the kernel (or other privileged utilities) to ensure the portability of the code. An API can also provide an interface between a high level language and lower level utilities and services which were written without consideration for the calling conventions supported by compiled languages.

ArcIMS

The solution for delivering dynamic maps, GIS data and services via the Web. It provides a highly scalable framework for GIS Web publishing that meets the needs of corporate Intranets and demands of worldwide Internet access. ArcIMS services can be used by a wide range of clients including custom Web applications, the ArcGIS Desktop, and mobile and wireless devices. Using ArcIMS, city and local governments, businesses, and other organizations worldwide publish, discover, and share geospatial information.

ArcSDE

 A server software product used to access massively large multiuser geographic databases stored in relational database management systems (RDBMSs). It is an integrated part of ArcGIS and a core element of any enterprise GIS solution. Its primary role is to act as the GIS gateway to spatial data stored in a RDBMS.

DDE

The Data Delivery extension is an extension product to ArcIMS. It is architecturally independent of ArcIMS but can be configured to work with ArcIMS. The integration

requires minor modifications to the ArcIMS HTML Viewer. The ESRI ArcIMS Data Delivery extension enables geographic information system (GIS) data publishing for ArcIMS. It allows ArcIMS users to make their data available in a wide variety of standard spatial formats.

FME

FME (Feature Manipulation Engine) is an intelligent translator for spatial data and doesn't stop at the simple translation of one GIS or CAD format to the other. Using the FME 'factories' it is possible to define rules to handle very complex models. The FME Suite (Feature Manipulation Engine) consists of three main components: a translator, a transformation editor and a viewer. FME Suite also includes developer tools and coordinate conversion support.

FTP

File Transfer Protocol, or FTP, is a protocol used to upload files from a workstation to a FTP server or download files from a FTP server to a workstation.

GIS

a system of hardware and software used for storage, retrieval, mapping, and analysis of geographic data. Practitioners also regard the total GIS as including the operating personnel and the data that go into the system. Spatial features are stored in a coordinate system (latitude/longitude, state plane, UTM, etc.), which references a particular place on the earth.

HTTP

Abbreviation of HyperText Transport Protocol, the protocol for moving hypertext files across the Internet. Requires a HTTP clien*t* program on one end, and an HTTP *server* program on the other end. HTTP is the most important protocol used in the World Wide Web (WWW)

JAVASCRIPT

A scripting language developed by Netscape to enable Web authors to design interactive sites. Although it shares many of the features and structures of the full Java language, it was developed independently. JavaScript can interact with HTML source code, enabling Web authors to spice up their sites with dynamic content. JavaScript is endorsed by a number of software companies and is an open language that anyone can use without purchasing a license. It is supported by recent browsers from Netscape and Microsoft, though Internet Explorer supports only a subset, which Microsoft calls Jscript.

HTML

Abbreviation for **H**yper**T**ext **M**arkup **L**anguage, the authoring language used to create documents on the World Wide Web. HTML defines the structure and layout of a Web document by using a variety of tags and attributes. The correct structure for an HTML document starts with <HTML><HEAD<BODY> and ends with </BODY></HTML>. All the information the users would like to include in their Web page fits in between the <BODY> and </BODY> tags.

WGET

WGET is a network utility to retrieve files from the Web using http and ftp, the two most widely used Internet protocols. It works non-interactively, so it will work in the background, after having logged off. The program supports recursive retrieval of web-authoring pages as well as ftp sites.

TCP/IP

Abbreviation for Transmission Control Protocol/Internet Protocol. Two interrelated protocols that are part of the Internet protocol suite. TCP breaks data into packets. IP routes packets. TCP enables two hosts to establish a connection and exchange streams of data. IP specifies the format of packets, also called datagrams, and the addressing scheme. TCP/IP was originally developed by the U.S. Department of Defense.

URL

Abbreviation of **U**niform **R**esource **L**ocator, the global address of documents and other resources on the World Wide Web. The URL contains the protocol of the resource (e.g. http:// or ftp://), the domain name for the resource, and the hierarchical name for the file (address).