



White Paper



DAP – Large volume spatial data discovery and distribution over networks

*Ian N. MacLeod, Roger Amorim and
Nicholas Valleau, Geosoft Inc.*

SUMMARY

The Internet, Intranets and general globalisation of networking technology have produced a dramatic increase in the type and volume of geo-data that are available to geoscientists. The development of useful protocols and underlying technologies for computers to access and share geo-data, both privately within organizations, and globally on the Internet is key to our ability to use this information efficiently.

In this paper, we describe the Data Access Protocol (DAP), which is a suite of server applications that enable geoscientists to find and evaluate data, and automate windowing, reprojection and reformatting the data to suit a specific requirement. DAP technology addresses a variety of network situations including:

1. Simple web-browser based discovery and retrieval of data of interest in a specified format and coordinate system.
2. Support for Open GIS Consortium Web Map Server (WMS) interface to allow any WMS compatible application to retrieve "images" of the data for use as layers in a GIS application.
3. Direct support for DAP-enabled thick clients, such as Oasis montaj, to optimally retrieve data directly for their own use, and transfer data to a hosting DAP environment.

When communicating with DAP-enabled client applications, DAP addresses the movement of data (lossless compression, encryption and streaming) both to and from a data server over a network.

The core DAP protocol effectively abstracts data formats to allow client applications to work in whatever environment is required, and DAP servers to connect to data in whatever native format is in use by a hosting organization. This makes DAP suitable for use in many data storage environments. DAP also includes a number of spatially optimised data stores that can be used to deliver extremely high performance for data extraction and retrieval.

Key words: DAP, Data distribution, Network, Internet

INTRODUCTION

Geoscientists are certainly interested in obtaining whatever information and data may be useful in the pursuit of their project goals. This involves finding relevant data, evaluating its usefulness, and retrieving the data or information to be used in the context of a project. By connecting a computer to the Internet, or an Intranet, the available data becomes any information that can be obtained from information servers connected to the same network. We call the finding, evaluation and retrieval of these data the “Data Experience”. DAP (Data Access Protocol) and related technologies address the challenges this experience presents to the geoscientist and aim to make it as efficient and effective as possible.

The Geosoft Oasis montaj system is an example of a client application that uses DAP. When a computer is connected to a network that also hosts DAP servers, a DAP client like Oasis montaj will use the DAP protocol to communicate with and retrieve data of interest from DAP servers. The DAP protocol enables the rapid search for information of interest, evaluation of that information, then retrieval of that information efficiently, securely, and in context for the project at hand.

However, not all data consumers have access to a DAP-enabled network application, though all will have access to a standard Web browser. DAP supports an Open GIS Consortium (OGC) standards based approach to allow web applications to include a data browsing portal and data extraction mechanism.

DESIGN CHALLENGES

The data experience involves three distinct activities – discovery of what data are available that may be of interest; evaluation of those data to determine their usefulness for the task at hand; and finally exploitation of the data through retrieval and use (McLeod, 2000). When applied to the breadth and volume of geological and geophysical data, each stage presents its own challenges, which revolve around three important issues:

1. Data Volume – geophysical data in particular can be extremely large. Airborne geophysical survey databases can easily exceed 10 gigabytes in size, and geophysical grids and images are often

in the 10s of megabytes in size. Consequently, techniques that optimise the delivery of data are very important.

2. Data Formats – data continue to be created and used in a variety of formats, and accessed through a variety of application programming interfaces (APIs). Formats and APIs are important both to a data server, which may use any of a variety of archiving and data storage architectures, and to the data client that must use data in its own specific environment.
3. Coordinate Systems – spatial data must also be used in the map coordinate system of a project, which is often different from the coordinate system of a data archive.

The next section describes key DAP design components in the context of these issues.

DAP Network Architecture

DAP development began in 1999 with the goal of developing technology to support the discovery, evaluation, and retrieval of data by so-called “thick” clients. In contrast to a simple web browser application (a “thin” client), a thick client is a data processing application that runs on a workstation connected to a TCP/IP network. Geosoft’s Oasis montaj is an example of a thick client for data processing and analysis. GIS systems such as MapInfo and ArcGIS are also thick clients. The most important distinction between thick and thin clients is the CPU and memory capabilities that are available to a thick data processing workstation. More recently we have developed a “thin” client web browser for DAP, which is described in the next section.

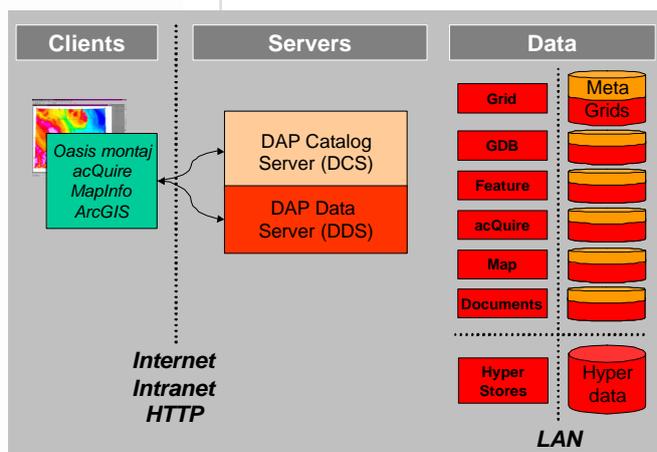


Figure 1. DAP network topology schematic illustrates how a “thick” client workstation accesses data from a DAP server over a network. DAP servers in turn connect to any type of data through plug-in data access components.

Figure 1 shows a simplified schematic of the DAP architecture in a thick client environment. The following sections discuss specific design issues and solutions that became apparent during the research and development of DAP.

HTTP and the TASK Protocol

DAP technologies use Geosoft's TASK protocol to communicate over networks. TASK is a very simple and robust mechanism for client and server computers to communicate and share information securely using the HTTP protocol. HTTP is the basic protocol used by web browsers, and building TASK on top of HTTP allows DAP installations to work within a variety of firewall configurations that limit Internet access to Web browser applications. TASK has been in use in Geosoft's commercial web applications since 1998.

Using TASK/HTTP, only a client can initiate a dialog with a server by sending a TASK that instructs the server to do something. The server responds with zero, one or more TASKs for the client to perform. For example, a client may send a data request TASK, but the server may require user authentication, so it would respond with a TASK to get a user name and password from the client, and a TASK to retransmit the original data request together with the authentication information.

All TASK communication uses via a compressed binary data stream, which provides certain level of security against nominal investigation of network traffic. However, for more robust security, TASKs can be implemented using a full Public Key Infrastructure (PKI) security that conforms to whatever security requirements an organization may have.

Data Streaming

TASK supports true data streaming between a server and a client that requests data. This has significant performance benefits because it allows a server to start compressing and shipping data as it is retrieved from a data source, and the client can start to use the data as it is received. A more conventional approach requires the server to make a second copy of the data to be transmitted, and the receiving client to create yet another local copy which is then converted to the format required in the client.

CPU Load Balance

A more conventional approach to the data delivery problem places all the processing demand on the server. This processing includes data windowing and resampling, reformatting, reprojection, and compression. However, this conventional design does not recognize the potential to harness the client CPU to ease the load on the

server. DAP was designed to deliver data to thick clients, which by definition have a reasonably powerful CPU and resources. DAP takes advantage of this by splitting the work between the server and the client. In a typical DAP installation, a DAP client will perform the work of reprojecting and reformatting data, which tends to be more CPU and resource intensive than the data extraction and windowing task performed by the server.

Server Load Balance

Processing load can also be shared among multiple servers. The task of managing the data catalog and responding to spatial or keyword queries, potentially from many simultaneous clients, demands a rapid response. DAP places this load on the DAP Catalog Server (DCS). When data are requested, the task of windowing, compressing and streaming the data is placed on the DAP Data Server (DDS), which can reside on a different computer. Further, different data sets may be hosted on different servers, but all can share the same catalog server.

Neutral Data Format Across the Network

DAP itself has little knowledge of the data that are actually being transferred across the network. It only knows that data of a certain basic type are being sent. Such data types currently include grids, images, line databases, point data, maps and documents, and all the metadata and attributes associated with these data. The extensible nature of DAP allows new data types to be added as they are needed.

The job of understanding actual data file formats is handled by specific software plug-ins that connect the server to real data. These plug-ins read data and convert it to a neutral format as it is compressed and streamed across the network by the DAP server. On receipt, a DAP client converts the neutral format to an application-specific format for use on the client. In this way, DAP servers have the ability to connect any type of client application to almost any type of data.

In fact one area that we see a lot of interest in DAP is to solve legacy data format issues. Rather than being forced to convert legacy data from existing formats to a suit a new archiving or processing environment, DAP enables an organization to simply connect to the data in its existing format and thus insure access to the data in any future environment.

Coordinate Reprojection

We deal with spatial data, and the coordinate system context of spatial data is a critical, though complex issue. A key feature of DAP is its ability to deliver data in the coordinate system of the client regardless of the coordinate system of the original data. DAP uses the comprehensive coordinate system projection engines that have existed in Oasis montaj for a number of years (conforming to EPSG, 2002), which allow for on-the-fly reprojection of located data, map data, grids and images.

Hyper Stores for High Performance

For many sets of data, the information is accessed infrequently, and performance of the data retrieval system may not be a significant issue. For example, connecting and extracting data from a relational database system (RDBMS) through an SQL interface is very powerful, but can be time consuming. RDBMS systems are ideal for managing a wide variety of data types, and are optimised to deliver small amounts of information on a frequent basis. In a geo-data retrieval environment, we need to be able to deliver very large volumes of data on a much less frequent basis.

To address this performance issue, a DAP data administrator may choose to mirror certain performance sensitive data into “hyper-stores”, that are specifically optimised to deliver optimum windowing and extraction of very large data. A good example of this are the world GLOBE topography data hosted on Geosoft’s public DAP server (www.geosoft.com/dap/map). This is a 3-gigabyte grid of the topography of the world sampled at a 30-second interval. These data are stored in a DAP hyper-grid (a hyper-store for grids and images), and browsing the world at any resolution using the Geosoft web portal appears almost instantaneous.

Disconnect

DAP was primarily designed to deliver data of interest to workstations that need to work with that data. This fundamental focus separates DAP from web-only techniques that require a full-time connection with a data server in order to work with the data. Once data are retrieved from a DAP server, a client computer can disconnect and use the data. This is very important for current and foreseeable exploration work, much of which is carried out on laptop computers away from a central office.

DAP FOR WEB BROWSERS

Although DAP was initially designed to deliver data to DAP-enabled client applications, many organizations that publish and distribute data have a need to expose and deliver the same data through a web browser. Figure 2. shows the web interface to Geosoft’s public DAP server, which contains a variety of data sets that are of interest to the exploration community.

Figure 3. shows a schematic of the DAP Web Map Server that can be added to an existing DAP environment. This includes three additional components:

1. The DAP Map Server is a web page hosting server based on the open source Map Server technology (Lime, 1998-2002). This server in turn supports an OGC/WMS standard interface to render images to its web pages. This server is constructed as a simple frame that can be added to existing web page applications, or added to an existing internal company portal.
2. The DAP WMS Server is an Open GIS Consortium standard compliant WMS server that performs the job of rendering DAP data to an image in response to standard WMS requests (de la Beaujardière, 2002).
3. The DAP Extraction Server supports the work required to extract data from a DAP Data Server and deliver it to a web browser client. This involves reprojection and reformatting to a desired format, then compression into an industry standard ZIP file, and delivering that file to the client.

By adding a DAP Map Server to a DAP installation, the same underlying data management strategy can be used both to deliver real data directly to thick data clients and to publish the data on the web.

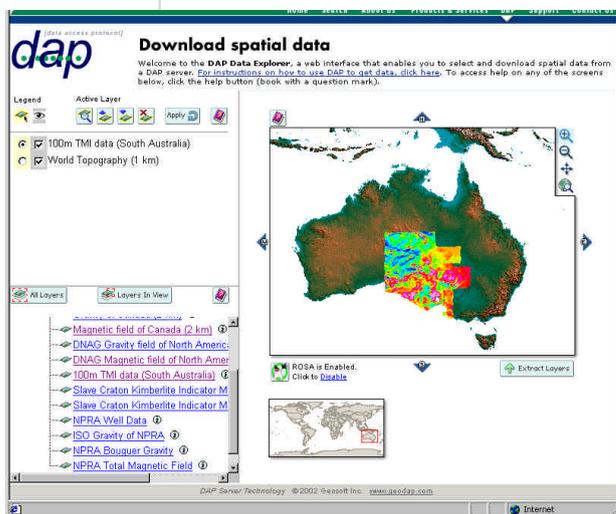


Figure 2. This shows the Geosoft public DAP server accessed through a standard web browser. The page is generated by the DAP Map Server. The [Extract Layers] button will prompt the user to select a layer to extract, the coordinate system and format desired.

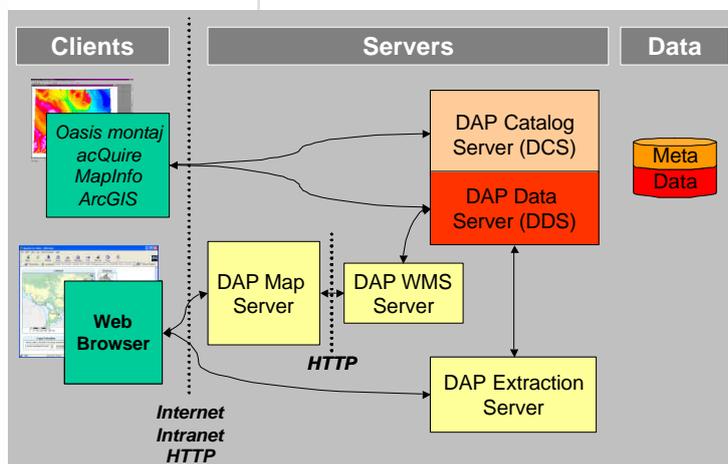


Figure 3. DAP may be configured to support both direct access from a thick client and a web browser using a standards-based OGC/WMS compliant interface. The DAP Map Server supports a browsing frame on a web page, and the DAP Extraction server performs data windowing, reprojection, reformatting, compression and delivery of a ZIP file to a web client.

OTHER ISSUES

DAP Metadata

Metadata that describes data plays a key role in the cataloguing and browsing of data. In DAP, all such metadata are enclosed in a Geosoft META object, which provides an XML compatible way to structure and organize data within a hierarchical data structure. The structure of that data is defined by the needs of the data, not by the software that is using the data. This allows the flexibility to fit DAP to any data model without imposing a metadata standard. Should metadata standards be in place (for example FGDC, XMMML or GeoXML), these standards can be described within the META objects used by that organization. Because META is XML compatible, this also allows organizations to adopt XML based standards as they develop in the future.

Versioning

In any data exchange environment that involves the movement of information between computers, the issue of how to maintain version compatibility is critically important. One cannot install new DAP server versions that have extended capabilities that “orphan” all older client applications. Similarly, newer DAP client applications must be able to communicate with any version of a DAP server. This problem is addressed by adopting Geosoft’s object-oriented versioning technology. This is a mature version-based serialization model that has been in use by Geosoft applications for almost 6 years. Serialization is the conversion of internal data structures to an external form to be placed on the wire, in a file, or directed to some other sequential storage. The Geosoft version technology

allows older applications to work with newer objects by simply ignoring information that is not understood. Similarly, newer applications can always work with older serialized objects. This provides the flexibility to advance the capabilities of both DAP servers and DAP clients independently.

Standards

Standards are particularly important in network environments where a variety of clients must access different sources of data. The Open GIS Consortium (OGC) is taking a lead role in the development of useful standards for working with GIS type data on the Internet. For example, the OGC Web Map Server HTTP standard has already been well established, and this standard was used to support the DAP Map Server technology. However, standards are slow to develop, and it is always uncertain which standards will actually become important in the future.

DAP has been designed to embrace standards, but not be dependent on any specific standard. By supporting a neutral data mechanism across the network, the responsibility to conform to a specific standard or file format is delegated to the client receiving data. Oasis montaj, for example, supports more than 40 different file formats and many standards with the intention of continuing to follow important standards as they develop.

In addition, DAP servers connect to host data using independent data plug-ins that can be constructed to connect to any data format or data standard, even across a network using emerging and future Internet protocols such as the OGC Web Feature Server standard. In summary, DAP has the ability to embrace any standard, but is not dependent on, nor does it impose a specific standard.

CONCLUSIONS

Data Access Protocol (DAP) technology addresses the problem of finding, evaluating and delivering very large volume spatial data and their associated metadata between a network server (Internet or Intranet) and intelligent client applications that wish to use that data. The design allows for efficient movement and use of data and metadata through the use of data streaming, compression and possible encryption of the information.

Data standards are not a part of DAP, but have an important role in connecting DAP technology to existing and developing standards at the server and at the client. DAP also addresses the requirement for evolving the capabilities of the technology through the use of a simple object model and robust versioning technology.

REFERENCES

- McLeod, B., 2000, Geospatial Data Access and Delivery – Open Access to Data, The SDI Cookbook, Chapter 6, 74-91.
- Lime, S. et. al., 1998-2002, WebServer Open Source development environment (<http://mapserver.gis.umn.edu>).
- EPSG, 2002, European Petroleum Survey Group Geodesy Parameters, (<http://www.epsg.org>)
- de la Beaujardière, J., 2002, Web Map Service Implementation Specification, Open GIS Consortium (<http://www.opengis.org/techno/specs/01-068r3.pdf>)
- FGDC, 2000, Content Standard for Digital Geospatial Metadata. (http://www.fgdc.gov/publications/documents/metadata/workbook_0501_bmk.pdf)