DODS: PROVIDING DIRECT ACCESS TO DISTRIBUTED RESEARCH DATA RESOURCES

By George Milkowski

MANY EXCITING ADVANCES in oceanography are occurring because of the data being acquired from newly developed instruments and observing systems. However, the volume and diversity of the data being generated by these systems presents researchers with formidable data management problems. Adding to the complexity of the data management problems is the fact that many oceanographers also want and need access to the data of other researchers or national oceanographic archives. In order for oceanographers' research endeavors to keep pace with rapidly changing technologies it is critical that researchers be provided convenient and easy access to data.

. . . while numerous . . . data systems were being developed there had been no coordination to permit interoperability . . .

This paper describes the design of a distributed data system for oceanographic data that addresses this problem. The system, currently under development, will enable research oceanographers to interactively access and directly import distributed, on-line science data using their own personal research analysis and processing applications. The system is being developed jointly by researchers and staff at the University of Rhode Island Graduate School of Oceanography and the Massachusetts Institute of Technology.

The realization that such a system was needed became evident in 1992 at The Oceanography Society meeting held in Seattle. At that meeting a number of on-line data systems were demonstrated that illustrated the potential of providing access to research data over the Internet. However, no pair of the systems demonstrated could communicate easily. It was clear that while numerous useful oceanographic data systems were being developed there had been no coordination to permit interoperability between them.

In October 1993, the issue of developing an interoperable, distributed system for accessing oceanographic research data was explored in detail at a workshop organized with the help of The Oceanography Society and held at the W. Alton Jones Campus of the University of Rhode Island. The workshop was made possible with funds from NASA and NOAA. To obtain a diverse and representative perspective of different group needs regarding the application and management of oceanographic data, oceanographic researchers, data system developers, and data archive managers were invited.

The workshop had three primary objectives: 1) to develop a vision of a distributed data system for oceanographic research data, 2) to generate a set of requirements for such a system, and 3) to identify system architectures that would meet these requirements. A paper, titled "Report on the First Workshop for the Distributed Oceanographic Data System" (Cornillon, *et al.*, 1993) of the workshop has been published and can be accessed via the Internet[1]. At the end of the workshop, the Distributed Oceanography Data System (DODS) development team was formed. The development team was tasked with the responsibility of generating a detailed system design and building the core components of DODS.

In this article I will summarize the motivation for embarking on the development of DODS and provide an overview of the design and implementation strategy.

Motivation

The motivation for the development of DODS was spurred by the rapid increase in the number of data sets available on the Internet coupled with the lack of coordination between the systems that were developed to access and/or distribute these data. Particularly within the oceanographic community, researchers were making large data sets available on-line (e.g., University of Hawaii's AVHRR ftp site, University of Colorado's SAND-DUNES, Scripps Institution of Oceanography's SSABLE, and University of Rhode Island's XBrowse). Within the federal government, agencies were exploring development of distributed data systems within the context of Global Change

G. Milkowski, Graduate School of Oceanography, University of Rhode Island, Narragansett, RI, 02882-1197 USA.

[1] http://lake.mit.edu/dods.html

(e.g., NASA's Mission to Planet Earth, NOAA's Climate and Global Change) and federal archives were attempting to provide on-line access to archives. However, it appeared from the outside that these federal efforts were being undertaken with little actual developmental coordination.

There was no coordination between researchers who were developing their own systems, and these systems were incompatible. This was true even in the case where the systems were providing the same data sets (e.g., the systems mentioned above providing AVHRR data).

Another issue of concern to oceanographic researchers was the sense that the data resources of individual researchers were being overlooked within the scope of the federal data management programs. This was considered a serious oversight, since researchers produce data as well as use it. In fact, researchers tend to invest a considerable amount of effort "cleaning up", calibrating, and processing raw data for use in their research. The time and knowledge they put into massaging their data increases its scientific value significantly, making it more valuable to the research community as a whole. The federal data systems that were being developed had not identified the need to integrate access to individual data sets within their systems. Thus high quality data from researchers would not be readily available through these systems. Eventually, the raw version of an investigator's data would end up in one of the national data centers where it would be available to the federal data systems; however, the valuable undocumented knowledge that the scientist initially invested in preparing these data for his or her research would be lost to the system and the community.

DODS: An Extension to Existing Interfaces

DODS is not a single stand-alone data management or archive system nor is it a system for creating data base schema and inventories or for managing archival data sets. Rather, DODS is a method for directly accessing data on the Internet. It is a *data access protocol* that provides both a common functional interface to data systems with on-line data and a well defined data model with which to represent data on those systems. It is designed to be integrated with already existing user applications and resource management systems; not to replace them. DODS extends the operational capabilities of existing systems in two important ways: first, it transforms them to a distributed client-server system and second, it provides a system-independent integrated view of their resources.

To implement these extended operational capabilities DODS provides both a defined data delivery model and software tools. The data delivery model provides the necessary infrastructure and semantics for application clients and data servers to interoperate with one another. The tools enable data systems developers to create servers that generate a canonical representation of their data resources, in many cases without modification of the storage form of data sets. In addition, the tools, through the use of customized translators, will allow data users to utilize their data analysis and processing applications as interpreters of the canonical data representation provided by the servers. DODS then is envisioned as a cooperative network of both large and small autonomous systems which interoperate via the DODS data access protocol and supported interfaces.

In order to understand the tangible objectives of DODS, it is useful to pose a research scenario that will help to illustrate the functional concepts that are being implemented within DODS.

A physical oceanographer at the University of Rhode Island (URI) has put together a program of research that involves the tracking of isopycnal floats within the Gulf Stream and the Western North Atlantic. The parameters dealt with are float position (determined from sound source telemetry), time, temperature, and pressure. In addition, CTD, XBT, current meter, and satellite data are often used in the analysis. The researcher utilizes a suite of processing and analysis applications customized specifically for these research data, and uses them on an almost daily basis as new float data come in.

Currently this researcher is involved in a program to study the dynamics of the North Atlantic Current in the region east of the Canadian Maritimes. Not having worked in this area before he or she might be interested in acquiring historical data to gain some long-term perspective on the major dynamic processes occuring in the region and to determine how best to organize and execute a field study with his isopycnal floats. WOCE CMDAC current meter data from Oregon State University, GTSPP temperature and salinity data at NODC, and URI's satellite SST data are available on-line and are data sets that might contain information of interest. Researchers must use their own processing and analysis application for looking at any historical data that are relevant.

This researcher is also interested in exchanging data with other researchers who are collaborating in the North Atlantic Current Study. These data sets are the most relevant to his research and are only available through proprietary arrangements with his colleagues. Again, researchers will want to use their own software applications for conducting analyses.

. . . DODS is a method for directly accessing data on the Internet.

The DODS Implementation Scenario

DODS client libraries are linked to the researcher's processing and analysis applications, in effect making the applications procedure

calls, such as open, read, plot, etc., surrogates for DODS supported operations. DODS server libraries are installed at the systems that provide on-line access to data as well as at the systems of the research collaborators. The researcher at URI interactively analyzes the historical data from the three different on-line data systems by providing a WWW universal reference locator with embedded search constraints as the argument to one of his application procedure calls (e.g., read). The underlying DODS application transparently submits a request from the client application to the appropriate DODS data server. The DODS data server takes advantage of the data system's querying capabilities to locate the data of interest and translates those data into a DODS specified canonical form. The DODS server then transmits the results to the client. The DODS client application translates the canonical representation into the format that the initiating application procedure (e.g., read) expects and then imports the data into the application. The researcher can then continue processing and analysis, using preferred tools for doing research.

The same is true for data that he or she wishes to exchange with collaborators. They access remote data over the Internet and have it imported into their systems in the format that they support. The data system access, query, data translation, and transfer are all transparent to the researcher. Each researcher maintains data in the format that he or she utilizes locally without needing to be concerned about the format requirements of other researchers.

Each researcher maintains data in the format that he or she utilizes locally . . .

Current Data Access Scenario

Before the researcher can use local applications to review the historical data there are a number of preliminary steps that must be gone through. First he must access each data system either through TELNET or WWW and utilize each system's query interface to locate data of interest. Next he or she must transfer the data to the local system, typically utilizing ftp to copy the data files. GTSPP data must also be uncompressed. If the data are in a format not supported by local applications he or she must either reformat the data or recode, recompile, and relink the application. Finally, a GTSPP file contains the global coverage of temperature and salinity data for a 3 month period and is large (10,000 records, 10MB). The researcher will want to extract the data of interest and discard the rest of the file, if only for space considerations since a year's worth of data would require 40MBytes of disk space. Having gone through these steps the data can now be imported into the researcher's applications for analysis.

To exchange data with collaborating researchers, the URI researcher must go through many of the same steps that were required to get the historical data. He or she could make arrangements with colleagues to ensure that the data were provided in a form that local software could read, however, this could lead to multiple copies of the same data set in cases where more than one format were required and would complicate the data management burden on researchers who wish to provide their research data to other researchers.

It is important to reiterate that, in the DODS implementation above, the URI researcher has not needed to modify the analysis application nor to learn how to use a new program or system to access distributed data resources from both large data archives or individuals. The only procedure necessary to make the researcher's application DODS-compliant was to relink it with DODS client library software. Another important feature is that the very application the researcher turns to every day for conducting research processing and analysis has been transformed into an agent that supports access to distributed data resources. And finally, regardless of the form that the data is in at the remote server source, when it is imported to the client application it is translated into the format that the calling procedure expects.

Current Status of DODS Development Effort

The DODS development team has made significant progress toward the goal of developing the system envisioned at the workshop. As a result of the development team's evaluations, along with ideas and comments contributed from the workshop attendees and other interested individuals, many of the ideas put forward at the meeting have evolved during the formative stages of the design process. For example, an open dialogue with individuals outside the development team was particularly helpful to the design of the data model. The following specific issues have been addressed by the development team since the workshop[2]:

- Specified a DODS system architecture based on the client-server model. With the success of systems such as NCSA Mosaic, gopher, and URI's XBrowse a client-server based approach was considered the most versatile and extensible.
- Designed the DODS data delivery architecture. The data delivery architecture specifies the configuration of client and server components, the operations performed by the system, and where those operations are performed.
- Designed a DODS data model, which can support a wide range of earth science data types and structures. The DODS data model is the central component of the DODS data access protocol. It provides the means for translating between different data access mechanisms.

[2] For those interested in learning more about the detailed technical aspects of the DODS development effort, on-line documentation is maintained at http://lake.mit.edu/dods.html.

- Implemented a prototype of DODS using remote procedure calls (RPC) to determine feasibility of remote data access through existing interfaces such as netCDF.
- Identified HTTP as the network transfer protocol that will be used by DODS. HTTP was selected for a number of reasons. It is already a widely used protocol and is effectively managed by developers at NCSA, CERN, and elsewhere. Many researchers already have installed HTTP client and server software on their systems (e.g., NCSA Mosaic and NCSA HTTPD) and are familiar with its use. In addition the Common Gateway Interface (CGI) mechanism of HTTPD is flexible enough to allow sophisticated specialization of the server.
- Identified two candidate system applications for prototype development: netCDF and JGOFS. These two interfaces have different underlying data models and are representative of different types of data systems. JGOFS supports a relational data model, whereas netCDF is designed for gridded data. Implementing these two different applications will present a broader overview of the technical challenges to extending DODS beyond these interfaces.

Currently, the JGOFS and netCDF client libraries and data servers are being coded. A DODS tool kit composed of core components that can be used to build DODS client libraries and data servers for different systems is also under development. The tool kit will provide modules that manage the network communications, implement the DODS data model, manage URLs, implement the data access protocol operators, etc. Prototyping of parts of the system will be taking place in the winter of 1995.

Summary

DODS is being developed to help address the need within the oceanographic community for researchers to access data resources easily over the Internet. DODS developers are creating a system that will allow researchers to take advantage of the rapidly increasing availability of data on-line by implementation of a data access protocol within a client-server based infrastructure. There are two unique aspects of the DODS system design: 1) it utilizes researchers' data analysis and processing applications for accessing distributed data resources by transforming them into client agents within DODS, and 2) it supports a system-independent view of the data resources available through its servers. Rather than trying to replace or build a better system, DODS is purposefully designed to be integrated with other already existing data systems and user applications and to take advantage of and extend those systems' capabilities.

Acknowledgements

I wish to thank James Gallagher for his careful review and critical input to this paper. Funding for the workshop and subsequent DODS development came from NASA grant NAGW-3784.

References

Cornillon, P., G. Flierl, J. Gallagher and G. Milkowski, eds., 1993: *Report on the First Workshop for the Distributed Oceanographic Data System*, 29 September–1 October 1993, W. Alton Jones Campus, University of Rhode Island ❑