Report on the First Workshop for the Distributed Oceanographic Data System

29 September – 1 October 1993

W. Alton Jones Campus

University of Rhode Island

West Greenwich, Rhode Island

Edited by:
Peter Cornillon
Glenn Flierl
James Gallagher
George Milkowski

EXECUTIVE SUMMARY

The Oceanography Society, with funding from NASA and NOAA, is planning a workshop series to define the structure of a client-server based distributed system for access to oceanographic data over the Internet and to develop and test a prototype. The underlying idea of such a system is that individual scientists as well as national archives will become providers of data. Any scientist or archive willing to make their data generally available over the network would install a server on their CPU providing access to their data. Researchers would then be able to obtain any of the data available to this system through a client running on their own CPU.

The series will consist of three workshops.

The first workshop was held at the University of Rhode Island's Alton Jones Campus from 28 September to 1 October 1993, reported herein. The first day was dedicated to questions related to system architecture. The second day focused on communication objects and the third on defining a prototype based on the previous day's discussions. A system programmer hired as part of this effort will be responsible for implementing the prototype.

The second workshop will be held in the summer of 1994. It will focus on progress made in implementing the plans/issues discussed in the first workshop or problems encountered in this. In particular, the prototype outlined in the first meeting will be presented and discussed with some data sets. At this point it should be clear what needs to be done at each data archive site. Following the second workshop, the programmer will go for short periods to the participants' labs to help them get things going or will work with them over the Internet.

The third workshop will be held at the next TOS meeting, late spring 1995. Its purpose will be to present the model to the oceanographic community.

The workshops will be run administratively by The Oceanography Society. Scientific direction will be provided by a steering committee consisting of Glenn Flierl of MIT, Ken MacDonald of NASA (EOSDIS), Jim Holbrook of NOAA's Pacific Marine Environmental Laboratory and Peter Cornillon of the University of Rhode Island.

This is the report of the first workshop, intended to begin the development of a Distributed Oceanographic Data System (DODS). The body of the report is a summary with liberal interpretation of the proceedings of the workshop. Appendix A is a list of requirements that was abstracted from the minutes of the workshop (Appendix B) and from the flip charts developed in each of the breakout groups.

During the meeting, we developed a set of requirements — what the users and data providers would like to see from a DODS. These can be summarized as follows:

- To be successful, scientists must find the system useful and turn to it when they wish to collect data together for their research. This implies that the system must provide access to both archive and PI-held data sets. It must be easy and natural to work with, not only for other persons' data but for their own as well. The system must provide simple ways for scientists to distribute their data and to submit it to the archive centers.

- The system must make it easy to locate the appropriate information. Various procedures for searching, browsing, and refining searches must be possible.

- Interfaces for programming languages must be provided so that data can be incorporated directly into analysis routines. The system must support multiple interfaces, from simple commands to a GUI.

- To be successful, archive centers must perceive the system as a good way for individuals to retrieve information from their archives. In addition, it should encourage and facilitate submission of data.

- The system must be easy to install and extensible (meaning that new servers, new data types, new browsing procedures, new user interfaces, new data filters, ... can be added at any time).

Originally it was intended that the architecture to be used for this system would be defined at the workshop. This was however not practical given the time available, and discussions with regard to system architecture begun at the meeting have continued in the interim. Appendix C contains a brief outline of the three different systems under consideration. The prototyping effort has begun and the communications structure for the system is being defined and implemented.

Each participant was asked to prepare a short summary of their current interests. Appendix E contains these summaries. Appendix D is a listing of the participants.

Everyone involved in the meeting with an Internet address was placed on a distribution list. The purpose of this list was for messages of very broad interest to the group such as when the next meeting would be held. In addition, several participants expressed an interest in being involved in the more detailed discussions related to the development of the DODS that were to follow the meeting. A smaller distribution list was formed for this group. Some of the discussions alluded to in the previous paragraph have taken place via this list. There has also been some discussion in this forum related to data structures.

To receive reports and other updates via the dods-report distribution list, send an email message to `dods-report-request@dcz.gso.uri.edu` with 'subscribe' as the subject or in the body of the message. If you would like to participate in the more detailed discussions between designers and implementors, join the 'dods' list by sending a similar message to
`dods-request@dcz.gso.uri.edu`.

A postscript version of this report can be obtained via anonymous ftp at
`zeno.gso.uri.edu` in `/pub/workshop1/`.

Contents

1 Introduction

The Oceanography Society, with funding from NASA and NOAA, organized a workshop to explore issues associated with a distributed data management system for oceanographic researchers. The workshop took place at the W. Alton Jones Campus of the University of Rhode Island on 29, 30 October and 1 September 1993. The workshop was the first in a series to promote the development of a Distributed Oceanographic Data System (DODS). The long term goal of this series of workshops is to develop a system which will provide direct access to oceanographic research data over the Internet. The first workshop focused on the specification of the system requirements as the initial step in the development of the DODS.

Data providers, systems developers and research scientists from government agencies, academic institutions and private corporations attended the workshop. They provided a comprehensive perspective on the current state of oceanographic data management systems as well as the expertise for productive discussions. The format of the workshop included both small focus groups tasked to address specific issues in detail and plenary sessions which provided a forum for the general discussion of topics and review of presentations made by the focus groups.

This report of the first DODS workshop summarizes the discussions that took place at the meeting and documents the system requirements derived from those discussions. The report is organized into six sections: Workshop Goals, Workshop Organization, Motivation, Vision, General Implementation Issues and Requirements.

2 Workshop Goals

The three primary goals of the workshop were: 1) to develop a vision of a distributed oceanographic data system, 2) to specify the requirements for that system, and 3) to define a system architecture capable of accommodating these requirements. The vision expresses the overall objective of the workshop series. It also provides a model that can be decomposed into its separate parts for the purposes of planning a development strategy. Specifying the system requirements and system architecture are tangible, short term goals that are viewed as the necessary first steps in the development of the DODS.

2.1 System Vision

Developing a vision of the DODS at the workshop was important for a number of reasons, it helped to:

- Define the problem
- Clarify the solution
- Abstract different functions

- Show the synergism of components
- Form the foundation of the workshop discussions

The first two points were critical to establishing a common level of understanding regarding the purpose of the meeting as a whole. This was a difficult task considering the diverse perspectives and backgrounds of the attendees. For example the term 'distributed system' has one meaning to a systems engineer and a very different meaning to a person who is principally a data provider. Systems engineers who work with computer networks generally think of a distributed system as a system where "...the existence of multiple autonomous computers is transparent to the user" (Tanenbaum, 1989). Many data providers use a much broader definition which includes multiple data systems residing on different computers that may be connected via a computer network. The vision helped to clarify the meaning of words and concepts.

As expected, the system vision continuously evolved throughout the course of the workshop. However, considerable emphasis was placed on maintaining a research oceanographer's perspective as the focus during the vision's development. There was a simple reason for this; the purpose of the workshop was to develop a system that will be used! Intentionally focusing on researchers' problems helped ensure that the requirements produced by the workshop would address those issues. This problem-oriented approach provides a basis for developing tools that address a known research problem and are immediately useful for that purpose.

Feedback from system developers and data center managers, within the framework of the research oceanographer's perspective, was used to constrain the system vision. By constantly referencing the system vision, the discussions were prevented from becoming debates of system applications in the abstract. By participating in the development of the vision, meeting attendees were encouraged to explore new and innovative solutions to scientific data management problems. Defining these solutions lead to descriptions of the requirements necessary to develop a system.

2.2 System Requirements

The second goal of the the workshop was the specification of requirements for a system based on the vision. The requirements for the DODS are derived from the discussions that took place at the workshop and are meant to describe the external behavior of the system. Most were formulated within the focus groups and then presented in the plenary sessions for general discussion and review. This process was very dynamic since there were significant interdependencies between how one focus group's requirement would modify the issues another focus group was addressing. Therefore, the plenary sessions served as a forum to resolve discrepancies between different focus group approaches. Appendix I gives the raw requirements as extracted from the meeting. The raw requirements were used to develop the formal set of requirements which are presented in the Requirements Section. These requirements are the basis for the development of the DODS design plan, to be generated following the workshop.

2.3 System Architecture

The final goal of the workshop was to specify a system architecture that would support the DODS vision and satisfy the requirements. Constraints placed on the architecture are that it must be compatible with current systems that would participate in the DODS and be capable of accommodating future systems (large and small) as based on the system vision. This is obviously an important consideration if the DODS is to be successful.

Discussions related to the system architecture helped to resolve where in the system specific functions might take place. In some instances this provided relevant feedback to the requirements specification process. For example, we considered whether data manipulation functions, such as file decompression, should take place at the remote or local CPU and the trade off in each instance.

3 Workshop Organization

The workshop was two and a half days long. The first two days were devoted to the discussion of oceanographic data systems and their functions. The half day session at the end of the meeting was spent summarizing the previous two days' discussions and synthesizing them into requirements for the development of a distributed oceanographic data system. Each of the first two days focused on a single topic. The topic for Day One was system architecture, and for Day Two it was data objects and communication protocols.

Each of the main topics for the first and second day of the workshop was initially introduced by a plenary session. These introductions framed the topic in the perspective of a research oceanographer. Following the morning plenary sessions, the workshop participants broke up into focus groups to examine the topic from a different and more narrowly focused perspective. Afternoon or evening plenary sessions were used to summarize the results of each of the focus groups to the workshop as a whole.

There were three focus groups which met on both of the first two days of the workshop; data providers, system developers and data users. Each of these groups corresponded loosely to the different roles of participants at the workshop. A fourth focus group met on the first day of the workshop to discuss data objects and protocols in use by existing distributed data systems. Each focus group had a group leader who stimulated discussion and summarized the groups' debates for the plenary sessions.

4 Motivation

The workshop series was motivated by the rapid increase in the number of data sets available on the Internet coupled with a lack of coordination in the systems being developed to access and/or distribute these data.

Recently, many oceanographers with large data sets have been making their data sets available on-line. At the same time federal agencies have begun exploring the development of a distributed data system for global change research and federal archivists have begun investigating on-line access to data in federal archives. These efforts are however being undertaken with little to no coordination. For example, most research oceanographers are potential data providers as well as data users; a point overlooked by the federal agencies in their development of an earth science data system for the 1990's. In addition, because research oceanographers collect, calibrate, process and analyze their own data, they are intimately familiar with its strengths and weaknesses. This means that data sets derived from the raw data by the researcher/collector often include his largely "undocumented" knowledge of the raw data. So, although the raw data may be transferred to federal archives for distribution to the community at large, oceanographers will often prefer to deal with data sets that are acquired from a colleague whom they know and whose judgment about data quality they accept. This is especially true in large interdisciplinary experiments such as SYNOP, JGOFS, or WOCE that have been conceived and undertaken by a group of scientists at a number of different institutions. To make full use of their data, each of the scientists will in general require data from one or more of their colleagues. In order to address this problem, groups of researchers have begun investigating or have already developed data systems that provide for easy distribution of data between colleagues on the same project. The JGOFS data system is an example of this. The difficulty is that these systems are being developed independently and, in particular, accessing data held in one of these systems by another system is difficult to impossible.

Although many research efforts require the acquisition of new data, most also make use of existing data held in the national archives. Access to these data is an absolute necessity. A consortium of federal agencies recognize this and have indicated that they will work toward a system providing seamless access to data held in federal archive, but progress toward the actual development of such a system has been slow at best. The perception from the outside is that each agency is focusing on systems that will meet their needs with little actual interaction between the agencies with regard to these systems and how they will operate together.

Much of the above became evident through a group of demonstrations of on-line data systems assembled at the recent Oceanography Society meeting held in Seattle. Of the eighteen systems presented, ten of which were operational at the time of the meeting, no pair could communicate with ease! It also became evident at the meeting that the rate of development of such systems was going to increase rapidly over the next several years. The potential explosion of available data sets along with the fact that there currently exist few systems capable of delivering the data, point to the present as a window of opportunity with regard to the development of a truly distributed data system: the hardware exists, individuals as well as institutions are willing to make their data widely available and advances in software design make a distributed system practical. This workshop, the development effort following, and the later workshops in the series have been designed to address the remaining required ingredient — cooperation in the design of a system to access these data.

5 Vision

The vision of data access that emerged from the workshop was that of oceanographers interactively moving data from data sets located on remote systems directly into their analysis packages for examination or using local applications to acquire and process these data in programs tailored to their research problems. From the oceanographer's perspective, the distributed data to be accessed has a consistent form and structure making it straightforward to locally manipulate data from different sources. The system is viewed as being integrated with the oceanographer's own data system and applications environment so that commands, operations and applications for accessing, processing and analyzing research data are the same whether accessing locally stored data or data held remotely. In addition, the system is a tool which enables researcher to solve problems by providing access to data both as they experiment with analysis ideas and when they are ready to actually process the data.

It became clear at the workshop that this last point, the research scientist's approach to problem solving was not well understood. Scientific research is generally carried out in a tentative fashion at first and, as the researcher gains a better understanding of the issues and the data involved, in a more assertive mode. In the early stages the scientist repeatedly asks "what if" questions, often searching out additional data to corroborate an observation or to answer a question not originally posed but clearly related to the problem at hand. This approach to problem solving is seen as being fundamental to the design of a system that will efficiently meet the scientist's research needs.

To help render the vision presented in the first paragraph of this section more tangible, we present it in the context of a system in which all data are either held locally or reside on remotely mounted disks. First, however, consider how a scientist accesses and analyzes data held locally on-line; this may help clarify the point raised in the previous paragraph. There are two general models to data access in commercially available and/or public domain analysis packages. We refer to those based on an identify-load-command-manipulate paradigm (e.g., MatLab) as Class I systems, and those based on an identify-command-load-manipulate paradigm (e.g., FERRET) as Class II systems. The major distinctions between the two classes are when and where data that the user wants to access can be sub-sampled. In Class I applications the user defines the data (location and format), the identify step, then directs the application to load the data. The application reads the entire data resource into memory; then the user is free to issue commands and manipulate, sub-sample or display the data. For

Class II applications the user again identifies the data resource location and format, but instead of loading the data immediately, the application provides verification of the resource's existence. When the user issues a command to manipulate and/or display the data, the application retrieves only the data which has been specified. In Class I systems the data are sub-sampled within the application program's memory, for Class II systems the data are sub-sampled prior to being placed in memory. In both cases, the user identifies the data resources of interest and then specifies operations to be performed with the data. Both systems allow the researcher to easily access locally held data and provide a straightforward method for operating on multiple types and multiple formats of data resources.

Now consider the case where one of the data resources resides on a disk that is connected to another, remote system. Assuming that the structure of these data is well defined (and known) and that the data can be ingested by the analysis package, the disk containing the data might be remotely mounted via the network file system (NFS[1]) at which point the data appear as a locally accessible resource to the researcher's analysis application (regardless of class). The user specifies these data resources in the same fashion as was done in the simple case of locally held data and the same range of operations is available.

Next, imagine that all disks with oceanographic data are NFS mounted to the user's system and that the analysis application contains a listing of all data sets, their formats and the NFS mounted disk on which they reside. In such a world, if the researcher was interested in contouring all XBT derived temperatures at 500m and plotting the results on a satellite derived SST field, the analysis application would access all the NFS mounted disks with XBT data, import the appropriate subsets, remove duplicate XBT values from different sources, perform the 2D contouring and plot the results on the SST derived field. Conceptually, all oceanographic data resources are local and available to the investigator's analysis application.

6 General Notes on Implementation of the DODS

The DODS is envisioned as a system with the functionality presented in the previous section. The actual implementation of the system need not however rely on remote mounting of disks containing the data of interest; there are other software implementations that would accomplish the same objectives. The image of remotely mounted disks was used in that it is a simple extension of concepts that we all use in our current approach to data analysis and hence renders an understanding of the vision straightforward.

In choosing an actual implementation strategy, an important underlying principle of the DODS effort is the belief that much of the work of putting the system together

[1]NFS, because it is widely used and well known, is used here to illustrate the concept of a virtual file system. There are other implementations of virtual file systems that may in fact be more appropriate to the case in hand

will also be distributed (as are the data). How compatible specific system implementations, such as the NFS virtual file system example above or an object oriented client-server system, either of which could support the functionality required for the DODS, are with regards to distributed development must be considered in the design of the DODS. The relevant issues are how will DODS evolve within a rapidly changing technological environment and how the responsibility for that evolution is distributed? Virtual file systems or client-server approaches will provide different answers to these problems. For example, in the case of virtual file system packages, system-to-system communication software is maintained and controlled by a third party, whereas a client-server system developed in-house would presume that the communication software is developed by the DODS development team and maintained by the DODS community. This issue brings to light questions concerning software maintainability, extensibility and obsolescence.

A second consideration with regard to the choice of a system relates to the requirement that it handle very large data sets, data volumes too large to acquire over communal networks. In these cases, data would be staged by the resource system to removable media (e.g., tape, CD-ROM or optical disk) that are then made accessible to the user's local system. Sophisticated users, who must process large volumes of data in the assertive phase of their research, will require DODS to provide such capabilities. This means that the system must be capable of dealing not only with random access devices, but serial devices and, more importantly, there must be a way of transferring along with the data a means of accessing it. In a client-server implementation, the server, providing data from a remote system, resides on that system. The server may perform format transformations when the data are requested by a client. If a request for a large volume of data on removable media does not pass through the server, (i.e., if it is a straight file transfer to the media) a format transformation may be required on the user's system. Without knowledge of the structure of the data this is impossible. Furthermore, moving the server to the user's system may be quite difficult. There are several different fashions in which this problem could be addressed; some impact the architecture of the system itself while others impact the supplier or the user.

As indicated above, central to the success of the DODS is cooperation; the hardware, approaches to the software and accessibility of data are all realities. Although oceanographers require data from other disciplines and those in other disciplines may require access to oceanographic data, it was agreed at the meeting that the system would be designed to meet the needs of the oceanographer and would focus on access to oceanographic data. The group felt that a larger audience would jeopardize the success of the effort. This does not mean that non-oceanographers would be denied access to the system, nor that non-oceanographic data would be excluded from it.

Clearly, in developing the DODS different approaches must be investigated within the context of the issues raised above while at the same time satisfying to the extent possible the requirements defined at the meeting and presented in the following section.

7 Requirements

Software requirements describe the external behavior of a system. Ideally, requirements reduce ambiguity in the description of work to be done so that developers can evaluate the completeness of a design and users know what to expect in the finished system.

This section contains the requirements for the DODS as described by the DODS Workshop minutes. These requirements were developed by refining basic questions about the system's existence. The answers to those questions came both from the minutes and from our vision of the system.

The basic question that emerged at the workshop was: *What will it take to make the DODS to succeed?* It quickly became clear that the answer is quite straightforward. The scientist must want to use the system; it must be the first place that he or she will think of going to satisfy a data need. This then gave rise to a host of other questions: why will the scientist want to use the system? why will data archives want to make data accessible to the system? etc. The answer to each of these raises still more questions. This section is organized in terms of these questions beginning with the most general and moving to the requirements. The section is divided into subsections defined by a set of high level questions. For clarity, these are listed below prior to entering into the details of the responses.

Many of the items in this section have a tag that shows which items from the minutes match them. Appendix I contains a numbered list of requirements taken from the minutes (Appendix II). The tag numbers in this section match the item numbers in Appendix I.

7.1 Why will scientists want to use the system?

7.1.1 It provides access to data sets held by other scientists and to data sets held in the archives.

1. The system provides a self-consistent view of data.
2. The system supports browsing and searching the data. R: 21
3. The system provides a straightforward API that lets scientists write FORTRAN subprograms and C functions for use with off-the-shelf analysis packages as well as their own in-house analysis software. R: 23, 29, 51
4. The system provides command line interface tools that can be used from the UNIX shell. R: 51
5. The system will support other types of user interfaces including a GUI. R: 50
6. Data accessed are presented in a consistent, usable form, regardless of their storage form or location. R: 16

To obtain a self-consistent view of data requires:

- All data (remote and local) are accessed using the same commands. R: 12, 13
- The system can resolve synonyms when getting data. R: 6
- The system will support data set version numbering. R: 54
- The data can be selected with a simple query
  - Using boolean, relational, and functional operators.
  - The result of a query is the data it describes. R: 58
- The storage format of the data is hidden from the user.

To support browsing requires:

- Individual datasets may provide browse capabilities in various forms; the user should be able to take advantage of these

To support searching and location of data requires:

- The system provides a system-wide directory function. R: 2, 3, 7
  - The system knows about all of its available resources.
  - The system's data-set database is fully distributed (i.e., it is transparently spread out over some or all of the machines which make up the system). R: 8
  - The system's directory information is maintained dynamically. R: 9
  - The system automatically propagates information about additions, modifications or deletions to its data resources to the rest of the system.
  - The system may support search refinement. R: 5
- The location of the data is hidden — Software that the scientist uses will know how to communicate with other parts of the system — the scientist does not have to know. The scientist never has to think 'this is on another CPU'.
- The system enables querying of its data resources with user defined, multi-parameter searches.
  - The scientist can search 'keyword `relop|binop` value' of information extracted from data sets managed by the system.
  - The system supports 'unknown parameter' searches. R: 10
  - The system is able to resolve keyword synonyms in queries. R: 6
- The scientist can search plain text descriptions of data holdings.
- Co-located searches are fully supported by the system. R: 4

- If more than one data set must be used to satisfy a query, then the system must do that. R: 11
- Once found, the same data set may be accessed may times without repeating the initial search process. R: 24
- If a scientist knows exactly where a data set resides, the system can go directly to that data set without searching. R: 18
- Whenever data are managed by the system, they are automatically accessible to other users of the system. R: 30
- A provider may set limits to the remote access of his or her data sets managed by the system. Such access limits may apply to everyone or to everyone except a group of privileged users. R: 48, 56

The data is usable because:

- The system provides direct electronic access to data. The system uses the Internet to move data to the user. The existence of multiple computers within the system is transparent to the user. R: 28
  - There may be limits on the total quantity of information that can be transmitted over the network in response to any single query. These will be set by the individual provider. R: 25
  - Very large data sets may be transferred by mail; the user will be able to use these when they arrive with the same commands used over the network.
  - The scientist can access parts of a data set using a query.
- When the scientist gets data, it is returned in a self-describing form. By transforming all data, regardless of storage format, into a (canonical) self describing form the system only has to provide a parser for that form to correctly handle all data managed by the system.
- For this reason, it is simpler to combine different data sets and to co-locate data from two or more data sets.
- Once data has been accessed, the system makes it easy to import data into other applications. The operations performed to retrieve data make it straightforward to use those data with analysis packages. The API interface makes it simple to use systems like MatLab, ... and the command line interface makes it simple to pipe a data stream into UNIX or home-made filter programs. R: 50
- Data can be saved in files. The resulting files are self-describing and may accessed using the same API as remote data.
- Data users will be allowed to provide comments regarding individual data resources. R: 47

  - Feedback to data providers on problems with data quality or access.
  - Comments will be attributed to their authors, so the incentive to make meaningful comments will be fairly great.
  - The user comment capability would be enabled at the discretion of the data provider. R: 26

7.1.2 It provides an easy tool for managing their own data.

1. A single 'interface semantics' is used to locate and access data.
2. The system can use already developed data management tools or work with data organized using files and the UNIX file system. R: 50
3. The scientist can choose to ignore any of the features of the system while using any other features. R: 17, 20
4. Analysis tools can be added
   - The system's transformation of data includes format translation and subsampling. R: 55
   - The system includes the capability of adding new transformations
5. All data sets managed by the system are uniquely identified. R: 31
6. Although the system supports many features, a scientist is never *required* to make use of those features — A choice of options is always given. These options allow the scientist to choose a level of participation in the system. R: 26

7.1.3 It provides an easy means for them to make their data available to others.

1. Scientists may choose from self describing data formats — both standard and language-based types. R: 14, 22
2. Other storage formats may be added at any time
3. The system will be able to support providing data in a DBMS.
4. Data providers retain complete control of their data resources

The self-describing formats are often what the scientist is already using; then, files need only be moved to the appropriate directories within the system. Formats to be considered initially include

- netCDF, HDF, GRIB/BUFR, various ASCII tables R: 34
- Data providers will be encouraged to write text describing each data set and to use a self describing data format.

- Choosing to make the data set description file and/or using a self-describing data format will make it possible to use more of the system's features with the data. This is true for both local and remote users.

For data not already stored in a supported format, language tools will be provided so that the scientist can describe the data's structure to the system.

- Some ways of structuring data can be described to the system by creating template files using a text editor. This will *not* require any programming knowledge at all.
- Other ways of structuring data can only be described by writing programs.
  - If data needs unusual access procedures (e.g. to be accessed efficiently) the scientist will have to write a program that implements that procedure. We will make documented source code available to the scientist that will illustrate how we wrote the code that access files. The documentation will be more than just commented source code, it will be both annotated source code and a reference manual. R: 35

Data providers can include general and/or technical documentation on their data resources as an aid to users accessing their resources. Such documentation is provided at the discretion of the data provider for data users. They can choose to limit access to the data to only designated users. R: 26

7.1.4 It provides an easy method for submitting data to the archives.

- Since the national archives will be on the system, the archive center can use it to transfer data and write it into their own storage system. R: 57

7.2 Why will the data archive centers want to use the system?

- It allows them to distribute data with minimum effort and expense.
- It allows them to acquire datasets with little work.

- Data centers still have difficulties fulfilling their mandate — to provide data. R: 19, 30
  - Most data centers currently provide data through hard media (i.e., tapes and CD-ROMS). A number recognize the advantage to the research community of having data on-line and accessible over the Internet.
- Many data sets held in data center archives are high profile items. Providing access to these data sets will jump start the system.

- One of the responsibilities of data archival centers is to acquire data from scientist and programs. The system supports data centers not only providing data but also acquiring data for archival purposes.
- Data centers do not have to make use of all of the system's features when managing a given data set with the system. R: 17
- The system will support different kinds of data resource documentation.
  - The system will support maintenance of access statistics. Use statistics will be internally maintained at the discretion of the data provider. R: 53
  - The usage log for each data set will be available to the same set of users as the associated data set. R: 49

7.3 How will the system be installed?

1. Installing the system is easy
2. Extending the system is straightforward
3. The system is an open system.

The system is only supported on the UNIX operating system. Some features may not be available without a workstation and X11R5. R: 1

- All the programs that make up the system can be copied from an anonymous ftp server. Any third party libraries required to build the software will be available at this site as well. R: 36
- The system will be distributed in two forms: 1) precompiled binaries for popular workstations, and 2) source code. R: 36
  - The (supported) workstations include Sun Sparc: SunOS 4.x; DECstation: Ultrix; DEC alpha: OSF/1; SGI: IRIX; IBM RS6000: AIX R: 37
  - Source code will be available so that other platforms can run the system, but we will only support additional UNIX platforms if there is a significant call for such support and we have the resources. R: 38
- The bulk of the system will be written in ANSI C. Parts may use FORTRAN if standard, widely used, source code exits for common algorithms and if there are significant reasons not to recode it in ANSI C. We may also use a different language for any X11 interface we produce. Whether the system *can* be ported to OS 'Q' depends mostly on how similar 'Q' is to UNIX. Without knowing 'Q' in advance, it is impossible to say if the port can be done easily. Any cross-OS port will almost certainly require a systems programmer with experience.

- The software used to manage data may be more complex to build on unsupported platforms than the software used to access data on supported platforms. R: 38
- Enough system documentation will be provided so that others may develop software that will work with 'stock' parts of the system. This explicitly includes user interface software. R: 39, 50, 51

There will be no central management authority for the system. All nodes on the system appear equal. Responsibility for how the system is used and evolves is shared mutually by all participants in the system.

7.4 Possible design features of the system.

1. The system will not exclude the use of protocol translators located at the data source. R: 40
2. The design will not preclude adding protocol translators for efficiency on the user's software, but those translators must be optional (i.e., they must not be required to access any data). R: 41
3. The system must support filtering the data stream, both at the data and at the user end. R: 42
4. The system must support a base level communication procedure by which all parts of the system may communicate. R: 43
   - This communication procedure must include: Data set documentation (which may be null), Identifiers, attributes, data values/types, and structure. R: 44
   - This does not have to be the most efficient means of communication between parts of the system. R: 46
5. The system's components must support negotiation. Each component must be able to determine from other parts what level of support they provide for the various features of the system. R: 45
6. The design of the system must be both extensible and scalable. R: 52

References

[1] Tanenbaum, Andrew S. *Computer Networks, 2ed.*, Prentice-Hall:1989

A DODS Requirements from the Minutes

The following are requirements for the DODS from the workshop minutes. The list is complete in the sense that everything in the minutes that is clearly a requirement is included here. However, there may be important requirements that are not on this list. See Section 7 of this report for the complete requirements.

1. The system must be accessible from a workstation.
2. The system must make the location of remote data easy.
3. The system must make the location of local data easy.
4. The system must be capable of performing searches and co-location searches.
5. Refined searches are optional.
6. The system must be able to resolve keyword synonyms. This is true for both the object locator and data queries.
7. The system must provide 'meta data' in some way (i.e., location of data through descriptions of each server site's contents).
8. Data location will be supported with a distributed database.
9. The distributed locator database will be automatically updated.
10. The system must be flexible enough to handle unknown parameter searches.
11. If several data sets should be examined to satisfy a given query, the system must facilitate that.
12. The system must make acquisition of remote data easy.
13. The system must make acquisition of local data easy.
14. The system must support a variety of different data types.
15. The system must provide access to field/project data.
16. Data Providers — PI and centers — can provide a range of services.
17. The system will provide access to data held by individual scientists.
18. The system will provide access to data from government archive centers.
19. Not all data sets will be treated equally by the system.
20. The system will be able to browse data.

21. Data will be provided in a small set of standard formats.

22. Data will be directly accessible via API without first saving it to a file.

23. Data must be accessible via the search process or directly without first searching.

24. Servers may respond to some queries with a message rather than data (e.g., "You requested too much data", "Get data by ftp",...).

25. Data servers can provide varying service depending on the data set.

26. The system must support redundant data screening.

27. The system must deliver data electronically when possible.

28. Data acquired should be easily accessible to analysis packages (e.g., MatLab).

29. The system reduce the load of providing data to others.

30. The system must provide a global naming procedure for data sets (reduce name-space pollution).

31. The system is specifically for oceanographic data.

32. The system will allow other types (i.e., non-oceanographic) of data, but will not be deliberately designed with such data in mind.

33. The system must provide data translators for at least the following file formats: GRIB/BUFR, HDF, netCDF.

34. The system will provide support for data-archive resident software (e.g., servers) written by the DODS development team, others working in conjunction with the DODS Development Team, and others all on their own.

35. Software distribution must be easy for users to build — it must not require that they have a suite of other libraries. Instead use binary distribution or source distribution with all of the necessary libraries included.

36. Source code can be non-trivial to build if binary software is provided for Sun Sparc SunOS 4.1.x, DECstation Ultrix, Alpha OSF, IBM RS6000 AIX, SGI IRIX.

37. Server source code and/or binary software can be more complex to build/install than client given the relative complexities of the systems (but not more than 1 Programmer-Day).

38. The system should provide enough support for programmers that additional user interfaces can be constructed.

39. The design will not preclude accessing existing systems via protocol translators located at the server.

40. The design will not preclude *optional* client side protocol translators that improve efficiency. However, such translators *must* be entirely optional.

41. The system must support pre and post filters for the data stream.

42. The system must support a base level communication procedure.

43. The server's responses must include: Data set documentation, identifiers, attributes, data values/types, and structure.

44. The system's servers must support protocol negotiation.

45. The base level communication procedure does not have to be 'efficient'.

46. The system's log file allows users to comment on data sets.

47. The data provider must be able to determine what data objects are available and the order in which those data objects are delivered.

48. Usage log should be centralized (physically or logically).

49. The system must provide easy-to-use features for the novice and sophisticated features for advanced users.

50. The system must support several user interfaces: menus/hypertext and a programming language API.

51. The system design must be both extensible and scalable.

52. PI/providers must be credited for making data sets accessible.

53. The system must provide data set version numbers (along with processing algorithm version/revision control).

54. The system's distributed computing capabilities are limited to format translation and subsampling.

55. The provider must be able to restrict access in a flexible way.

56. The system should support data transmission from servers to national archives.

57. The system should not make unnecessary distinctions between 'data' and 'metadata'.

B Meeting Minutes

Distributed Oceanographic Data System Workshop

September 29 - October 1, 1993
W. Alton Jones Conference Center
University of Rhode Island

Day 1:

Introduction by Dr. Cornillon at 8:50

Visions of "Utopia", a perfect oceanographic data system, which will then be limited by some of the realities.

I need access to various data sets to design an algorithm to estimate mixed layer depths globally from scatterometer winds and AVHRR-derived day-night SST differences.

I will need other data: mooring data with good vertical resolution to help design the algorithm XBT/CTD data to help validate the algorithm

will also need the data to be coincident with scatterometer data and clear AVHRR day and preceding or night fields

have the best geographic coverage possible Such data does now exist at NODC, WHOI, JPL , URI and other sites

We envision a system accessed from our workstation that will allow us to locate and acquire these data sets including those held on our site quickly and easily

provide these data to our favorite analysis package in a format that the package recognizes.

We envision others accessing our data the same way (reducing the load of providing data to others that is currently an impediment to data exchange)

Assumptions:

The system will:

- focus on oceanographic data (Schramm- does this include met data?)
- be designed to serve the researcher
- be distributed (available to any data sites)
- based on a client-server model

- be dynamic and changing daily
- be easy to install (approximately 1 day for systems programmer)
- contain a variety of different data sets and data types
- not support distributed processing (at least not designed with this as a requirement.
- upwardly compatible

Question about acceptable time delay in locating and receiving data.
B. Douglas - It's easier to locate a large data set than a highly delimited small data base, for these are "well supported".
G. Flierl - question about "install times"
E. Dobinson - two other assumptions need to be added:

1) do we assume the user knows what data he/she wants?

 P. Cornillon - we will need a "front end" director to provide a location function

2) second assumption - is it always better to move GB of data around, or do we move the request to the site where this data is stored and do the processing on-site (move the request to the data site rather that the data to the request site)

W. Schramm - if your data are stored at multiple sites, this would be very hard.

D. Glover - Distributed Computing Environment - actual processing location should be transparent, system will optimize. Do we want to consider such as system?
P. Cornillon - We will restrict this morning's discussion to more traditional move request to data.

H. Debaugh - As long as you get your data, you don't care where you get your data from - it's much more complex to move your "program" to an unknown processing environment.

J. Corbin - we should not preclude the ability to do distributed processing but our system should not require it.
P. Cornillon - Lets assume for the near term discussion we do not have distributed processing— The Client/Server architecture allows flexibility in

- The data formats served the oceanographic data systems may be linked to other disciplines data systems in the future
- The user interface other user groups can build their own interface to the system (e.g. school teachers)

The workshop series:

- Objective - **To develop the base for our distributed oceanographic data system**
- Two workshops
  - Workshop #1 (now) - communications
  - Workshop #2 (summer 1994) servers
- Fourth TOS meeting - spring 1995 the rest of the world (systems programmer to implement the system designed in workshop #1 and to help data providers build servers)
- Coordinated by TOS (not a URI of JGOFS effort!)
- Steered by a steering committee
  - Glenn Flierl - MIT
  - Ken McDonald - NASA
  - Jim Holbrook - NOAA
  - Peter Cornillon - URI

Question about use and value of metadata

W. Brown - should capacity to have/use metadata be an assumption?
P. Cornillon - metadata varies so much in descriptions, ranging from what it is to how it was calculated. Issues of what the metadata is are very important.

R. Wilson - it is very important to have a browse capability of the host. This lets you verify the suitability and applicability of the data.
P. Cornillon - data browse will be an important requirement in our system. This will be an internet accessible system.

B. Douglas - question about data a PI is still working on and has not published. What happens when this data loses its identity in the system - since his career depends on proper recognition. How will originators get credit for participating in this system?
P. Cornillon - we will have a "voluntary" system. We want to have people use it because our system will help them. We're seeing a real change in how oceanographers make their data available.
B. Douglas - reality says we must have citations of how much our data is being used, and by whom.

B. Schramm - researchers need access to "operational" Navy and NWS data, yet we have constraints on who in the international community will have free access to this data.
P. Cornillon - we don't want "password" protection...

back to introduction...

This workshop is sponsored by TOS, not URI...

Time line viewgraph of workshop sequencing, with the goal of eventually developing our prototype client:

Timeline Figure

Workshop #1 Communication Objective

- Define the architecture of the System form of the Client/Server System
  - focus on the first day
- Choose the Communications Protocol for Data Objects passed between clients and servers -
  - focus of the second day

E. Dobinson - without rank-ordered system requirements from the scientists, its hard to come up with a desired system architecture.
P. Cornillon - how the different architectures scale will be an issue.
R. Chinman - isn't agenda for "data providers" much wider than just "what data will be available"?
P. Cornillon - We will let breakout groups define their own agendas.

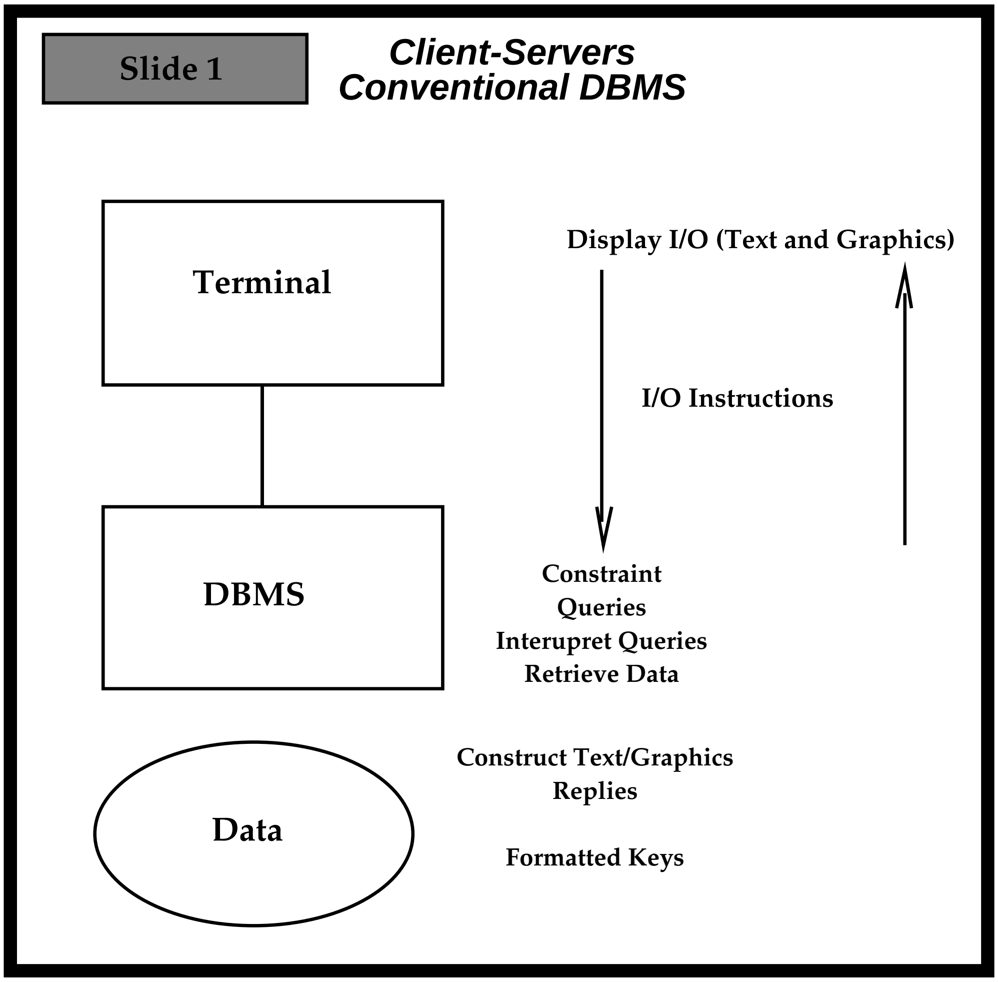
G. Flierl: Discussion of Distributed Systems Architectures

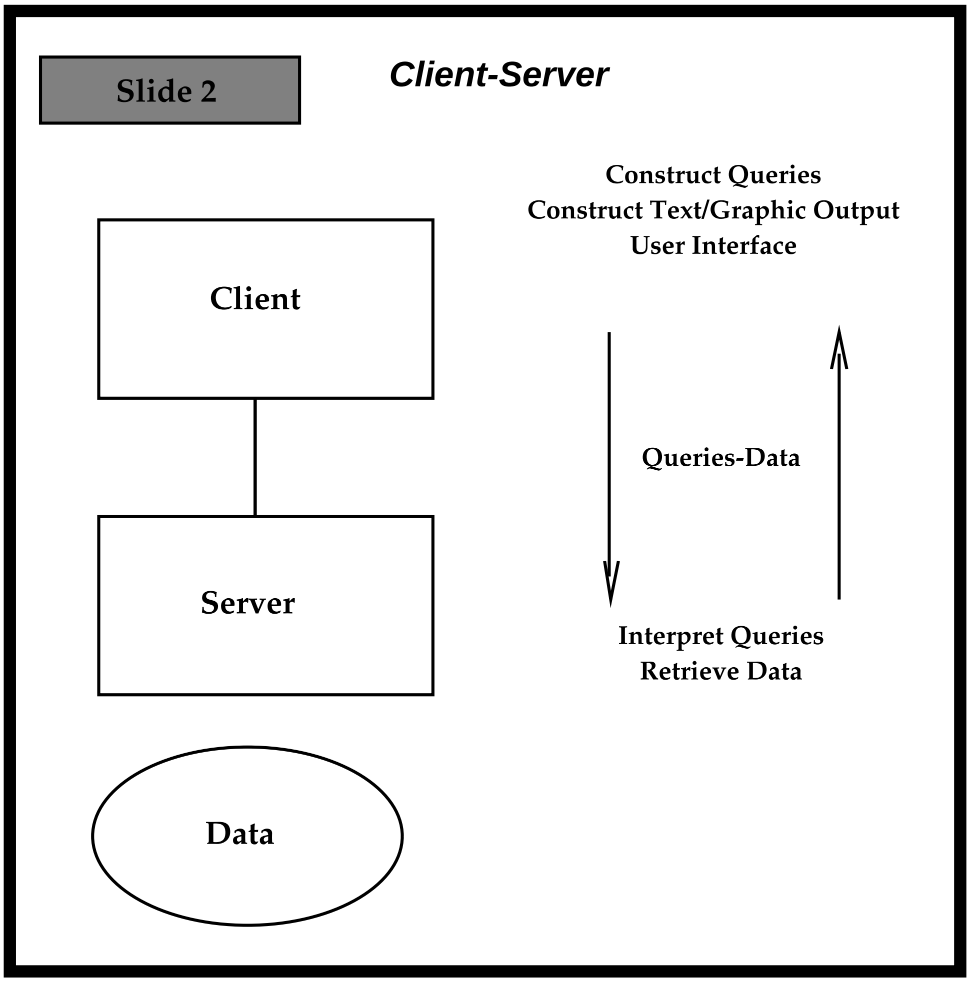
Where do existing systems fit in the context of the models we have discussed? (Slides are included in notes)

Slide 1: Conventional Data Base Management System –

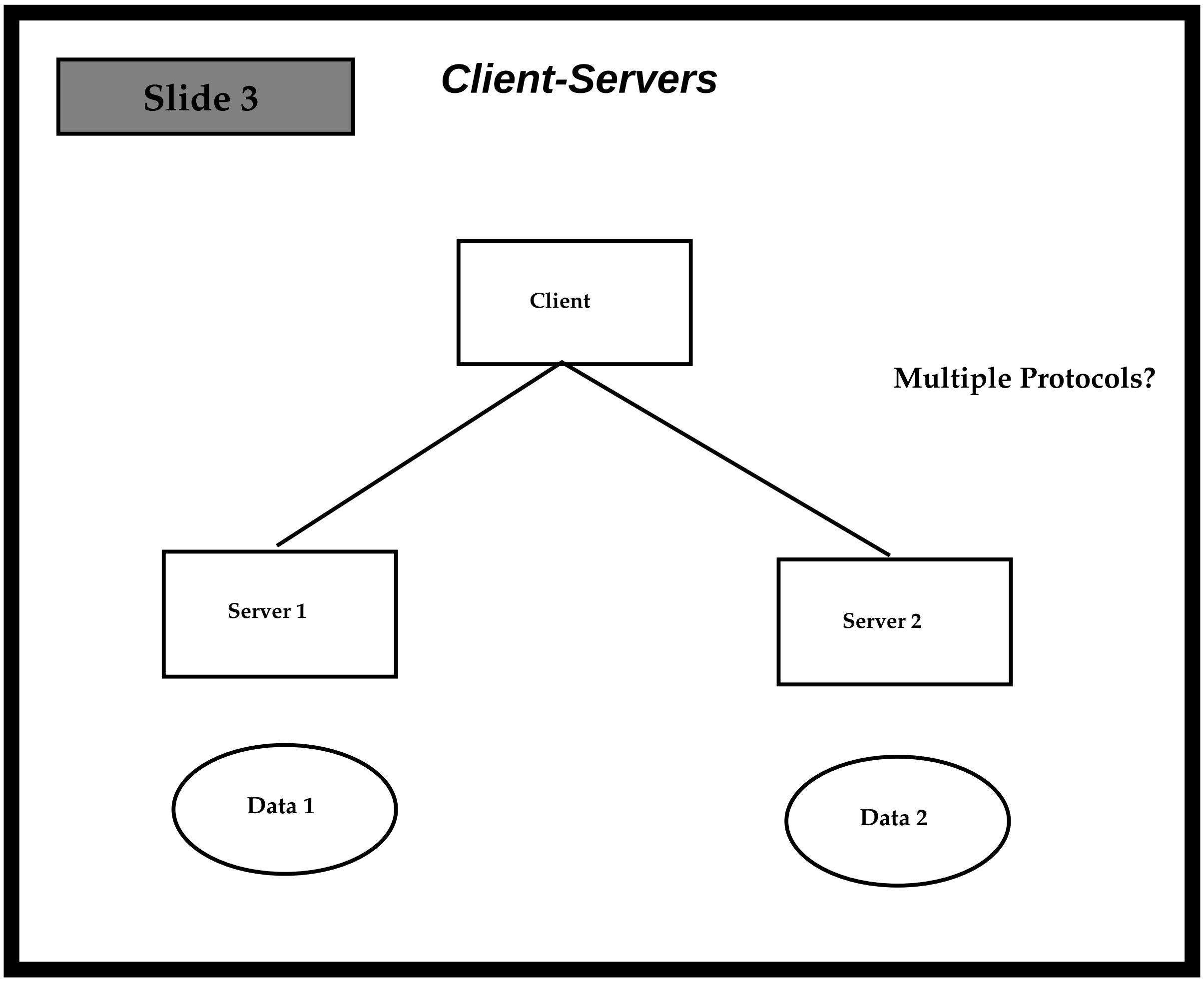
- Data held in a few formats specific to that system.
- Multiple data formats pose problems.

Slide 2: Client-Server model – Moves some of the processing functions to the client. No distinction between "data" and "metadata" .

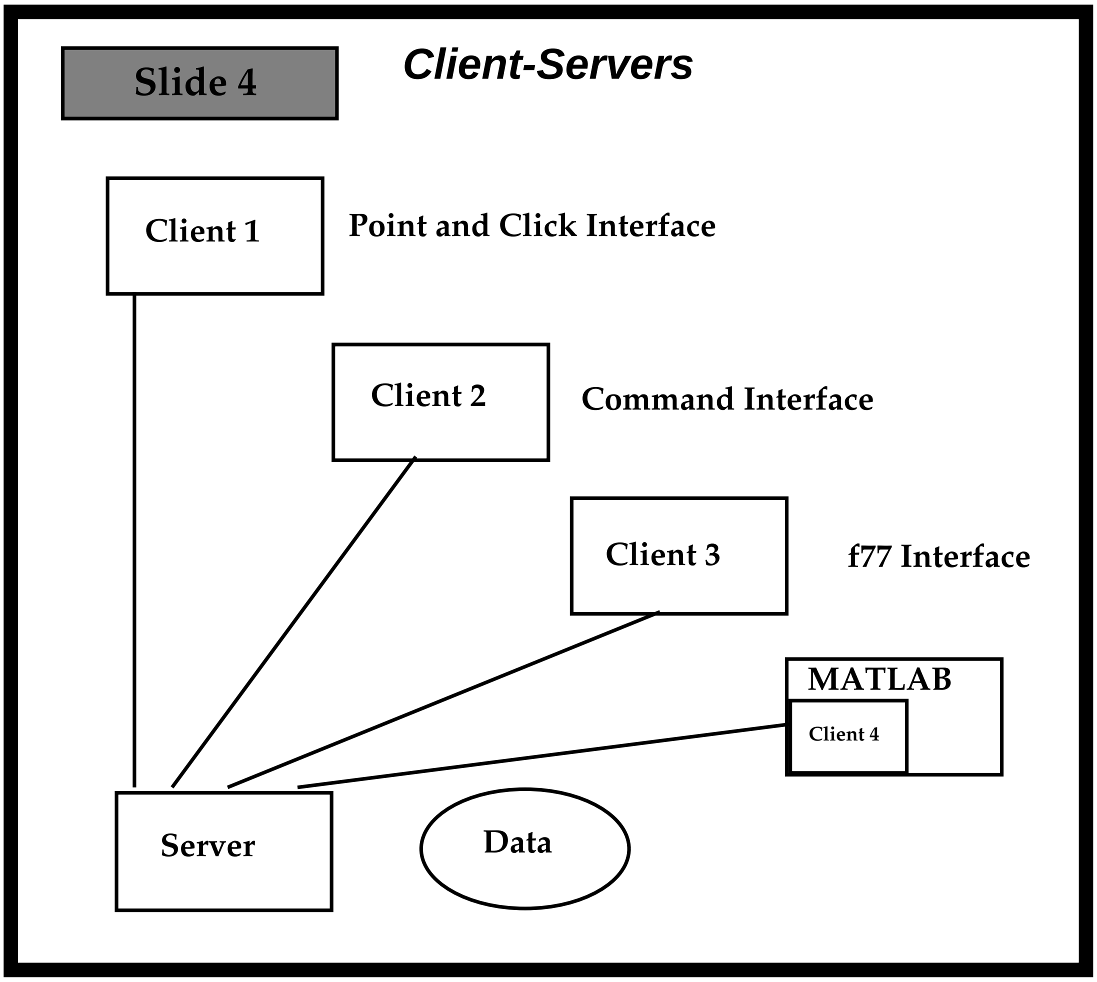

Slide 1
Client-Servers
Conventional DBMS
Terminal
DBMS
Data
Display I/O (Text and Graphics)
I/O Instructions
Constraint
Queries
Interupret Queries
Retrieve Data
Construct Text/Graphics
Replies
Formatted Keys


Slide 2
Client-Server
Construct Queries
Construct Text/Graphic Output
User Interface
Client
Queries-Data
Server
Interpret Queries
Retrieve Data
Data

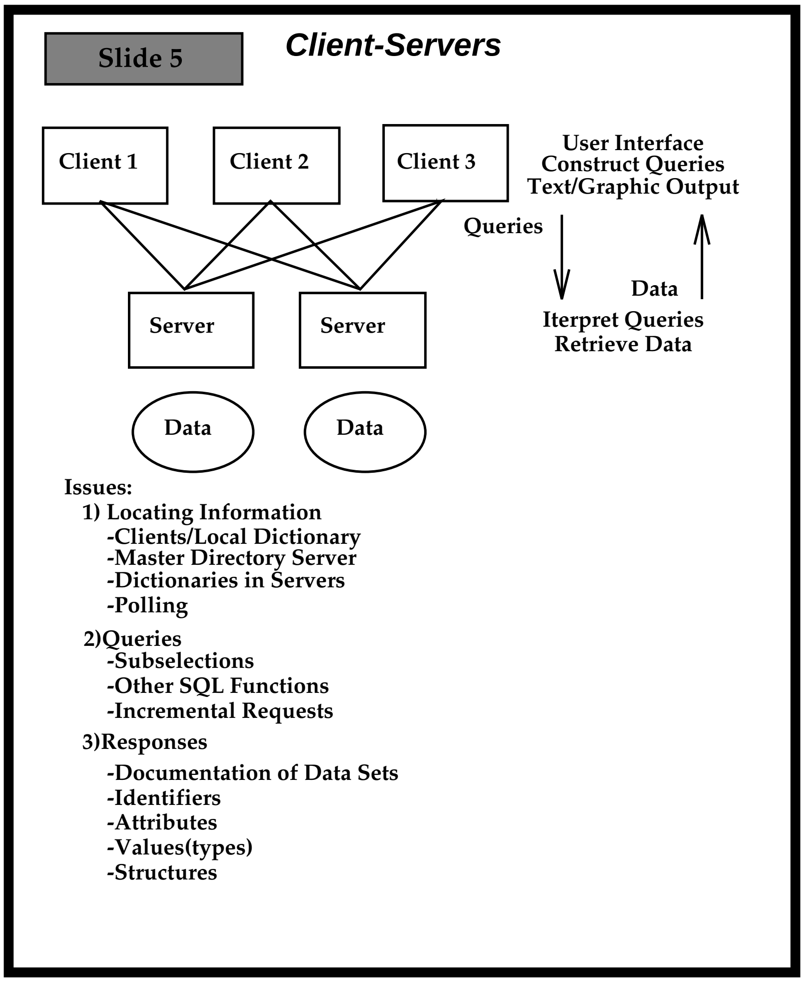
Slide 3: Client-Servers (One client supporting multiple servers, with possible multiple protocols)



Slide 4: Multiple Clients and one server; clients at different levels of complexity (e.g. PI, industry, educational)



Slide 5: Multiple Clients and Multiple Servers. Servers must interpret queries and retrieve data.

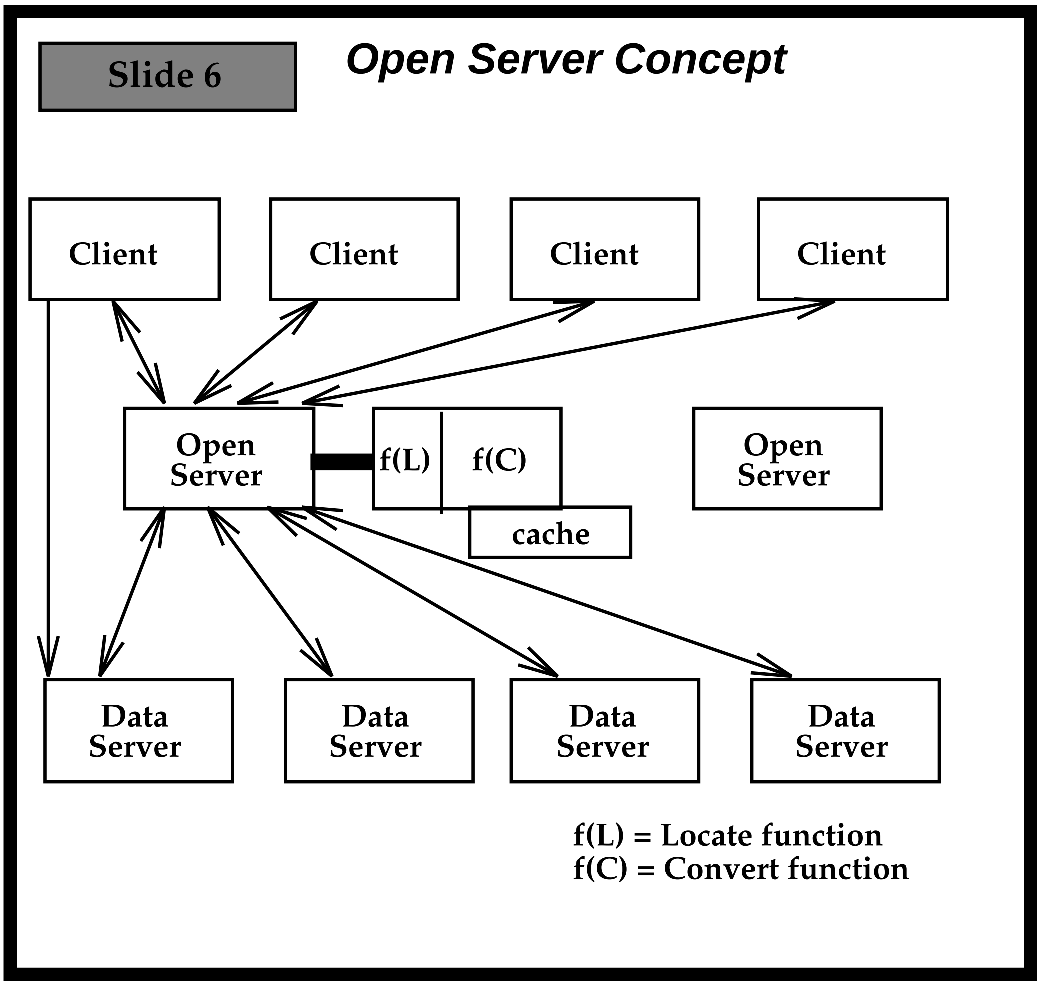


R. Chinman - if master directory acts as repository of data, is it serving as a layer between two clients? Issues - locating information; clients/local dictionary; master dictionary server; dictionaries in servers; polling

H. Debaugh - can we have an additional "open" server between clients. (will be addressed in subsequent slide)

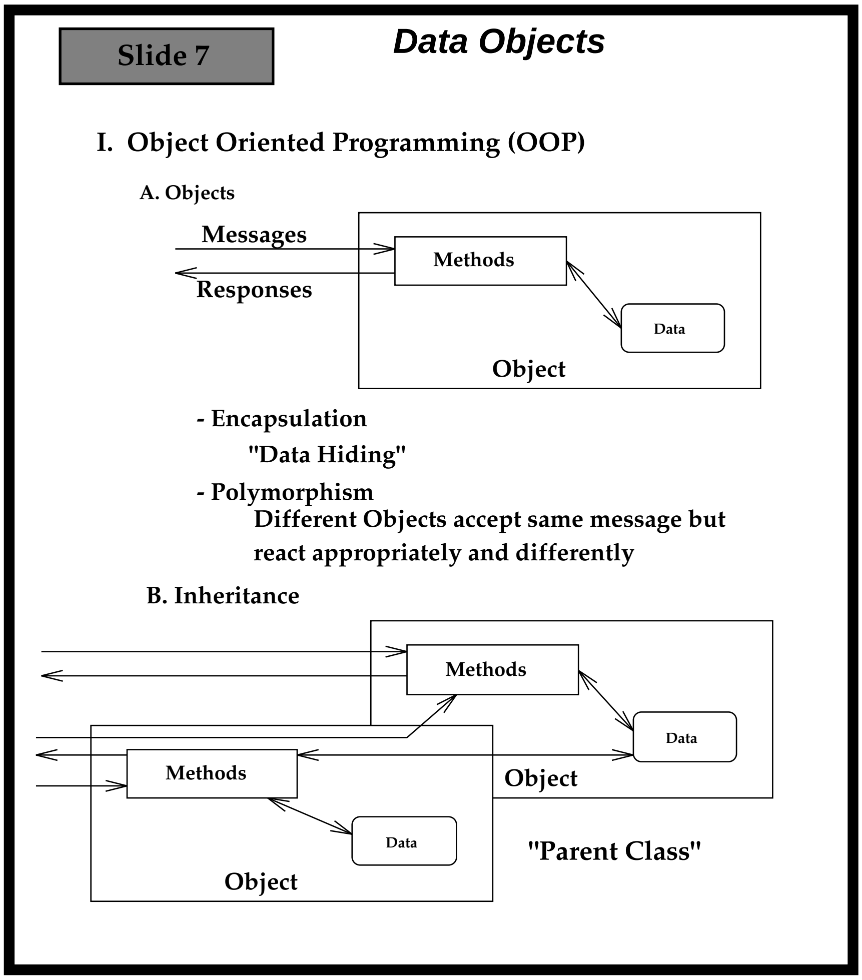
Slide 6:

- Types of Queries –
  - sub-selections
  - other SQL functions
  - incremental requests
- Types of Responses–
  - documentation of data set
  - identifiers
  - attributes
  - values/types (real numbers, integers..)
  - structure (most oceanographic data is not highly structured)



Slide 7: Data Objects – Object Oriented Programming overview,

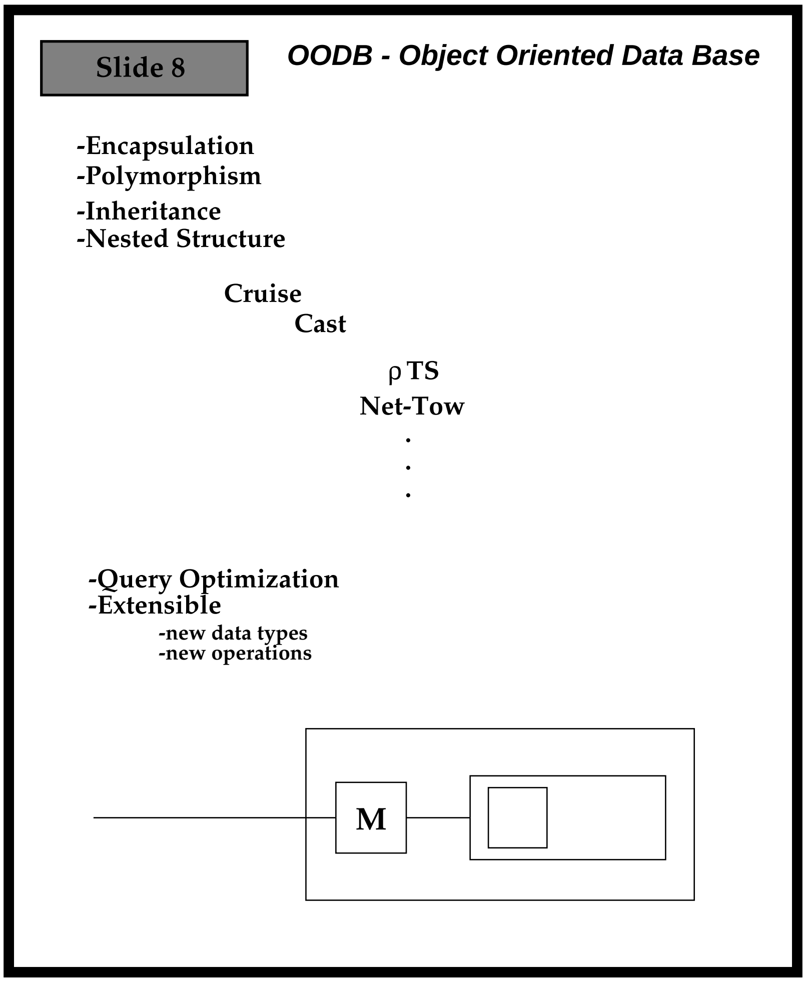
- program is built up of objects
- encapsulation (data hiding - you do not see internal information)
- polymorphism
- different objects accept same message but react appropriately and differently
- Inheritance – you can have a "parent" class with its own methods and data, and a sub-class with less capability. If the lower level cannot process the request, it will be passed up to the "parent" for processing.



Slide 8: Object Oriented Data Bases (OODB) –

- encapsulation
- polymorphism
- inheritance
- nested structure (inheritance)
- query optimization
- extensible
- new data types
- new operations

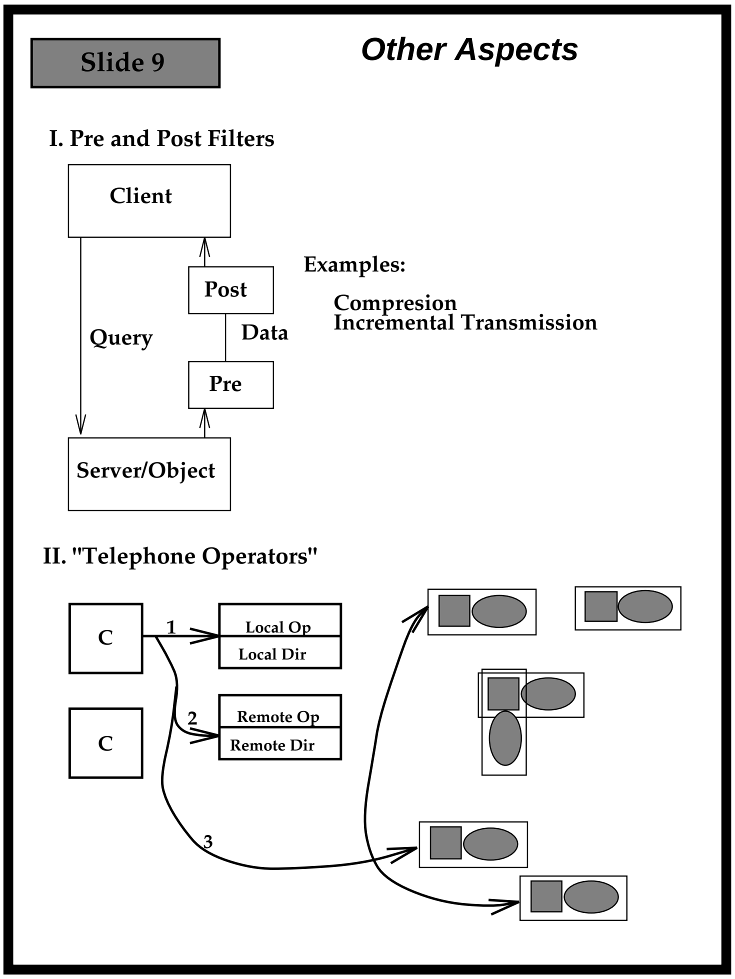
The vision which we are looking for is to define an appropriate combinations of the "clients-servers" structure in an OODB environment.



Slide 9: Other Aspects–

I. Pre- and post- filters (data compression)

II. Telephone Operators - one object may talk to multiple objects or sets to get the information out. Standard way to get data if you don't know where it is would be to call up master directory which will locate the data you want. Remote directory may update your local directory once the data are located. - keeping local system current will be a challenge, depending on frequency of use.



H. Debaugh - concern about data server doing compression and other tasks may cause problems. There may be other machines/programs in the link.

B. Schramm - a variety of providers will be available. Do we want to have different "classes" of servers instead of trying to put them all in one "box"?

G. McConaughy - are "server" and "object" the same thing?
G. Flierl - Yes, at this point in discussion.

H. Debaugh presented viewgraph interpretation of open server. (Slide 6 included in notes)

G. McConaughy - "scalability" is an issue of concern. We want to keep knowledge close to the data. "Intelligent" servers, "minimally- intelligent" routers, and "stupid" clients....

T. Kelly - open server lets you run more efficiently. The open server only has to order data once, and can then redistribute the data multiple times.

D. Collins - the open server provides an opportunity to bypass the "bottleneck" once you have established a repeated need for similar data. The "left" direct arrow on the diagram can be used then.

N. Soreide - what is in the open server? Does it have metadata, or is it just a "traffic cop"?
G. Flierl - Specifically, it would know where the data is, and would be able to respond to some data directly, but would have to go out and get other data.

Three important functions:**Location, Routing, and Caching**

J. Gallagher - maintaining open server with small number of formats on the bottom will be easy, but once the number of formats is large this will be hard.

(back to Flierl's slides)

III. Feedback, results from previous queries used to constrain/optimize current query

IV. Hypertext

Issues:

- which elements?
- directory protocol
- updates
- long-term stability
- redundancy
  - versions
  - updates

- protection/privilege
- credit for data sets

R. Chinman - what happens when an oceanographer takes their workstation with them - can be an issue for stability of a data base which may reside on that system.

D. Fulker - if result of a query gives you location of many different versions of the same data, then there is a problem. Must uniquely identify data sets.

B. Douglas - AGU has issued requirements to clearly identify data cited in research, in a manner such that another researcher could uniquely access it.

B. Douglas - "30-year" rule - will the data still have value in 30 years.

W. Brown - can't archived data be a distributed set of centers?

G. McConaughy - offered software (EOSDIS Version 0) as a base for the system we are trying to develop.

E. Dobinson - can we get a JGOFS overview?
P. Cornillon - JGOFS is just an approach we're looking at. It is not the model for the system we're developing. JGOFS requires a UNIX workstation, there is no automatic update. There are many things about JGOFS which should be improved.

B. Schramm - CD ROMs are becoming a common data format, and would be especially useful for maintaining "local" data. Can this group foster a format for CD ROMs?

J. Corbin - Do we envision unrestricted access and no data charges for this system?
P. Cornillon - it's an issue we need to address, but one which is fairly easy to look at.

L. Walstad - there must be some restrictions so that someone doesn't ask to download a 30 GB data set on internet.
P. Cornillon - X-Browse system allows certain access to the data for free, but you can't take the entire image archive and download it. The objective is to limit how long the line is used by one person.

Afternoon session

Broke into working groups for the afternoon

Working Groups:

- System Designers
- Data providers
- Data Users/Scientists
- Data Objects

Evening Session

Presentations by working groups

Data Users Working Group

W. Brown– Chairman

Data User Requirements

"Data" - primary "observations" and information

(O) *Must provide access to both investigator and archive-held data*

(1) Data Selection

- Ability to "locate" specified data in terms of: parameter, time-space window, type, source, QUERIES
- Refinement of search
- Coincident parameter search
- Flexibility to handle "unknown" parameter selections

(2) Data Acquisition

- Timeliness - electronic and/or mail delivery possible
- "Useful" Form - interactive and/or batch access
- Interface with simple programming languages

- Different formats
- Data subset selection possible

(3) Redundant Data Screening

(4) Data transfer protocol(s)– Investigator to Archive

————————-

Data Providers Working Group

R. Chinman–Chairman

Representation of the group:

- NOA-NOS, FNOC
- NODC
- NOAA-NOS (Ocean, Lake Level Div)
- PODAAC
- PI
- NOAA-NESDIS (DMSP, ERSY, METEOSAT, GMS)
- UNOLS R/V Tech
- Global Change Data Center GDAAC
- TOGA COARE

Dataset Classification:

- Open-ended
- Project (Closed)
- Rotating
- Orphan (incidental data to another study)
- Real-time

Providers - PI
Users - Data Center
repetitive data distribution tool data discovery tool (e.g. sociologists)

The types and kinds of data and providers and users *suggest/requires a relatively wide range of options/capabilities for distributing data*

What data providers need from or will do for the distributed ocean data system:

- Data for a wider community than ocean community alone
- Internet based
- Generation and distribution of metadata critical!! including citation info , algorithms, when and where data collected, data set version caveats
- MD-like locator needed for this system
- DIF-like metadata file needed for the locator, for the data file to be findable on the system
- Provider-specific capabilities for instituting restrictions, privileges, protection of datasets including log of users and activities
- Distribution of processing capabilities restricted to data format translation and subsetting

Discussion about data sets including appropriate credit for the PI who provided the data to the data center, and the requirement that a second investigator can obtain the exact same data set, apply the stated model or algorithm to it, and derive the same result.

P. Cornillon - MD is not populated as much as it should/could be because of the excessive documentation requirements for data sets

- Ease of installation - higher overhead of server installation acceptable, but ease still important
- Data format translators necessary, with at least the following translations:
  - GRIB/BUFR
  - HDF
  - net CDF

—————————————-

System Architecture Working Group

E. Dobinson– Chairwoman

Basic Constraints

1) Cost - reusable code, functions, systems

2) Schedule - 9 working months to build system

3) Extensibility (grow with time) and Scalability (start small then grow dynamically)

4) Simple to use, easy to install, easy to use

5) Dynamic - easy to grow and administrate

Given these constraints, what is essential functionality?

- Location and Search function primary - user must be able to locate and find data (look at USGS system?)
- Need an order function to provide the user with the data he or she wants..
- Services a server can provide may vary, based on the data set (BASIC CORE SYSTEM)

Discussion of Client/Server viewgraphs:

End of Day 1

Day 2:

Data Objects Working Group

G. McConaughy– Chairwoman

Started with an overview of existing systems.

Search to data access is range of service for most existing systems. EOSDIS is mostly search, JGOFS much more focused on data access. IDBMS mostly one large data base.

B. Schramm - net CDF (UniData), NEONS were left out.

EOSDIS Version 0 assumption is that you're sitting at your client, and it sends out a search message and gets a response message based on who it is talking to. None of these messages is visible to user. It is using ODL (Object Definition Language). A lot of info isn't seen by client.

MEDS/Gopher system - messages coming back are menus which are data services at sites. Inventory search and results and data order - you get back a "form" to fill out with data order and processing options. Some software is provided in the client for looking at data. It is less of a search system and more of an ordering system. Burden on data provider to hook in would be very small.

IDBMS has 9 nodes (ORACLE-based) with centralized node with Master Catalog where data are replicated and also used for data ingest. Each server holds a different type of data, user does not even know which node data is being accessed. Search can provide either metadata or primary data - similar to how EOSDIS works.

JGOFS - client has access to servers through calls. There is a server and a data dictionary which has a list of objects, as well as a path to tell how to execute the method. There will be different entries in the dictionary for data and methods. Tagged with the data are metadata. The client program connects to an executable remote – all of the dictionary serves to make this happen. Metadata can be searched to come up with data objects that may satisfy the users. Can list all variables which can be extracted from the data - including looking at multiple servers to do this based upon the entry in the dictionary. Focus of system is to support a PI. Every server has a data dictionary for all holdings.

R. Mairs - how would/could JGOFS send a message to EOSDIS?

G. Flierl - suppose you wanted SST within some latitude bounds, and you seek a yes/no answer about the data. This is very much the kind of information EOSDIS handles.

P. Cornillon - discussion about "capability" of a system, at the frequent expense of ease of use and menu-driven operations. EOSDIS may be the system designed for the broad range of users who desire ease of use, while a system like JGOFS is "harder" to use but of higher value to a PI who has more specific data requirements.

B. Starek - discussion of "query by example" on IDBMS. ARCInfo keys to databases (Common Production Tools - like methods).

P. Cornillon - our system must provide a very basic level of easy-to-use services yet have expert capability to support science users. A modified JGOFS may be a good starting point.

G. Flierl - discussion of JGOFS "Problems":

1) Feedback (in "client"?)

2) Limitations on query: <, =, >, <=, >=, ( ), &, | or ˜ but no more than 20 strings grouped together

3) Data types: all ASCII doesn't fully deal with matrices/tensors

4) No interactive retrieval (like XBROWSE which lets you "look" at an image as you begin to receive it and decide if it will meet your needs .

G. Flierl - Summary of support requirements

1) Multiple Clients/User Interfaces

- menus/hypertext ...feedback
- Fortran programmers ...simple full query specification

2) Data servers

a) inventory and location

b) browse

c) data, including points to other data

3) Pre/Post filters

4) Multiple Servers– routing

5) PIs adding servers– support multiple DBMS

L. Walstad - How will we handle requests for 2-3 GB of data?
G. Flierl - Server will respond, but indicate the data block is too big. It may suggest transfer by mail (tape).

P. Cornillon - list of problems he has encountered in using systems: One person wants "SST", a second searches for "sea surface temperatures". Variable names need standardization, and we need to make our locator object be "intelligent"

G. Flierl - discussion of data structure:

Hierarchical Data Structure Diagram

Afternoon session

The workshop will not try and endorse a specific data format in the afternoon discussions - this question cannot be resolved in a workshop of this scale.

We will break up into Users and Providers.

Discussion: How much of the definition of communication protocol must this group consider?

Developers can send out a bit stream from their servers in some format, and along with this is a structure (really a series of arguments) which defines this format. On the other end, the user must be able to "understand" this structure and apply it.

We're confusing communications protocols with the applications interface.

G. Flierl - API has two calls—

call and get variable names (strings, numbers)

call and get values (strings, numbers)

RPC vs Data Stream Diagram

Top figure illustrates an RPC-based system while the lower illustrates the data stream approach

N. Soreide - is "data volume" the amount of data (MBs) or number of data requests?
G. Flierl - A lot of research is being done via individual exchange of small data sets – not transfers of large volumes of data from archives. It is important the system support these "small" users...

J. Gallagher - MB and GB per day is "large"; less than MB is "small" (although these definitions are more commonly based on how long it takes to receive the data, which is driven by communications technology...)

Data Providers Subgroup Meeting - R. Chinman Questions: What Objects? What Structure? What is an object? - An object includes

a. inventory,

b. browse

c. data

(these are its attributes). To what extent are data providers willing to provide these? Inventory (for an image, for example) might be "is there an image?" Browse - defined as a mechanism for sub-sampling.

B. Schramm -- We need to look at two different cases. The Data Center will be significantly different from the PI in its implementation of these capabilities.
DATA: standard file format   location   disk space

Evening Session September 30

P. Cornillon presiding

System Functions

Selection

- Search (space/time, parameters, source, sensor, other), cross-inventory
- Refine search - optional but necessary (browse, quality, log)

Order

- Means of delivery and where
- Timelines
- Format
- Either from the selection process or directly (bypassing selection process)

Principles

Must provide access to PI data sets in addition to national archives (corollary) — people should want to use the system to manage their own data

Data Issues

Provider of the data must be able to determine what data objects (and order) to be delivered

Mechanism to resolve keyword conflicts (auto detection)

Global naming procedure for data sets

Central clearing house - system management (log)

We must avoid giving people a reason to NOT want to put their data into the server

Messages

Functional description of data, e.g., for location or defined range of formats

Options available from provider

Must allow (and in some cases encourage) a description of data set or provide a pointer to a description (search) – example, the "Gulf Stream Paths" data set.

Log File - Data set PI invites comments on data sets - central storage

Be able to explore the system

Discussion –

D. Fulker - provision of data needs to be in a form suitable for use in an applications program (API)

Data Providers - R. Chinman

Desire to remove all impediments to using data

Desirable functions and structure

inventory - locator

browse - refinement

data —*server*— standard format

inventory:

id

latitude

long

date/time

PI

sensor

parameters

cannot translate all native format data into one of the standard formats - PIs nor Data Centers So... use server software to do that and inventory (if necessary) Use existing inventories via server translators on new system.

Benefits

PI: fulfillment of contractual obligation to make data available to National Data Centers, access to Data Center data and other PI data

Data Center: get access to field project/PI-based data for their users

From the Data Providers (PI and Data Centers) will come a range of services

Ocean community should prioritize DATASETS and tell J. Gallagher

Data Centers provide an existing array of services and fold 2-3 into new system, e.g., EOSDIS design group, Emery EOSDIS testbed, NOAA DAC (includes NODC)

E. Dobinson - question about why this system (DODS) seems initially focused on large existing systems (EOSDIS, MEDS,NEONS.....) which have or will have established mechanisms for getting their data, instead of looking at developing a system uniquely but not exclusively designed to gain access to small, non-automated datasets held by PIs . Response from P. Cornillon centered on difficulty in becoming familiar with all of these different client/server system (or even knowing of them as they multiply) , making the data received from them suitable for use on your system, and having to wait until some of the big systems (e.g., EOSDIS (July 94)) are ready to become operational.

End of Day 2

Day 3

Conclusions

G. Flierl:

Next Steps:

0) Meeting report(s); mailing lists and telemail

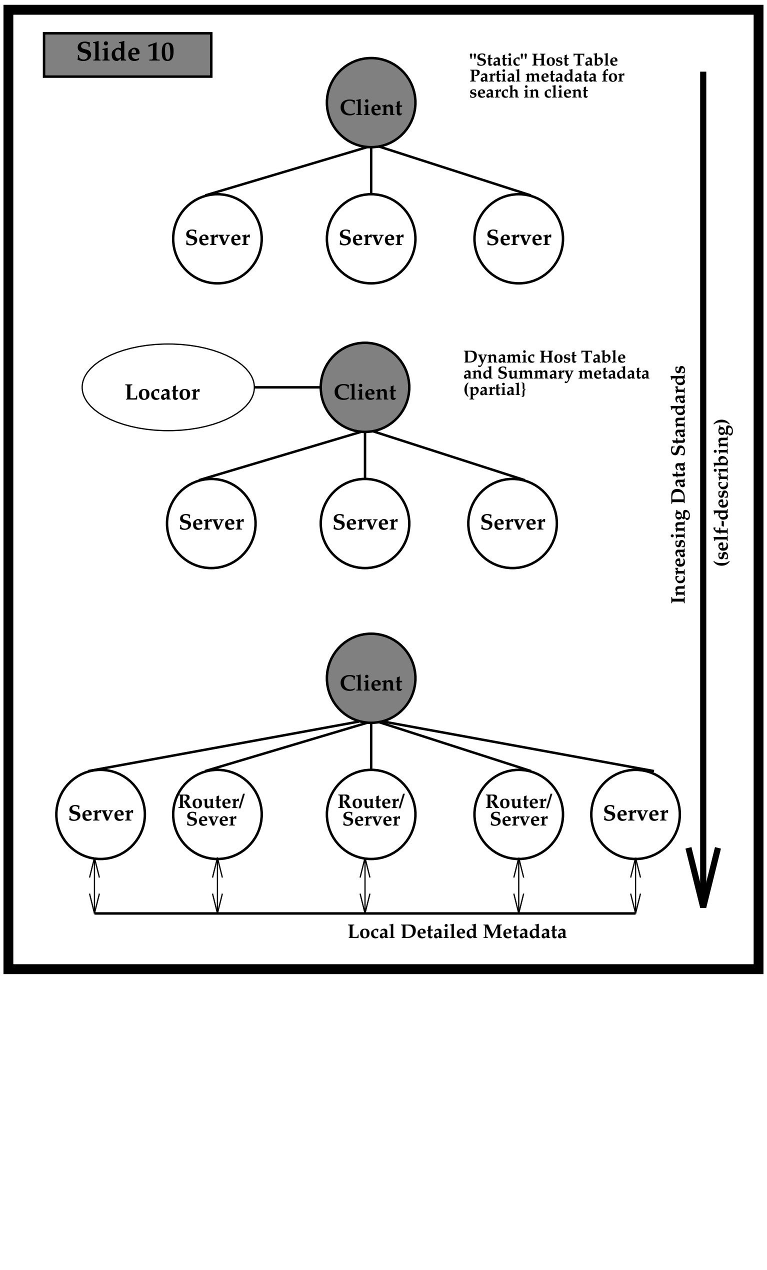
1) Architecture Strawman / Message Strawman
   - tell Gallagher
   - dialog or draft
   - join the development group

2) Testbeds
   - provide data/ work with URI/MIT
   - provide support (moral and personnel)

3) Next Workshop (planned elsewhere than Alton?)
   - focus on people who want to develop servers and clients

4) Colleagues

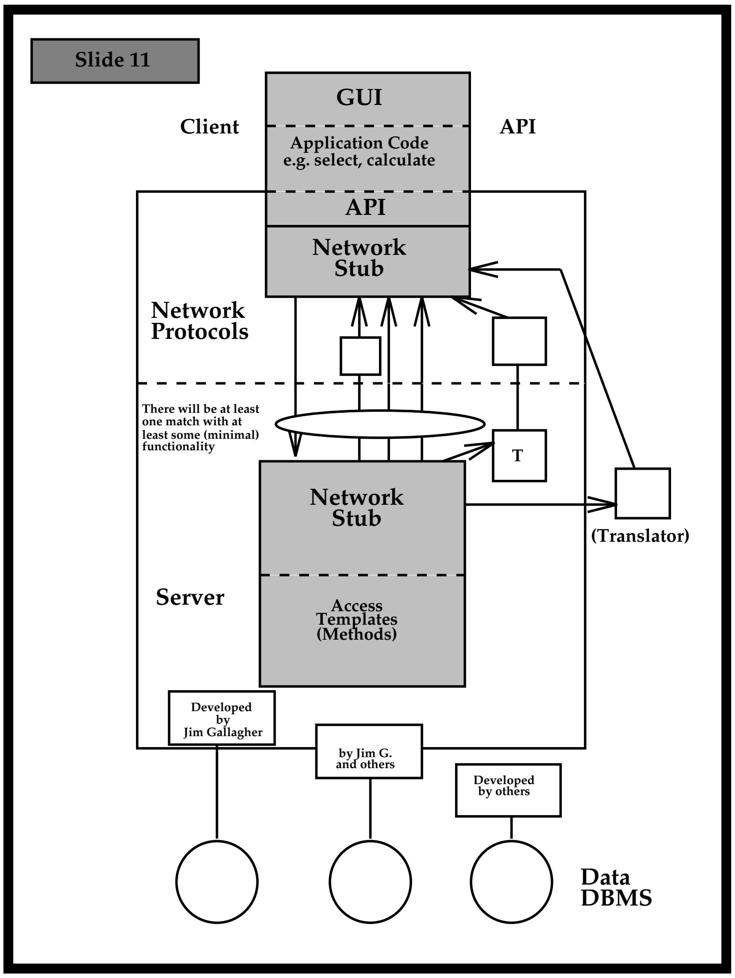
Software development will be undertaken in a phased approach. (Slide 10)


Slide 10
"Static" Host Table
Partial metadata for
search in client
Client
Server
Server
Server
Dynamic Host Table
and Summary metadata
(partial}
Locator
Client
Server
Server
Server
Increasing Data Standards
(self-describing)
Client
Server
Router/
Sever
Router/
Server
Router/
Server
Server
Local Detailed Metadata

R. Chinman - Is the name "DODS" appropriate? It focuses on oceans, yet there is a significant community outside "oceans" who may have interest in this work.

G. McConaughy - there may be an advantage to keeping it as an "oceans" system, for if we broaden its basis (as implied by a more encompassing name) it may sound like it will try and do things which are already being done.

Discussion and refinement of system graphic (Slide 11).



Is there a base level communication procedure? Will there always be a match?

1) May not be efficient
   - ASCII
   - XDR binary
2) Must be buildable from DODS distribution

P. Cornillon - One implementation option might be to just put an additional server on a site. We didn't consider this as much as we perhaps should have.

D. Fulker - discussed his concerns about locating translators on the "client" side.

The meeting concluded in general agreement of the system goals and functionality.

End of Day 3

C Proposed System Architectures

One of the explicit goals of the workshop was to recommend a system architecture for the implementation of the DODS. The system developers focus group discussed in detail several different system models on the first day. In their report to the plenary session, they recommended the client-server model as appropriate for the implementation of DODS. In this Appendix we elaborate on the client-server architecture and show three different ways in which it can be implemented.

The system shown in Figure 1 has translators located at the data servers which translate the format of the data resource into a canonical intermediate format used for transmission. The client component of this system reads and parses the data stream. User programs access the parsed data using a API. The data model implicit in the semantics of this API will closely match that of the canonical format used for transmission (although it could, in theory, be quite different, there is little reason to make it so).

In figure 2, the read and parse operations are moved out of the client API and into the data server. The data format implicit in the API is mapped to the data resource format either directly or using an intermediate format. Using an intermediate format would add complexity to design and might restrict the total number of data resource formats accessible. However, the presence of an explicit intermediate format would no doubt simplify support for different data resource formats and (possibly) APIs. This design is also capable of supporting limited random file access calls since it is not constrained by a serialized intermediate format.

The system in Figure 3 is significantly different from either of two preceding figures. Both figures 1 and 2 share a crucial feature; the lowest level of access to the system is through an API we provide. In order for programs to make use of the system directly they must be, at minimum, relinked with our API. The system in Figure 3 is designed to overcome this problem. Rather than develop a data specific API, the system in Figure 3 uses a special file system which has translators located in both the client and server components. The translators on the server side produce a data stream in response to requests for data from the translators on the client side. The translators on the client side are accessed not from an API we supply but from the UNIX file system calls (open(), read(), ...). In this system, users could choose from one of several translators on the client side so that the same file could be accessed as ascii, html, or netcdf, for example. The choice of translators would be accomplished using a special syntax for the file names. Implementation of this system requires modification to the Unix kernel.

Each of the three systems pictured here represent different tradeoffs in extensibility, generality and simplicity. The last design is the most complex to implement and maintain, since creating a file system and supporting that file system on several platforms is complex. System 1 and 2, however, are fairly simple to implement. This is balanced by the relative generality of the three basic designs. Systems 1 and 2 can only be used with programs we (or others) explicitly modify to use the DODS API.

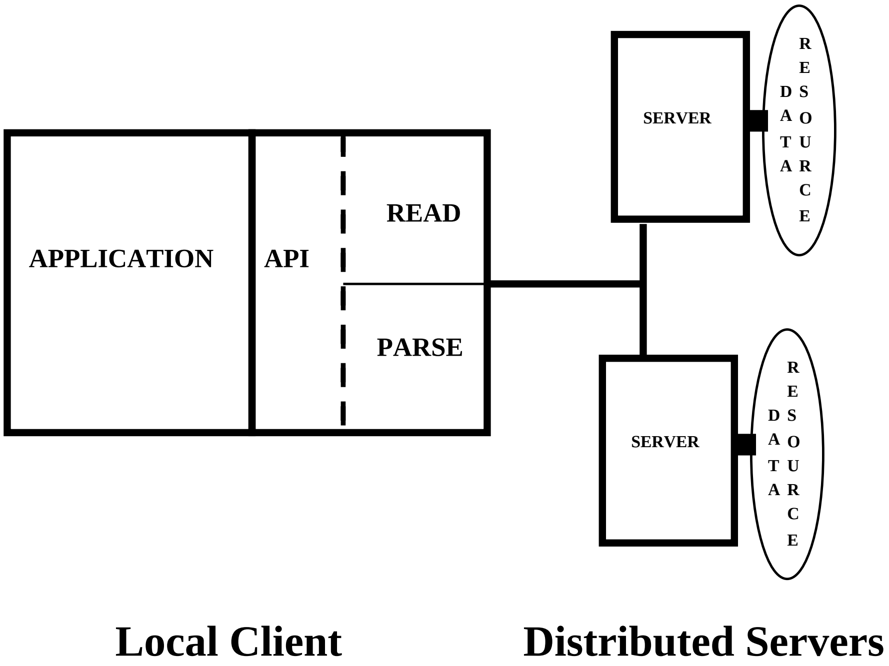


Figure 1: The Data Stream Client-Server Model for DODS

System 3, however, provides access to DODS data using Unix system calls (think of NFS) so programs which access DODS data will not have to be specially modified. Existing software would access DODS data using a special pathname which specifies the remote data (like pathnames of NFS volumes specify remote data) and both client and server format translators.

The above discussion presumes that arbitrary format translation can be accomplished (i.e., that any format can be translated to any other format). This is not true. However, if the choice of intermediate format(s) in the above systems is made wisely, a wide variety of data formats can be supported.

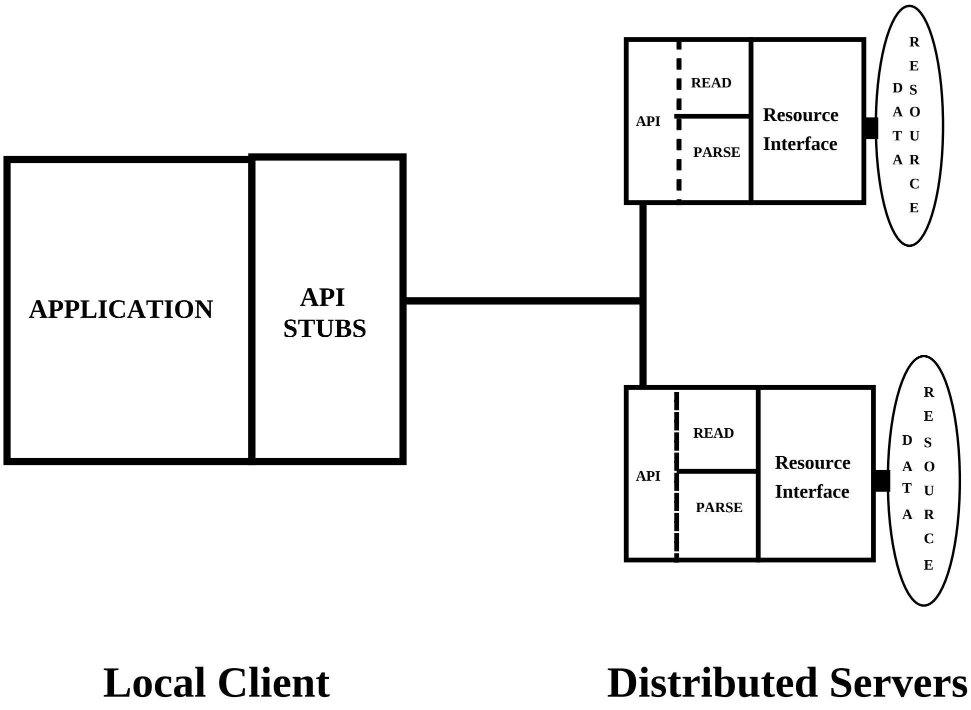


Figure 2: The API Client-Server Model for DODS

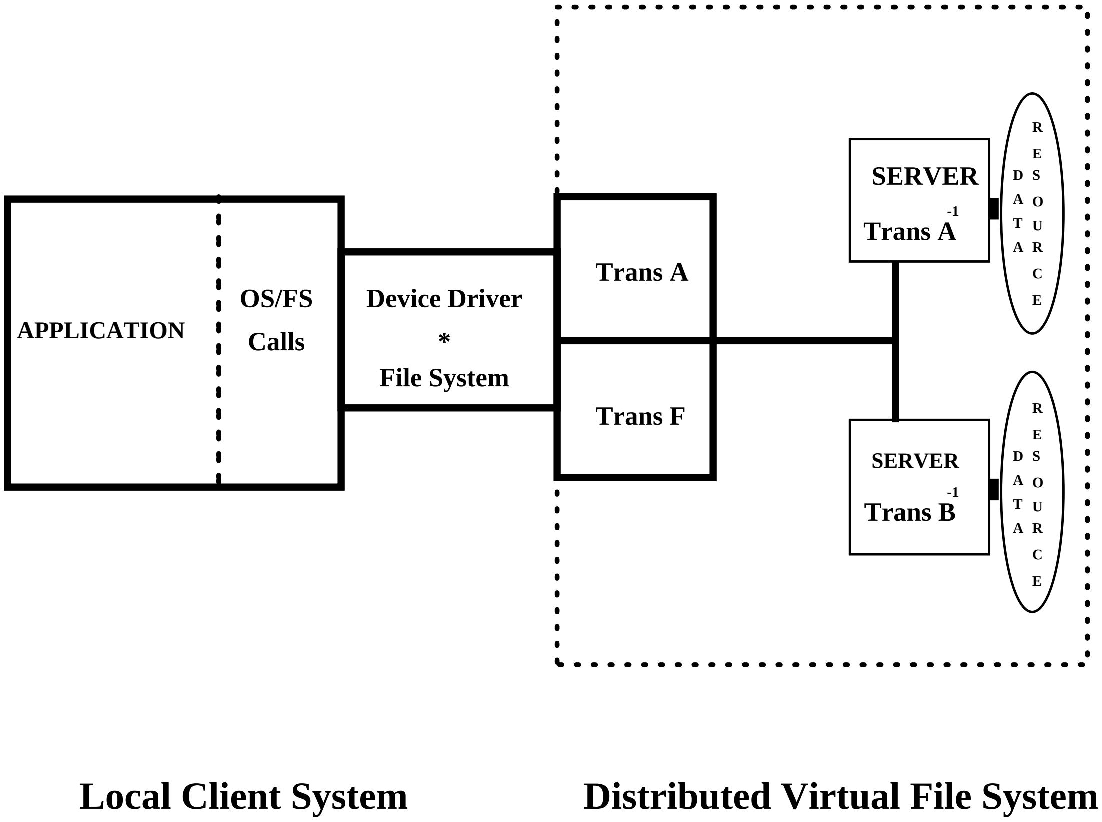


Figure 3: The Virtual File System Client-Server Model for DODS

D Participants

Addresses of participants can be found with their background statements.

BACON, Ian	ian@getafix.tsg.com
BASS, Bill	bill@eos.hac.com
BROWN, Wendell	wsb@panthr.unh.edu
CHINMAN, Richard	chinman@ucar.ncar.edu
COLLINS, Donald	djc@shrimp.jpl.nasa.gov
CORBIN, Jim	corbin@cast.msstate.edu
CORNILLON, Peter	pete@petes.gso.uri.edu
DEBAUGH, Henry	
DOBINSON, Elaine	elaine_dobinson@isd.jpl.nasa.gov
DOUGLAS, Bruce B.	DOUGLAS@omnet.nasa.gov
ENLOE, Yonsook	yonsook@killians.gsfc.nasa.gov
FLIERL, Glenn	glenn@lake.mit.edu
FRANK, George	george@galaxy.ngs.noaa.gov
FULKER, Dave	dfulker@unidata.ucar.edu
GALLAGHER, James	jimg@dcz.gso.uri.edu
GILL, Bob	gill@colts.nodc.noaa.gov
GIVEN, Jeffrey	jeff@gso.saic.com
GLOVER, Dave	david@plaid.whoi.edu
HANKIN, Steve	hankin@ferret.pmel.noaa.gov
HOGG, Roen	roen@oce.orst.edu
HOLBROOK, Jim J.	holbrook@pmel.noaa.gov
IRISH, Jim	jirish@whoi.edu
KELLEY, Tim	kelley@sanddunes.scd.ucar.edu
MAIRS, Rob	rmairs@saars1.fb4.noaa.gov
MILKOWSKI, George	george@zeno.gso.uri.edu
MILLER, Chris	miller@esdim1.nodc.noaa.gov
McCONAUGHY, Gail	gailmcc@boa.gsfc.nasa.gov
McDONALD, Ken	mcdonald@nssdca.gsfc.nasa.gov
NEKOVEI, Reza	reza@uri.gos.uri.edu
OLSEN, Lola	olsen@eosdata.gsfc.nasa.gov
RHODES, Judi	OCEANOGRAPHY.SOCIETY@omnet.nasa.gov
SCHRAMM, Bill W.	SCHRAMM@omnet.nasa.gov
SCHWENKE, George G.	schwenke@se.hq.nasa.gov
SOREIDE, Nancy	nns@noaapmel.gov
STAREK, Bob	
WALSTAD, Leonard	walstad@oce.orst.edu
WHITE, Warren	wbwhite@ucsd.edu
WILSON, J R.	R.WILSON.MEDS@telemail.nasa.gov

E Participant Background Statements

Ian Bacon

Personal-

Name: Ian Bacon

Title: Program Manager, NASA Programs

Affiliation: Telos Systems Group

Address: 14585 Avion Parkway, Chantilly, VA 22021

email: ian@getafix.tsg.com

phone: (703)802-1730

fax: (703)802-0718

Data System-

Data system name: Space Flight Operations Center, JPL

Discipline: Planetary Science

Data Managed:

Type of Data: all types returned from deep space probes
Inventory Meta Data [Y/N]: Y
Digital Data, Data Products [Y/N]: Y
Number of Data Granules:
Total volume of Data [Megabytes]: more like Gigabytes

Data Management Activities Summary-

Telos has experience with both distributed data systems and the object oriented world, through our work at JPL, and also through internal R&D projects. At JPL, we have been responsible for the Space Flight Operations Centre, previously called SFOC, now called AMMOS. This is a distributed system, comprised of over 500 UNIX workstations, spread across the country, networked together over the Internet. Telos designed a significant portion of this system. It is used to control spacecraft, but also to collect the level zero data, process them to level 1, and make them available in quick-look form to interested parties. We have also been responsible for the Command

Sequence Generator, used for Magellan and other systems. It takes maneuver profiles for various spacecraft, and converts them into the command sequences which are uploaded to the vehicles. SeqGen was designed using object oriented design techniques, and was developed in C++. It includes a small object oriented database which was developed specifically for the project by Telos.

On the East Coast, Telos has been working with object oriented database systems on a project originally designed to meet the needs of one of the Intelligence agencies. Project Headroom, as it is currently known, functions as a distributed database system, capable of extracting data from a variety of different database types (including Sybase and Oracle, as well as flat files, network databases, and others). The data can be accessed over a network, and are joined into a single data object, in an object oriented database. We are currently working on a prototype of this system for use with scientific datasets. The version which we are developing allows analysts to select data from multiple datasets without having to know specifically where they are located, effectively creating a fileless system. When the data are contained in monolithic files (such as HDF files) the system browses Metadata files, and extracts only those data points of interest to the analyst, so it is not be necessary to pull huge files over the Internet, and then manually extract the data. Data from several datasets can be joined into a single data object, when there are areas of commonality between them (such as dates or geographical locations, for example). The data, once retrieved, are available for processing by public domain or user specific tools, which can be incorporated into the object oriented database as methods. The client side includes a very powerful search tool, which is capable of performing contextual searches of text based information. We are also currently binding IDL into the client, to allow the analyst to use its powerful tools on the data object which the system builds. Our goal would be to distribute this system over the ECS network, with client systems at the SCFs, and servers at each of the DAACs.

Bill Bass

Personal-

Name: Bill Bass

Title: Principal Scientist

Affiliation: Hughes Applied Information Systems, Inc.

Address: 1616A McCormick Drive

email: bill@eos.hac.com

phone: (301)925-0304

fax: (301)925-0327

Data System-

Data system name: Eos Data and Information System Core System (ECS)

Discipline: Atmosphere, Ocean, Land, Cryosphere, Interdisciplinary

Data Managed:

Type of Data: From Raw Remote Sensed Data to Geophysical Parameters
Inventory Meta Data [Y/N]: Y
Digital Data, Data Products [Y/N]: Y
Number of Data Granules: 200000/day
Total volume of Data [Megabytes]: 1 TB/day

Data Management Activities Summary-

I am a system engineer working on the development of a data and information system for a repository and distribution system for a large quantity of remotely sensed data. I am currently investigating means of generalizing this NASA system to a global change data and information system and, beyond than, to a system in which users may easily become data providers as well as data users (sometimes called User-DIS).

Wendell S. Brown

Personal-

Name: Wendell S. Brown

Title: Professor of Oceanography

Affiliation: Inst. for the Study of Earth Oceans and Space, UNH

Address: OPAL/EOS Morse Hall, UNH, Durham, NH 03824

email: kmg@kepler.unh.edu

phone: (603)862-3153

fax: (603)862-0243

Data System-

Data system name: Environmental Data and Information Management System (EDIMS)

Discipline: Marine

Data Managed:

Type of Data: NOAA/realtime meteorological data
Inventory Meta Data [Y/N]: y
Digital Data, Data Products]Y/N]: y
Number of Data Granules: 8 files/day
Total volume of Data [Megabytes]: 0.002/file

Type of Data: Coastline/bathymetry
Inventory Meta Data [Y/N]: y
Digital Data, Data Products [Y/N]: y
Number of Data Granules: 30 files
Total volume of Data [Megabytes]: 1

Type of Data: UNB riverflow data
Inventory Meta Data [Y/N]: y
Digital Data, Data Products [Y/N]: y
Number of Data Granules: 1 file/day

Total volume of Data [Megabytes]: 0.001/file

Type of Data: AFAP dataset

Inventory Meta Data [Y/N]: y

Digital Data, Data Products [Y/N]: y

Number of Data Granules: 1 file

Total volume of Data [Megabytes]: 30

Type of Data: Massachusetts Bay dataset

Inventory Meta Data [Y/N]: y

Digital Data, Data Products [Y/N]: y

Number of Data Granules: 745 files

Total volume of Data [Megabytes]: 8

Type of Data: NOAA/NOCN SST dataset

Inventory Meta Data [Y/N]: y

Digital Data, Data Products [Y/N]: y

Number of Data Granules: 0-8 files/day

Total volume of Data [Megabytes]: 0.3/file

Data Management Activities Summary-

EDIMS CONTEXT

The Gulf of Maine Council on the Marine Environment (GOM/CME) has developed a ten-year Gulf Action Plan to address issues concerning the environmental health of the Gulf and management of the marine resource. Toward that end, the GOM/CME Working Group established a Data and Information Management Committee (DIMC) to develop an Environmental Data and Information Management System (EDIMS) for the Gulf of Maine region. I was chosen to lead a University of New Hampshire (UNH) effort to design and implement a prototype EDIMS which is now available for broad community use.

EDIMS CONCEPT

The EDIMS is designed to make a data directory and a relevant set of data bases accessible to EDIMS users. The prototype EDIMS is designed to be simple and flexible enough to accommodate anticipated changes. We expect that the eventual user group will include marine environment and resource managers; state and provincial planners; and ocean scientists and engineers. As users become more familiar with the prototype EDIMS they will articulate their needs more clearly. The development of EDIMS will be guided by that feedback.

The fully operational EDIMS will need to deliver to its users a large and diverse blend of archived and real-time data from operational and research sources. Thus EDIMS is structured around a decentralized database. Figure 1 illustrates schematically how data users/suppliers from the states, provinces and federal agencies will be able to exchange data and information via the Internet network. This approach relies on the effort and resources of the data users/suppliers and thus can be expanded relatively easily.

While the EDIMS data directory and a few special data sets will reside at a host computer site, most of the EDIMS databases will reside at remote locations. EDIMS data users/suppliers will link to the network and thus have access to all of the information and data at the EDIMS host site (presently UNH), as well as the remote EDIMS sites. As envisioned, EDIMS data users/suppliers will be asked to assume a major share of the responsibility for database quality and maintenance. They will be assisted by an EDIMS manager at the host site who will oversee the operations of the EDIMS and implement improvements to EDIMS. We expect that different data users/suppliers in the region will commit to and support such an EDIMS because of their need to access the comprehensive EDIMS database. This information will enable them to conduct research, protect public health, and/or manage the Gulf of Maine marine resource better than ever before.

PROTOTYPE EDIMS

A UNH development team has constructed and implemented the prototype EDIMS. The prototype EDIMS consists of a few representative databases (see below), which can be accessed by a broad user community. We have developed a set of protocols that will enable the group of data users/suppliers to (a) query a directory of the regional Gulf of Maine data sets; (b) electronically transfer a selected subset of these data to their own computing environment; and (c) communicate generally with the prototype EDIMS user community. We have selected a diverse set of databases to be part of the prototype EDIMS. They are:

- A Gulf of Maine database directory
- Documentation (incl. EDIMS User Manual)
- Gulf of Maine maps and bathymetry
- Massachusetts Bays Program physical oceanographic data archive
- Real-time satellite imagery and meteorology from the NOAA/NOS Ocean Products Division
- The New Brunswick Department of Environment real-time river discharges
- The USGS sediment texture data archive
- Dartmouth model "data" for the Gulf of Maine

- The Bedford Institute of Oceanography Atlantic Fisheries Adjustment Program (AFAP) hydrographic data archive

Our goal in the next six months is to add:

- A "Who's Who" in the Gulf of Maine
- The Gulf Watch mussel data
- The EOEA Massachusetts shellfish data archive

EDIMS includes a SQL/ORACLE-based data directory. A simple query system enables EDIMS users to browse the EDIMS database directory for information in user-specified time and space domains. This documentation directory and "capture" any of the electronic EDIMS databases. The prototype EDIMS electronic databases consist of a data description header and a flat ASCII data file.

In the prototype Gulf of Maine database directory, documentation and all but the Dartmouth databases reside on the EDIMS host client/server computer at the University of New Hampshire. (When fully implemented, most the EDIMS data bases will reside at their remote storage sites.) During the prototype EDIMS development, most the databases will be static, for the regularly updated NOAA and river discharge databases.

Internet is the conduit for the prototype EDIMS data and information. It is an established and well-documented international network, with data and mail transfer protocols that can be implemented on a variety of platforms. The Internet File Transfer Protocol (FTP) feature is used to retrieve selected data from the host and/or remote storage sites. In an effort to keep track of EDIMS use, we will monitor access to the EDIMS.

Donald Collins

Personal-

Name: Donald J. Collins

Title: Manager

Affiliation: Physical Oceanography Distributed Active Archive Center

Address: Jet Propulsion Laboratory m/s 300-323 4800 Oak Grove Drive Pasadena, California 91109

email: D.Collins/OMNET djc@shrimp.jpl.nasa.gov

phone: (818) 354-3473

fax: (818) 393-6720

Data System-

Data system name: Physical Oceanography Distributed Active Archive Center

Discipline: Physical Oceanography

Data Managed:

Type of Data: Satellite data of the oceans, including supporting data for verification. Higher level data products.

Inventory Meta Data [Y/N]: Y

Digital Data, Data Products [Y/N]: Y

Number of Data Granules:

Total volume of Data [Megabytes]: 10 Tb by 1995

Data Management Activities Summary-

The goal of the PO.DAAC is to serve the needs of the oceanographic, geophysical, and interdisciplinary science communities which require physical information about the oceans. This goal will be accomplished through the acquisition, processing, archiving, and distribution of data obtained through remote sensing, or by conventional means, and through the provision of higher level data products to the scientific community.

The PO.DAAC presently serves the broad scientific community, responding, without charge to the user, to requests for complete data sets and for subsets of data based on temporal and spatial criteria specified by the user. The PO.DAAC is responsible for the acquisition of well documented satellite ocean data products at all levels, from existing visible, infrared, passive and active microwave sensors, the distribution of these holdings to the scientific community, and the provision of data product documentation.

The Earth Observing System (EOS) Project at the Goddard Space Flight Center (GSFC) contains as one element the Earth Science Data and Information System (ESDIS). The ESDIS has been formulated as a distributed system, consisting of central functions at GSFC and Distributed Active Archive Centers (DAAC) at eight sites throughout the United States. The Jet Propulsion Laboratory (JPL) has been designated as the site for the Physical Oceanography Distributed Active Archive Center (PO.DAAC). The activities at each of the DAACs have been separated into Version 0 activities and into later versions. The concept for Version 0 is that these activities evolve from the capabilities of the present systems that have existed at the selected sites from discipline specific data systems and data activities. Version 0 will be operational in July, 1994 as a, "working prototype with operational elements", with prototype capabilities for an Information Management System (IMS), a Data Archive and Distribution System (DADS), and a Product Generation System (PGS).

Version 1 will be operational at the end of FY-97, utilizing hardware and software delivered by the EOS Core System contractor. The transition between Version 0 and Version 1 will be conducted without an interruption of service to the scientific community. Version 1 operations will be conducted by the PO.DAAC.

The activities of the PO.DAAC are focused on the provision of pre- EOS data sets to the scientific community during Version 0, on the establishment of an operational status for the PO.DAAC by July, 1994, and on the transition between Version 0 and Version 1.

The PO.DAAC will provide the data archive and distribution services for the TOPEX/POSEIDON mission, including the generation and publication of the Merged Geophysical Data Record, through the end of the nominal mission in August, 1995, and through the end of any extended mission period, presently assumed to be August, 1997.

The PO.DAAC will archive and distribute the data from the AVHRR Oceans Pathfinder, and will assume the continued AVHRR Oceans Pathfinder data production following the initial production of the data sets for the period 1981-1995 by the Pathfinder Task. The transition is assumed to occur at the beginning of FY-96. The Pathfinder Task will assume responsibility for a 1 km U.S. coastal data product in FY-95, and will continue to produce this data product after that time.

The PO.DAAC will continue the provision of data archive and distribution services to the U.S. WOCE and Scatterometry teams for the ERS-1 low bit rate data sets, and the provision of these same services to the U.S. teams for ERS-2, with a probable launch in mid-1995.

The PO.DAAC will assume responsibility for data packaging and formats for the NSCAT data products, and for the archiving and distribution of these products to the NSCAT Science Working Team, and to the scientific community. The PO.DAAC activities include preparation for, and the support of, the NSCAT mission, with launch in February, 1996. The PO.DAAC will also assume responsibility for the development of higher level data products as determined by the SWT and the PO.DAAC User Working Group.

The PO.DAAC will be responsible for data processing, archiving, and distribution for the EOS Altimeter mission, including all data products from level 0 through the Sensor Data Record and Geophysical Data Record. This responsibility will include the production of value added data products, and the archiving of ancillary data and algorithms required for reprocessing of data. The extent of the PO.DAAC role will be determined during the Phase A and Phase B studies, scheduled to begin in late FY-93. During this period, the respective roles of the PO.DAAC and NOAA will be determined relative to the mission data.

The PO.DAAC will be responsible for data processing, archiving, and distribution for the SeaWinds mission, including all data products from level 0 through the Sensor Data Record and Geophysical Data Record. This responsibility will include the production of value added data products, and the archiving of ancillary data and algorithms required for reprocessing of the data. The extent of the PO.DAAC role will be determined during the Phase A and Phase B studies, scheduled to begin in late FY-94. During this period, the role of the PO.DAAC will be determined relative to the mission data.

The PO.DAAC will continue to publish the TOGA CD-ROM series throughout the International TOGA period, ending in FY-96. The PO.DAAC will publish other data sets as recommended by the Science Working Teams of the missions which we support, and by the User Working Group, including the SSM/I Oceans Products in FY-94, the Altimetric CD-ROM in FY-95, the West Coast Time Series CD-ROM in FY-95, and additional data sets as identified.

James H. Corbin

Personal-

Name: James H. Corbin

Title: Director

Affiliation: Center for Air Sea Technology, Mississippi State University

Address: MSU - CAST, Bldg. 1103, Room 233, Stennis Space Center, MS 39529-5005

email: j.corbin (OMNET) corbin@cast.msstate.edu (INTERNET)

phone: (601)688-2561

fax: (601)688-7100

Data System-

Data System Name: Navy Environmental Observational Nowcast System (NEONS)

Discipline: Oceanography, Meteorology

Data Managed:

Type of Data: Operational model runs from FNOC
Inventory Meta Data [Y/N]: N (?)
Digital Data, Data Products [Y/N]: Y
Number of Data Granules: 150 per day; X/Y grid
Total Volume of Data [Megabytes]: 20MB per day; 21 day rotating archive

Type of Data: Model results from Data Assimilation and Model Evaluation Experiments (DAMEE) project
Inventory Meta Data [Y/N]: Y (?)
Digital Data, Data Products [Y/N]: Y
Number of Data Granules: (?) 10**5
Total Volume of Data [Megabytes]: (?) GB

Type of Data: Observational data for verification in DAMEE
Inventory Meta Data [Y/N]: Y (?)
Digital Data, Data Products [Y/N]: Y
Number of Data Granules: (?) 10**5; profiles, tracks, etc.

Total Volume of Data [Megabytes]: (?)

Data Management Activities Summary-

A major effort of CAST is development of distributed data-bases and data systems. For the last four years the CAST research staff has made major contributions to the enhancement and applications of the Naval Environmental Operational Nowcast System (NEONS) developed by the Naval Research Lab at Monterey California. NEONS is an interface to a relational database management system. It provides "C" and "FORTRAN" Applications Programming Interface (API) to the RDBMS, and also models typical oceanographic data types, namely grids, point observations in time, profiles, tracks, and images. Through CAST efforts, NEONS has been implemented at the U.S. Naval Oceanographic Office (NAVOCEANO), the Fleet Numerical Oceanography Center (FNOC), the National Climate Data Center, and the Naval Research Lab Stennis Space Center (NRL/SSC) among others. Through the support efforts to NRL/SSC, NAVOCEANO and FNOC, CAST will have near real-time access to meteorological and oceanographic observational and Navy operational model output data sets that could be included in the distributed system.

CAST is also in the final developmental phases of a "client-server" version of NEONS called "netNEONS" which obviates the need for running a local RDBMS. This allows local client applications to access the remote RDBMS transparently via the remote NEONS server. The experience gained with netNEONS, coupled with what CAST has learned in the development of a prototype Network Data Browser and associated applications could prove valuable input to this workshop. This prototype allows the scientists to browse and retrieve model output fields stored in UNIX files scattered across the network. The meta information on the contents of the files are registered in an RDBMS which the user queries via a GUI. This is a contents based browse system being developed for the NRL/SSC Ocean Sciences Group. The NRL scientists run many numerical models of differing versions. The Network Data Browser helps keep track and manage outputs from the different model runs for each of the different modeling groups and facilitates cross-group data browsing and retrieval. The final version is intended to be an end-to-end client-server system.

CAST has also developed and implemented an operational data-base for NAVOCEANO to manage and distribute the ocean profile data from the Master Oceanographic Observational Data Set (MOODS). This entailed migrating nearly two and a half million ocean profiles from files-based system on UNISYS to the EMPRESS Distributed RDBMS running on a Cray Y- MP and SUN front end. CAST is also teaming with University of Colorado on development of an Altimetry Data Processing and Analysis System (ideally this will be a distributed system).

Peter Cornillon

Personal-

Name: Peter Cornillon

Title: Professor of Oceanography

Affiliation: URI/GSO

Address: 112 Watkins, Narragansett Bay Campus, URI, Narragansett, RI 02882

email: pete@petes.gso.uri.edu

phone: (401)792-6283

fax: (401)792-6728

Data System-

Data System Name: xbrowse

Discipline: Physical Oceanography

Data System Name: Global AVHRR Database

Discipline: Physical Oceanography

Data System Name: InSitu

Discipline: Physical Oceanography

Data Managed:

Total Volume of Data [Megabytes]: Gigabytes

Data Management Activities Summary:

At The University of Rhode Island I am involved with three different data access systems. Each system has a different scope, but all are accessible using the Internet.

The Global AVHRR database locates existing HRPT and LAC passes both at a number of institutions around the world. The database is automatically updated using the Internet. Users can submit SQL queries to the database using a captured account. In addition, the database can be used to co-locate AVHRR and XBT data using a data server we developed and installed at NODC. The AVHRR database currently points to over 140,000 HRPT/LAC passes.

The Xbrowse system provides realtime access to our 20,000+ archive of 5km AVHRR data of the western North Atlantic. This client- server system provides the user with a simple interface to a flatfile database which can be searched by date only. In response to a query the user is presented with a list of image names, any one of which may be examined. Xbrowse uses progressive transmission to ameliorate different (often low) levels of available network bandwidth at users sites.

We have also developed a set of small in situ data servers which are accessed transparently using a captured account. This system provides access to several different types of in situ and model data. It uses a hierarchical searching mechanism specifically tailored to these data sets.

In addition, we have an extensive in house archive of satellite data, both processed and raw.

Henry A. Debaugh

Personal-

Name: Henry A. Debaugh

Title: Database Administrator for the Ocean and Lake Levels Division (OLLD)

Affiliation: NOAA

Address: NOAA/NOS, Routing code: N/OES2x1, Room 7209, 1305 East-West Highway, Silver Spring, MD 20910

email: Internet not available yet (We are hoping for November.) Use omnet, care of d.beaumariage

phone: (301)713-2884

fax: (301)713-4437

Data System-

Data system name: National Water Level Data Center (proposed name)

Discipline: Oceanography

Data Managed:

Type of Data: Water Level Time Series data

Inventory of Meta Data [Y/N]: Y

Digital Data, Data Products [Y/N]: Y

Number of Data Granules: ? (OLLD maintains a world wide network of water level observation stations app: 200 at present)

Total Volume of Data: app. 3000 megabytes

Type of Data: Ancillary Atmospheric and Oceanographic Times Series Data

Inventory of Meta Data [Y/N]: Y

Digital Data, Data Products [Y/N]: Y

Number of Data Granules: ? (OLLD maintains a world wide network of water level observation stations app: 200 at present)

Total Volume of Data: app. 1000 megabytes

Type of Data: Water Temperature & Density

Inventory of Meta Data [Y/N]: Y

Digital Data, Data Products [Y/N]: Y

Number of Data Granules: ? (OLLD maintains a world wide network of water level observation stations app. 90 collect Temperature & Density data)

Total Volume of Data: app. 200 megabytes

Data Management Activities Summary:

I am the database administrator for the Ocean and Lake Levels Division. I have a primary responsibility for the management of OLLD's data resources.

OVERVIEW OF OLLD'S NATIONAL WATER LEVEL DATA CENTER [2]

The Ocean and Lake Levels Division in the Office of Ocean and Earth Sciences, National Ocean Service, NOAA, is presently developing the Data Processing and Analysis Subsystem (DPAS), a critical component of the Next Generation Water Level Measurement System (NGWLMS). DPAS is a fully integrated, state-of-the-art computer system that will perform the following functions of the NGWLMS; data acquisition, data processing, analysis and quality control, database management, field requirements assessments, logistics control, administrative activities, and data dissemination. DPAS is scheduled to be completed early in 1994 and operational after thorough system testing has been successfully completed. Some parts of DPAS, such as data acquisition functions, have been operational since early in 1992. DPAS will provide on-line access to the Division's very large and valuable database for both in-house and external users. OLLD analysts and oceanographers will use DPAS software to derive standard Division products and also to perform ad-hoc analyses and perform specialized services.

The Division is responsible for the collection and subsequent processing and analysis of water level and related data from the coastal areas of the United States, including Alaska, Hawaii, and U.S. territories in the Pacific, and the Great Lakes. Through OLLD's participation in the Global Sea Level Program field stations have already been installed in several foreign nations and more installations are planned for the near future. NGWLMS field units, which are replacing existing water level gages based on antiquated technologies, automatically measure and record water level data and associated data quality assurance parameters, and other ancillary environmental data (such as wind speed, direction and gusts, barometric pressure, etc.). These data are transmitted from each remote site every three hours via GOES satellites to NESDIS satellite downlink facilities at Wallops Island, Virginia. DPAS automatically calls NESDIS every hour to download data collected since the last call and then automatically performs preliminary quality control checks and generates several reports which advise NOS personnel on the operational status of the entire system. Every 2 weeks (as presently scheduled) DPAS also automatically interrogates they field units directly and downloads data that was not collected via satellite transmission for whatever reason.

[2]proposed name

The nucleus of DPAS is Sybase, a high-performance relational database management system (DBMS) which manages and maintains OLLD's database which will exceed 10Gb of data in a few years. This database will store and maintain not only data from the new NGWLMS field units but much of OLLD's historical data as well. Sybase provides many other important features such as security and control, server enforced integrity, high data availability, and window-based tools. Sybase also runs on a variety of hardware platforms and operating systems making it highly portable and providing an easy migration path to more powerful platforms. Sybase is used in a client-server architecture. For DPAS development, a VAX 4000 Model 500 computer acts as a database server running Sybase server software; this processor will be upgraded when the system becomes fully operational to increase performance. Other VAXes are networked to the database server to provide various network services. Application software, consisting of integrated commercial and customized software, runs on client workstations which are 486-based PC's running the OS/2 operating system for in-house work. DPAS will have the flexibility to allow external client workstations with other hardware and software configurations to access the database server via wide-area network (WAN) and download selected data to be used as needed.

When completed, DPAS will automatically perform many functions of OLLD that are now manual processes. Routine data processing and quality assurance tasks will be done autonomously. DPAS will provide more capabilities to both in-house and external users and allow ad-hoc analyses to be accomplished with relative ease. The extensive data archive of OLLD will be directly and readily available to external users and data from NGWLMS field units will become available in near-real time.

Elaine Dobinson

Personal-

Name: Elaine Dobinson

Title: PO.DAAC Deputy Task Manager

Affiliation: JPL

Address: 4800 Oak Grove Drive, Pasadena, A 91109

email: elaine_dobinson@isd.jpl.nasa.gov

phone: (818)306-6269

fax: (818)306-6929

Data System-

Data System Name: EOSDIS Physical Oceanography DAAC

Discipline: Computer Science

Data Managed:

Type of Data: Physical Oceanographic Data Products
Inventory Meta Data [Y/N]: Y
Digital Data, Data Products [Y/N]: Y
Number of Data Granules: Approx. 30,000 Granules
Total Volume of Data [Megabytes]: Approx. 125 Gigabytes

Data Management Activities Summary-

I am currently working on the Version 0 implementation of EOSDIS. I led the IMS work at the PO.DAAC, co-lead the IMS system-level data dictionary development activities, and am currently deputy manager for the DAAC. The PO.DAAC archive is an upgrade of the NODS (NASA Oceans Data System) implemented in INGRES and about to be ported over to UNIX on SGI.

Prior to my working on EOSDIS I was the lead database designer of the science data catalog for the Planetary Data System. I am also the supervisor of the Archive Data Management Group in the Science Data Systems Section of JPL.

Glenn R. Flierl

Personal-

Name: Glenn R. Flierl

Title: Professor of Physical Oceanography

Affiliation: MIT

Address: 54-1426, MIT, Cambridge, MA 02139

email: glenn@lake.mit.edu

phone: (617) 344-2728

fax:

Data System-

Data system name: JGOFS

Discipline: Bio, Chem, Phys Oceanogr.

Data Managed:

Type of Data: station, time-series, model ...
Inventory Meta Data [Y/N]: y
Digital Data, Data Products [Y/N]: y
Number of Data Granules:
Total volume of Data [Megabytes]: growing, but not huge

Data Management Activities Summary-

A Distributed, Object-based Data Management System for JGOFS

Glenn Flierl, James Bishop, David Glover, Satish Paranjpe

Large oceanographic programs such as JGOFS (The Joint Global Ocean Flux Study) require data management systems which enable the exchange and synthesis of extremely diverse and widely spread data sets. We have developed a distributed, object-based data management system for multidisciplinary, multi-institutional programs. It provides the capability for all JGOFS scientists to work with the data without regard for the storage format or for the actual location where the data resides. The approach used yields a powerful and extensible system (in the sense that data manipulation operations are not predefined) for managing and working with data from large scale, on-going field experiments.

In the "object-based" system, user programs obtain data by communicating with a program (the "method") which can interpret the particular data base. Since the communication protocol is standard and can be passed over a network, user programs can obtain data from any data object anywhere in the system. Data base operations and data transformations are handled by methods which read from one or more data objects, process that information, and write to the user program (or to the next filter in the series).

We have written methods for various ASCII and binary databases, and built transformation routines for doing mathematical operations, dynamic height calculations, and data joins.

George M. Frank

Personal-

Name: George M. Frank

Title: National Geodetic Survey Database Administrator

Affiliation: DOC/NOAA/NOS/C&GS/NGS

Address: 1315 East West Highway - Station 9127 Silver Spring, Md. 20910-3282

Email: george@galaxy.ngs.noaa.gov

Phone: 301-713-3251

Fax: 301-713-4172

Data System-

Data System Name: National Geodetic Survey Integrated Data Base (NGSIDB)

Discipline: Geodesy

Data Managed:

> Type of Data: Geodetic Survey Data
>
> Inventory Meta Data: Yes
>
> Digital Data, Data Products: Yes
>
> Number of Data Granules:
>
> Total Volume of Data (MB): 6000

Data Management Activities Summary:

The National Geodetic Survey is a Federal organization that was established in 1807. It was given the responsibility to map the coastline of the United States so that the nation's shipping industry could safely navigate its waters. This mandate has resulted in the determination of horizontal coordinates of latitude and longitude for over 300K points and the vertical coordinates for approximately 1000K points. This information is the basis for the National Geodetic Reference System from which all U.S. mapping efforts should begin.

The NGS data holdings consists of all the geodetic surveying information that NGS has accumulated from private, local and federal government and its own efforts during the period since 1807. The data types consist of coordinates of latitude, longitude, and elevation; descriptive text describing the location and physical characteristics of the points; and observational data such as gravity, directions, distances, elevation differences, and satellite observations including doppler, very long baseline interferometry (VLBI), and global positioning system (GPS).

The NGS data is stored and managed by a relational database machine, a Britton LEE IDM 700, otherwise known as a Sharebase or Teradata. This database server is accessed through a local area network of Unix workstations and PCs using the Structured Query Language (SQL). SQL is used in either an interactive mode or in Fortran or C application programs. The NGS LAN is a TCP/IP twisted pair ethernet consisting of over 100 PCs and 35 UNIX workstations such as Suns, HPs, and Sun clones. The NGS LAN can be accessed with dial-in capability or through Internet.

For those outside users that have not been granted direct access to the NGS LAN or database system, information can be obtained by contacting the NGS Information Center by telephone. The Information Center provides a wide variety of data in various formats. GPS orbital data which is not in the database can be obtained through the Information Center or through the Coast Guard bulletin board.

The NGS database system is currently undergoing a transition to a Sybase DBMS system that resides on a Sun multiprocessor server. Sybase possesses features such as ANSI SQL and Open Server, that will permit NGS to more easily share its data with other database systems. The multiprocessing capability of this system will also increase the performance and therefore the availability of the NGS data.

NGS and another organization within NOAA, the Ocean Lake and Level Division (OLLD),is proposing to create a distributed data system. NGS and OLLD independently maintain their own data holdings using Sybase DBMS. Each realizes that the other possesses data that they could use. It is their goal to create a system that would provide a link between them. This would require the creation of metadata for their databases, a network accessible graphical user interface (GUI) that would allow users to locate and obtain data easily and rapidly, and the physical link that could be provided through Internet or dial-in access.

David Fulker

Personal-

Name: David Fulker

Title: Director, Unidata Program Center

Affiliation: University Corporation for Atmospheric Research (UCAR)

Address: P.O. Box 3000, Boulder, CO 80307 or for UPS, etc: 3300 Mitchell Lane, Suite 170, Boulder, CO 80301

email: fulker@unidata.ucar.edu

phone: (303)497-8650

fax: (3)497-8690

Data System-

Data System Name: Unidata

Discipline: Atmospheric Science

Data Managed:

Type of Data: Surface, Soundings, Grids, Images

Inventory Meta Data [Y/N]: N

Digital Data, Data Products [Y/N]: Y

Number of Data Granules: NA (real-time flow)

Total Volume of Data [Megabytes]: ¿100 MB/day

Data Management Activities Summary-

* UNIDATA Overview Factsheet

Unidata is a nationwide program to help university departments acquire and use atmospheric data. The Unidata Program Center (UPC) is managed by the University Corporation for Atmospheric Research (UCAR) in Boulder, Colorado, and sponsored by the National Science Foundation (NSF). Personnel and other resources required for university participation are provided by the institutions themselves. The NSF and university resources are complemented by contributions from private industry.

Unidata Provides:

- Real-time weather data (via satellite) at group discount rates;
- Software to display and analyze those data;
- Consultation on the necessary hardware and software;
- Training workshops on how to install and use Unidata software;
- Ongoing support by and for the Unidata community of users.

Products and Services

- Current Data via Satellite:

  Established Data Products. Unidata participants receive National Weather Service data at discounted rates. These include the Domestic Data Service (conventional surface and upper-air observations for the U.S.), the International Data Service, and the Numerical Product Service (gridded analyses and forecasts from the National Meteorological Center and the European Centre fro Medium-Range Weather Forecasting).

  Research Data Products. A special Unidata broadcast channel carries a range of data prepared at the University of Wisconsin-Madison under contract with the UPC. These include satellite and radar images, as well as the conventional meteorological data used with the Unidata McIDAS software/ This channel also carries special products not generally available through the National Weather Service, such as wind profiler data.

- Software Tools:

  netCDF [described below]

  The Local Data Manager (LDM) [described below]

To licensed universities, Unidata distributes software applications for displaying and analyzing the data captured by the LDM. These packages are WXP, the Weather Processor (developed by Purdue University), GEMPAK (developed at NASA/Goddard Space Flight Center), and YNOT (developed by MacDonald Dettwiler and Associates). Unidata also distributes McIDAS-X and McIDAS-OS2 (developed at the University of Wisconsin-Madison Space Science and Engineering Center), which analyze and display data from the special Unidata/Wisconsin broadcast channel.

Support:

The Unidata Program Center provides full support for all the software packages it distributes. Full support includes: consultation, training workshops, software maintenance, and documentation.
Costs:

Unidata software, training, and support are provided at no charge to universities. Unidata arranges a group discount rate for data services. Universities assume the costs of purchasing equipment, hiring site and system administrators, subscribing to data services, and all travel and accommodations associated with training workshops.

Some of the Unidata activities most relevant to a distributed data systems workshop are associated with two software systems, netCDF and LDM, described below.

* UNIDATA netCDF Factsheet

Overview:

The Network Common Data Form, or netCDF, package is a software package that standardizes how scientific data are stored and retrieved. More than a data format, the netCDF package is a set of programming interfaces that can be used with widely varying scientific data sets by machines of widely varying architecture.

Features

- Standardized Data Access

  Unidata's library of netCDF subroutines insulates applications from the underlying data format. Multidimensional floating- point, integer, and character data can be stored and retrieved using these functions. Data are accessed by specifying a netCDF file, a variable name, and a description of what part of the multidimensional data is to be accessed. Multidimensional data may be accessed one point at a time, in cross- sections, or all at once.

- Self-Describing

  Variables and their associated dimensions are named. Information about the data, such as what units are used what is the valid range of data values, can be stored in attributes associated with each variable. The processing history of a data set can be stored with the data.

- C and FORTRAN Compatible

  The netCDF subroutines can be invoked from either C or FORTRAN, and data stored using one language may be retrieved in the other.

- Machine-Independent

  The format underlying the netCDF package employs an open standard known as XDR (for eXternal Data Representation) that renders netCDF files machine independent. The netCDF package is particularly useful at sites with a mix of computers connected by a network. Data stored on one computer may be read directly from another without explicit conversion.

- Portable

  The software has been used successfully on a broad range of computers, from PCs to supercomputers.

- Benefits
  - Reusable Applications
    Unidata's purpose in creating the netCDF library is to generalize access to scientific data so that the methods used for storing and accessing data are independent of the computer architecture and the applications being used. In addition, the library minimizes the fraction of development effort devoted to dealing with data formats.
  - Reusable Data
    Standardized data access facilitates the sharing of data. Since the netCDF package is quite general, a wide variety of analysis and display applications can use it. The netCDF library is suitable, for example, for use with satellite images, surface observations, upper-air soundings, and grids. By using the netCDF package, researchers in one academic discipline can access and use data generated in another discipline.

* UNIDATA LDM Factsheet

Overview

Unidata's Local Data Manager (LDM) software acquires meteorological data and shares these data with other computers on a network. The LDM handles data from National Weather Service data streams, including gridded data from numerical forecast models.

A client ingester handles a specific data feed. It scans the data stream, determines product boundaries, and extracts products, passing selected products to one or more LDM servers.

The LDM server processes the raw data passed to it by one or more ingesters and converts that data into a form that can be used by applications programs.

Features – The LDM is

User configurable: the LDM server can be instructed to append a particular product to a file; save a product in a particular form, such as Unidata's netCDF; execute an arbitrary program with the data product as input; store or retrieve products in a simple database by key; and/or pass data along to other running client programs.

Site configurable: data captured on one machine can be stored on other machines on a network. This means that data ingest functions can be separated from storage and use functions, allowing sites to tailor their LDM system to their capacity.

Extensible: new client decoders can be added easily, including decoders for archival data.

Event-driven: the system captures data in real time.

Unidata is currently engaged in exploiting the architecture of the LDM software to build a system for distributing real-time data via the Internet. The principle is that products will fan out from the source through several tiers of cooperating LDM computers, each of which relays data to several "neighbors" on an event-driven basis. In this way, hundreds of end-user sites can be served promptly (i.e., within a few seconds of data arrival) without the kinds of traffic jams that would arise if all sites were to contact a single server at the same time (i.e., at the time of data arrival).

The distributed LDM system is undergoing tests at a dozen or so sites, and the results are sufficiently encouraging that we are planning eventually to replace the satellite data broadcast service with this Internet Data Distribution (IDD) system.

James Gallagher

Personal-

Name: James Gallagher

Title: Programmer/Analyst

Affiliation: URI/GSO

Address: 110 Watkins, Narragansett bay campus, URI, Narragansett, RI. 02882

email: jimg@dccz.gso.uri.edu

phone: 401.792.6939

fax: 401.792.6728

Data System-

Data system name: xbrowse

Discipline: Physical oceanography

Data Managed:

Type of Data: Sea surface temperaure (SST)
Inventory Meta Data [Y/N]: Y
Digital Data, Data Products [Y/N]: Y
Number of Data Granules: O(10k)
Total volume of Data [Megabytes]: O(10)

Type of Data: SST, Raw satellite
Inventory Meta Data [Y/N]: Y
Digital Data, Data Products [Y/N]: Y
Number of Data Granules: O(10)
Total volume of Data [Megabytes]: O(1)

Data Management Activities Summary-

Xbrowse was designed so that users who are not physically at an archive site can efficiently access, review and retrieve image and in-situ data necessary for their work. Xbrowse uses the Internet to provide real- time access to image archives. Xbrowse is different from other remote browsing systems in that it provides more than a fixed resolution preview of each image being browsed. With xbrowse, the user views the image at progressively increasing levels of resolution – up to the resolution of the archived data if desired. Images that are of no interest to the user, for example because of cloud cover in a region of specific interest or because of data drop out in part of the image, can quickly be passed by at low resolution. Images of greater interest can be allowed to progress to higher levels of resolution. Further, the user can stop the progressive transmission of the image at a (low) level of resolution and mark out an inset area in the main image for viewing at higher levels of resolution. Since researchers often need only a portion of the image area at full resolution, this progressive transmission/composite image approach to browsing makes it feasible to provide effective real-time, remote access to the archived data and allows xbrowse to be used as a data access tool and not just a data ordering tool. In addition, xbrowse can access in situ metadata and overlay that data on the imagery being displayed.

The xbrowse software is based on the client/server model of cooperating, independent processes. Client software, essentially the user interface, is installed on a user's machine. Server software at archive sites responds to commands from the client software to send the data requested by the user. The client software on the user's machine can interact with many different servers at different sites. Although at present only the two data servers mentioned earlier are operational – the server at the University of Rhode Island which provides AVHRR SST images and a server at the National Ocean Data Center which provides XBT metadata – it is expected that other data servers will be added at other archives in the future.

Both the client and data server are available via anonymous ftp at zeno.gso.uri.edu.

Jeffrey Given

Personal-

Name: Jeffrey Given

Title: Geophysicist

Affiliation: Science Applications International Corporation

Address: MS A-2 10260 Campus Point Drive, San Diego, CA 92121

email: jeff@gso.saic.com

phone: (619)458-2656

fax: (619)458-4993

Data Management Activities Summary-

Our group at Science Applications International Corporation has approximately 40 professionals involved with all aspects of the management, processing, and wide-area distribution of geophysical data. The representative data management activities include:

1. The Center for Seismic Studies (CSS), Arlington, VA. This is a seismological data center that we have developed for ARPA over the past decade. The activities of the CSS include (1) Near real-time, world-wide data acquisition over a TCP/IP WAN; (2) Data processing and analysis by independent, cooperating institutions in Norway, Russia, and the US, with automated data migration between the sites; (3) data-centric software architecture based on a distributed Oracle DBMS and supported by several application interfaces; (4) many gigabytes of randomly and readily accessible time-series data stored on an optical jukebox and tightly integrated into a near-real time processing system; (5) User interfaces built with X/Motif for data selection, browsing, and interpretation; (6) Distribution of these data to the world-wide community of seismologists.

2. Sequoia 2000. SAIC has been an industrial partner for two years in this state-wide project of the University of California. We are deeply involved in developing DBMS and application software to support the project oceanographers and climate researchers involved in global change science. We are working especially closely with NOAA staff and NOAA- supported researchers at SIO.

3. NOAA. Under contract to Scripps Institution of Oceanography, SAIC is prototyping a distributed data access system.

4. Acoustic Thermometry of Ocean Climate (ATOC). Under contract to Scripps Institution of Oceanography, SAIC is developing a data center for the ARPA-sponsored ATOC project. The data to be managed includes large volumes of real-time acoustic data collected from distant, distributed, data acquisition points, and numerous other oceanographic and atmospheric data sets. A fundamental project requirement is that access to these data be available to a global community of users for acoustic analysis and global-change studies.

David M. Glover

Personal-

Name: David M. Glover

Title: Research Specialist

Affiliation: Woods Hole Oceanographic Inst.

Address: Dept. of MC&G, WHOI, Woods Hole, MA 02543

email: david@plaid.whoi.edu

phone: (508) 457-2000

Data System-

Data System Name: JGOFS DBMS

Discipline: Biological, Chemical, and Physical Oceanography

Data Managed:

Type of Data: station data and time series data
Inventory Meta Data [Y/N]: y
Digital Data, Data Products [Y/N]: y
Number of Data Granules:
Total Volume of Data [MB]: approx. 50MB and growing

Data Management Activities Summary-

A Distributed, Object-based Data Management System for JGOFS

Glenn Flierl, PI, James Bishop, David Glover, Satish Paranjpe

We have been involved in developing a distributed, object-based, multiple client, multiple server front-end for a relational DBMS. This system provides to the user (the scientists in the JGOFS project) a format/location transparent means of access to the continually growing JGOFS database. This access allows the synthesis of fairly diverse data set types into new science-driven products without the user worrying about where the data is located or in what format it is stored. This is achieved by using "methods" that are executable bits of code that know about the data set's particulars and the requests are directed by a "server" that acts as a telephone operator knowing where the data is located. Since none of the data manipulation operations are predefined the system is flexible and extensible.

I also sit on the EOSDIS Advisory Panel (aka the Data Panel) for the EOS Investigator Working Group (IWG) and chair the Users Working Group (UWG) for the JPL Physical Oceanography DAAC. The Data Panel has been instrumental in the crafting of the EOSDIS requirements that went into the RFP that was consequently won by Hughes. Now the Data Panel "stands guard", as it were, to insure that the distributed, evolving DIS we recommended actually comes about. The Users Working Group (UWG) advises the physical oceanography DAAC at JPL.

Steve Hankin

Personal-

Name: Steve Hankin

Title: Computer Scientist

Affiliation: NOAA/Pacific Marine Environmental Laboratory

Address: 7600 Sand Point Way NE, Seattle WA, 98115

email: hankin@pmel.noaa.gov

phone: (206)526-6080

fax: (206)526-6744

Data System-

Data system name: FERRET and TMAP Data Base

Discipline: Oceanography

Data Managed:

Type of Data: gridded data products
Inventory Meta Data [Y/N]: Y (exists)
Digital Data, Data Products [Y/N]: Y
Number of Data Granules: 250 (files)
Total volume of Data [Megabytes]: 5000 Mbytes

Type of Data: gridded model outputs
Inventory Meta Data [Y/N]: Y (exists)
Digital Data, Data Products [Y/N]: Y
Number of Data Granules: 3000 (files)
Total volume of Data [Megabytes]: 250,000 Mbytes

Data Management Activities Summary-

The Thermal Modeling and Analysis Project (TMAP) at NOAA/PMEL in Seattle is a numerical ocean modeling and data analysis group emphasizing upper ocean processes and ocean climate physics. Typically TMAP's model experiments are performed on Cray and Cyber supercomputers and produce mid-size output data sets (one to five gigabytes). TMAP also performs observational data studies to produce data sets suitable for forcing and validating the models. These analyses are carried out on networked workstations. The program FERRET was developed by TMAP as the primary software tool for analysis.

FERRET is a workstation-based, interactive visualization and analysis environment that permits users to explore large and complex gridded data sets. 'Gridded data sets' in the FERRET framework include climatological data products (e.g. Levitus climatologies), binned monthly observation summaries (e.g. COADS), operational model outputs (e.g. NMC & ECMWF), diagnostic model outputs (e.g. GFDL MOM), suitably prepared 'section' data, and time series and vertical profiles (viewed as singly dimensioned grids). FERRET's gridded data sets can be one to four dimensions - usually (but not necessarily) longitude, latitude, depth, and time - where the coordinates along each axis may be regular or irregularly spaced. A data set may contain mixed 1, 2, 3, and 4-dimensional variables. Axes of the same orientation may also differ - models often require staggered grids and gridded data products have few standards for temporal or spacial resolution.

FERRET was designed to function in a distributed data environment. Since gridded data sets are frequently multi-gigabyte in size the program contains logic to manage a data set as a network-wide, distributed collection of files. Furthermore, FERRET communicates with its data base using a two phase approach that is suited to optimized, wide-area access. In phase one FERRET queries the meta-data to determine the description and range of the available variables. In phase two, after a user has specified a calculation, FERRET issues requests for only the minimally required subset of the data needed for the calculation, optimizing with respect to parameters that control access speed. Memory caching is used to minimize the network bandwidth requirements.

FERRET is an excellent environment for browsing gridded data sets. Graphical displays of data sections (and extractions of these sections to files) may be created with single commands. Graphical displays are automatically labelled with complete and unambiguous documentation. Additional variables may be overlaid - again with full documentation provided automatically. A (prototype) point and click graphical user interface (GUI) has been developed to make browsing simple for novice users. The development of this GUI is expected to continue under a project funded by the NOAA ESDIM program to provide Internet-wide gridded data browsing and analysis services from the NOAA Seattle campus.

FERRET development has emphasized interoperability standards. FERRET data sets are normally stored in the Unidata netCDF format, a self-documenting, publicly available format supported by the meteorology community. FERRET graphics are layered upon the International Standards Organization's Graphical Kernel System (ISO/GKS) which provides network compatibility through the X-windows standard, and also supports a wide collection of other protocols: PostScript, HPGL, CGM, Tektronix, and Versatek to name a few. FERRET is written in transportable FORTRAN 77 and ANSI C. It has been ported to SUN, DEC/Ultrix, SGI, VAX, and Macintosh (beta version) with IBM RS6000 and DEC/Alpha versions planned for the near future.

FERRET offers a flexible environment for data analysis; new variables may be defined interactively as mathematical transformations of variables from data sets. Complex analyses may be defined through hierarchical variable definitions. A symmetric, 4-dimensional command syntax is used to designate arbitrary rectangular regions in 4-space and "IF-THEN-ELSE" logic permits calculations to be applied over arbitrarily shaped regions. An assortment of data smoothers and data gap fillers compliments the normal range of mathematical analysis tools. FERRET's scripting language facilitates large, batch-style calculations and enables FERRET to interact with specialized packages such as Matlab or GMT and to use custom-written routines.

The TMAP group maintains a large and growing data base of gridded data products - approximately 5 Gigabytes in size presently. These data are stored in random access format for quick, record-level access by FERRET and are maintained on-line. Much of this data base is expected shortly to be available over Internet through FERRET.

A partial list of the data sets includes:

- COADS monthly average surface marine observations (1946-1991)
- Esbensen & Kushnir global ocean surface heat budget
- ECMWF/TOGA global 12-hourly surface analysis (1985-92)
- FSU tropical Pacific wind stress (1961-92)
- NODC Levitus climatological global ocean atlas
- NGDC ETOPO 5 minute relief of the surface of the earth
- NMC blended monthly average SST (1982-1992)
- NMC monthly upper air winds and OLR analysis (1968-88)
- Oberhuber atlas of heat, buoyancy and turbulent kinetic energy
- Rasmusson and Carpenter tropical Pacific El Nino composite analysis
- Richardson monthly climatological global ocean surface currents
- Sadler tropical Pacific winds (1979-90)
- US Navy FNOC global 6-hourly surface winds (1982-92)

A partial list of the sites using FERRET includes:

- NCAR
- MIT

- Los Alamos National Laboratories
- NOAA/Pacific Marine Environmental Laboratory
- NOAA/Alaska Fisheries Science Center
- NOAA/Geophysical Fluid Dynamics Laboratory
- NOAA/Atlantic Ocean Marine Laboratory
- University of Washington (6 departments)
- University of Hawaii
- Naval Post-graduate School
- Florida State University
- University of Rhode Island
- University of British Columbia
- Texas A and M

FERRET is freely available over the Internet via anonymous FTP on node abyss.pmel.noaa.gov.

Roen Hogg

Personal-

Name: Roen Hogg

Title: Faculty Research Assistant

Affiliation: Oregon State University

Address: College of Oceanography Oceanography Administration – Building 104 Corvallis, OR 97331-5503

email: roen@oce.orst.edu

phone: 503-737-4414

fax: 503-737-2064

Data Management Activities Summary-

The objective of our project is to use object-oriented technology to develop a user-friendly system that will allow scientists to interact with oceanic data and test hypotheses at the workstation. In addition, this system will facilitate multidisciplinary ocean field experiments by allowing scientists to communicate the results of their research to internal and external user groups.

The proposed system will provide intuitive and relatively transparent access to existing analysis systems, numerical models, imagery data, data acquisition systems, heterogeneous databases, and communication systems. Given the large disk storage and computing capability (between 2-3 Giga flops) required by existing 3-D modeling and analytical tasks, the system will provide access to the Oregon State University Oceanography computing facilities (SUN Sparc and IBM UNIX- based workstations, massively parallel CM-5 Connection Machine, and 100GB optical data storage capabilities).

Some of the major features associated with the system include the following:

1. Derivation Analysis

   The system will process requests for the selection and formatting of relevant data. In particular, the system will process requests to transform:

   (1) level 1 data (sensor data) into level 2 data (geophysical data)

   (2) level 2 data (geophysical data) into level 3 data (higher-order data)

2. Research Analysis

   The system will support the research analysis common to most research projects. This includes the following:

(1) useful numerical analysis and display of data sets

(2) useful data queries

In addition, the system will store a description of how each data set was derived.

3. Visualization

   The system will support graphic and video images. These images will include plots (e.g., profile, time series, drifter, imagery, grids, station), satellite images, and video. The system will provide the following functionality:

   (1) interface with the appropriate graphic/video generating programs

   (2) store a description of how each image was derived

   (3) store any free-form text a scientist wants to associate with a particular image

   (4) provide useful query capabilities

4. Communication

   The system will provide computer networking and transmission of data and results to internal and external user groups. The system will support the following types of communication:

   (1) e-mail

   (2) video

   (3) audio

   (4) data (files)

5. Ocean Model

   The system will interact with various existing models by providing specifications of input and output parameters.

James D. Irish

Personal-

Name: James D. Irish

Title: Research Specialist

Affiliation: Woods Hole Oceanographic Institution

Address: 307 Smith

email: jirish@whoi.edu

phone: (508)457-2000 Ext. 2732

fax: (508)457-2195

Data System-

Data system name: Not using a formal data system at present, but starting to formulate one for use in 1 year.

Discipline: Physical Oceanography

Data Managed: Water Velocity, Temperature, Pressure, Conductivity, Optical and Acoustical Time Series

Data Management Activities Summary-

As a student at Scripps I worked with Walter Munk and his BOMM system of data analysis and archiving. I carried many ideas (and much code) to APL/UW where I constructed a similar system to analyze and store data using the computer systems there. At UNH Wendell Brown and I merged our ideas and the SIO routines to an Ocean Analysis Software Package with archiving capability, but the system had no on-line search capability. A search and retrieval system was started, but not completed before I left UNH. At UNH I also developed moorings systems which telemetered data via ARGOS and GOES satellites, and packet radio back to the laboratory in real-time. Since I have been at WHOI, I have also used cellular phone techniques for returning data from the field. A PC and mainframe based system was developed at UNH for retrieving the near real-time data daily from the field, editing, normalizing, and storing it in ASCII and binary data files for analysis and archiving. However, we were never able to get funding to set up a real-time data base for others to access. As part of my funded GLOBEC activities on Georges Bank, I am again deploying moorings with GOES and ARGOS telemetry (and possibly acoustic telemetry from bottom instruments) for several years, and plan to work with the Georges Bank Data Management Office at WHOI (Wiebe, Flierl, Brown, and Lynch) in establishing a real-time data retrieval, processing and distribution system on my computers for use within the Georges Bank GLOBEC data system using the ideas and concepts developed and tested by this group.

George Milkowski

Personal-

Name: George Milkowski

Title: Data Systems Manager/Developer

Affiliation: University of Rhode Island

Address: Graduate School of Oceanography Narragansett Bay Campus Narragansett, RI 02882-1197

email: george@zeno.gso.uri.edu OMNETg.milkowski

phone: 401 792 6939

fax: 401 792 6728

Data System-

Data system name: Global 1km AVHRR Inventory

Discipline: Physical Oceanography

Data Managed:

Type of Data: High Resolution Raw/Level 1b AVHRR

Inventory Meta Data [Y/N]: Y

Digital Data, Data Products [Y/N]: Some

Number of Data Granules: >200,000

Total volume of Data [Megabytes]: 700 Gigabytes

Data Management Activities Summary-

The Global 1km AVHRR Inventory was developed to provide, within a single inventory, a comprehensive listing of available high resolution AVHRR data from major archives around the world. The inventory includes listings of Level 1b data holdings from U.S. and foreign archives. Those archives with holdings listed are; NOAA/NESDIS, USGS EROS Data Center, University of Rhode Island, University of Miami, European Space Agency (ESA's inventory is composed of AVHRR data collected and archived by contributing European country agencies and their data centers), Australian CSIRO. The Japanese Weather Service is currently working on providing listings of their holdings. An on-line, simple to use, user interface provide access to the Global 1km Inventory and permits user specified searches based on time, geographic location, sensor, archive, platform, sensor data collection mode and data processing level. The inventory is held with a relational data base management system. Information on how to access and use the inventory can be obtained through anonymous ftp at zeno.gso.uri.edu in /usr/spool/ftp/pub/. Access to the inventory is also possible through the NASA Climate and Global Change Master Directory.

Contributing archives periodically transfer listings of their new acquisitions to the Global 1km AVHRR Inventory. These updates are coded in the internationally recognized format developed by the Committee on Earth Satellites Work Group on Data Catalog Subgroup specifically for AVHRR inventory update exchange.

Other scientific data management activities include the development and management of an XBT client-server application with NOAA/NODC that provides the location of XBT drops based on user supplied spatial and temporal windows. This application was designed such that the client application can either stand alone or be incorporated within another application. An example of this latter implementation is URI's XBROWSE where the XBT client-server application dynamically displays the location of XBT data correlated with sea surface temperature imagery.

Christopher Miller

Personal-

Name: Christopher Miller

Title: Physical Scientist

Affiliation: NOAA/NESDIS/Environmental Information Services

Address: 1825 Connecticut Ave., NW, Suite 506, Washington, D.C. 20235

email: C.Miller.NOAA (Omnet)

phone: (202)606-5012

fax: (202)606-0509

Data Management Activities Summary-

The responsibilities of the Environmental Information Services Office encompass the activities of the NOAA National Data Centers (National Climatic Data Center, National Oceanographic Data Center and the National Geophysical Data Center) and the Environmental Services Data and Information Management (ESDIM) Program, which is a cross-cutting program addressing NOAA-wide data management issues. ESDIM's current focus is rescuing critical NOAA data at risk of being lost and improving access to NOAA data and information. The head of Environmental Information Services, Mr. Gregory Withee, is, also, responsible for the Information Management element of the NOAA Climate & Global Change Program. Information Management serves the Climate & Global Change Program through its support of the needs of the multiple science elements of that program (generation of long-term climate and global change data sets; development of information management systems in connection with specific science program objectives). Information Management is, also, a focal point for the Global Change Data and Information System (GCDIS), which is envisioned as an interagency gateway for access and delivery of data and information products. A key goal is to promote interoperability among the existing heterogeneous systems.

William Schramm

Personal-

Name: William Schramm

Title: Chief, Ocean Applications Branch

Affiliation: NOAA

Address: 2560 Garden Road, Monterey, CA 93940

email: OMNET/W.SCHRAMM

Phone: (408) 647-4206

Fax: (408) 647-4225

Data System-

Data system name: NEONS database management system

Discipline: Synoptic Oceanography and Meteorology

Data Managed:

Type of data: Numerical ocean and atmospheric products (gridded fields from Navy and NOAA sources)

Inventory Meta Data? yes

Digital data, data products? yes

Number of data granules: 7025 fields

Total volume of data: 84 MB

Note: 1035 new fields (12.5 MB) are entered into the DBMS/day. Some are retained for two days and some for thirty days.

Type of data: SSM/I (SDRs, TDRs and EDRs)

Inventory Meta Data? yes

Digital data, data products? yes

Number of data granules: 14 satellite passes/day

Total volume of data: 150 MB

Note: This data changes daily.

Type of data: Synoptic ocean and atmospheric observations

Inventory Meta Data? yes

Digital data, data products? yes

Number of data granules: 16,000 observations

Total volume of data: 1.8 MB

Note: Observations are retained for two days.

Type of data: DMSP IR and visual images

Inventory Meta Data? yes

Digital data, data products? yes

Number of data granules: 79 images

Total volume of data: 1.6 MB

Note: Images are retained for two days in compressed format.

Data Management Activities Summary-

BACKGROUND

The Naval Research Laboratory (NRL) in Monterey CA has developed a powerful database management system for environmental data called the Naval Environmental Operational Nowcasting System (NEONS). The system was developed to manage the three basic types of environmental data; observations, images and gridded data. NRL uses NEONS to support research and development programs such as the development of satellite data processing software. A typical R&D application of the system at NRL is to develop "virtual sensor" information based on combinations of satellite data that do not exist from any single satellite. It is the design of NEONS, however, and not the NRL applications that makes the system of interest to others including NOAA facilities.

Unlike most data management initiatives, which are developed for a specific application, NEONS was developed to be very flexible and versatile. As a result other institutions have found the system to be useful and because of the Navy's Technology Transfer program NRL has been very cooperative in making the software available. Recently the Navy announced that NEONS will be used for operational database management on Cray supercomputers at the Fleet Numerical Oceanography Center (FNOC) and the Naval Oceanographic Office at Stennis, MS. Within NOAA the NEONS software was first installed at the Ocean Applications Branch (OAB) of the National Ocean Service. OAB was established in Monterey to support Navy/NOAA cooperative programs and NEONS is being used to support civilian distribution of FNOC data, analyses and forecasts via the Navy/NOAA Oceanographic Data Distribution System (NODDS). OAB later contacted other NOAA offices to make them aware of the availability of NEONS and to promote sharing of software and data resources between Navy and NOAA. In late 1991 OAB arranged for the installation of NEONS at the National Climatic Data Center (NCDC) in Asheville NC for support of the Global Climate Perspectives System (GCPS). Other participants in GCPS are the NOAA Climate Monitoring and Diagnostics Lab (CMDL) in Boulder CO and the Climate Analysis Center (CAC) in Washington DC. In February 1992 OAB helped install NEONS at CMDL and in August 1992 at CAC. In December of 1992 there was a second installation in Boulder at the Forecast Systems Laboratory where NEONS will be used in the MADER project. The most recent NEONS installation in NOAA was at the NMFS Laboratory in Hawaii. In addition to installations in the Navy and in NOAA, the system has been provided to; 1) Canadian Atmospheric and Environmental Services in Toronto and Vancouver, 2) Bureau of Meteorology in Melbourne, Australia, 3) South Dakota School of Mines and Technology, 4) Woods Hole Oceanographic Institution, 5) Cray Research Inc, 6) British Meteorological Office, 7) French Meteorological service and 8) World Laboratory in Italy.

NEONS TECHNICAL DESIGN

NRL designed NEONS for fast, efficient operation and compatibility with computer industry and international data exchange standards such as BUFR and GRIB. The system is built around the commercial database management system, EMPRESS, and operates on a variety of computers, from UNIX workstations to Cray supercomputers. Computer industry standards used in the design of NEONS include UNIX and SQL.

An important feature of NEONS is the storage of data in variable length binary strings. This is important because environmental data comes in a variety of record lengths. Another important advantage of the way in which NEONS stores data is that the data are addressed only to a minimal level of information in contrast to many other database systems which address data deeply, down to the report or even data value level. The approach used by NEONS greatly speeds up searches compared to other systems that are often burdened with high system overhead and frequent disk accesses. The third advantage of the NEONS approach is that by using binary compaction the system takes advantage of the great CPU speed of new RISC computers while at the same time minimizing the I/O time which is the critical limiting factor in modern DBMS systems.

The international data exchange standards used by NEONS are the binary formats adopted by the World Meteorological Organization (WMO) for global exchange of real-time weather data: Binary Universal Format for data Representation (BUFR) for observations and GRIB for GRIdded Binary numerical fields.

A NEONS NETWORK

With the expanding use of NEONS within the Navy, in NOAA and in other countries such as Canada and Australia, NRL and OAB have promoted the concept of a distributed network of NEON systems to facilitate the global exchange of weather and ocean data. In this concept, each office would continue to load and process its own data and in addition, would make the data easily available to others over INTERNET.

NRL has developed an X-Windows Data Browser to interactively browse files on NEON systems, search for particular data sets, and download data of interest. Using the browser, a user can specify the time and area where he wishes information. The browser then searches the database, either locally or remotely over INTERNET, to find satellite images, gridded model outputs or observations that fall in the desired time/space window. The user then interacts with the database to narrow the search to the actual data required.

The data can be downloaded in any of a wide variety of formats. The potential for such a network can be demonstrated by considering the gigabytes of climate data now being loaded by NCDC into their NEONS. Early this year NCDC described their work in the GCPS as follows: "NEONS is up and running. Sequences and parameters have been defined for the Global Historical Climatological Network (GHCN) data set (monthly global surface temperature, precipitation, and station and sea level pressure), Global Precipitation (GCPC) dataset, Cooperative Summary of the Day (TD3200) (daily max and min temps, precipitation, snow fall and snow depth), and CARDS. We are presently working on tying in the system with the Metadata portion of the STORM system (also an EMPRESS database system) to link the data to the station histories. We are also discussing with NRL the strong possibility of working with them on the development of an interactive interface between NEONS and NCAR Graphics for Lat/Lon/Time data) a system like the one they have developed for gridded data."

To promote the idea of a network of NEON systems, OAB started an OMNET bulletin board for NEONS users and, in cooperation with NRL and FNOC, hosted a NEONS Users Conferences in April 1992 and again in April 1993. The next NEONS users meeting will be held in conjunction with the AMS conference in January 1994.

SUMMARY

The Navy, through the Naval Space Warfare Systems Command, has invested over $4M in the NEONS program. Other government agencies should and can take advantage of this investment. Offices wanting more information about NEONS or wanting to install the system should write to OAB/NOAA, 2560 Garden Road, Monterey CA 93940.

Nancy Soreide

Personal-

Name: Nancy Soreide

Title: System Analyst

Affiliation: NOAA/PMEL

Address: 7600 Sand Point Wy NE, Seattle WA 98115

email: nns@noaapmel.gov

phone: 206-526-6728

fax: 206-526-6774

Data System-

Data system name: TOGA-TAO Display Software, EPIC

Discipline: Oceanogaphy

Data Managed:

Type of Data: oceanographic time series
Inventory Meta Data [Y/N]: Y
Digital Data, Data Products [Y/N]: Y
Number of Data Granules: aprox 3000 time series
Total volume of Data [Megabytes]: aprox 175 MBytes

Type of Data: oceanographic profile (depth-indexed) data
Inventory Meta Data [Y/N]: Y
Digital Data, Data Products [Y/N]: Y
Number of Data Granules: aprox 35,000 profiles (1 CTD=1 profile)
Total volume of Data [Megabytes]: aprox 530 MBytes

Data Management Activities Summary-

TOGA-TAO DISPLAY SOFTWARE

The TOGA-TAO Array consists of 63 moored ATLAS wind and thermistor chain and current meter buoys, spanning the Pacific Basin from 95W in the eastern Pacific to 137E in the west, transmitting data in real- time via the Argos satellite system. PMEL has developed the TAO workstation display software for distribution, and display, of the TOGA- TAO buoy data in a point-and-click environment. Data displays include buoy summary plots, buoy sensor plots, vertical stacks of plots from any collection of buoy/sensor pairs, as well as animations of TAO buoy data, operational ocean model analyses from NMC, TOGA drifting buoys, and climatological averages. TOGA-TAO buoy data displayed on the workstation are updated by acquisition from the Argos satellite shore- based computer during the previous night. The TOGA-TAO data sets and display software are available on PMEL's anonymous FTP on the Internet network, and have been distributed world-wide. Automated procedures provide remote users with updated data and graphics files. The TAO software is based on X-windows, and can be run on a local or wide-area network.

EPIC

The EPIC system was designed to manage the large volume of oceanographic time series and hydrographic data being collected by PMEL oceanographers participating in NOAA's large scale ocean climate study programs, such as EPOCS and TOGA. At present, over 35,000 data sets are on-line in EPIC at PMEL for retrospective analysis. EPIC is a complete system including a data selection module and a suite of over 100 graphics display and analysis programs. Supported data types include time series data, (such as temperature, wind and current time series from moored buoys), drifter data, acoustic doppler data, and profile data, (such as CTD, bottle and XBT data). System elements for data selection, data display, and data analysis, function independently. The system is well documented, with a user manual and extensive on-line help.

Robert Starek

Personal-

Name: Robert Starek

Title: Program Manager

Affiliation: Naval Oceanographic Office

Address: Code OTM, 1002 Balch Blvd, Stennis Space Center, MS 39522

phone: (601)688-5189

fax: (601)688-5701

Data Management Activities Summary-

I am the program manager for an effort to develop a geo-referenced, multi-disciplinary data base for the Naval Oceanographic Office. The system, entitled the Integrated Data Base Management System (IDBMS), will consist of Department of Defense (DoD) oceanographic and Mapping, Charting and Geodesy data. The architecture of the IDBMS is distributed in that data will be physically located closest to the point of greatest use on server class computers. A majority of the data will be stored in a relational data base and will be accessed and managed using a client- server based concept. Tightly coupled to the actual data will be a catalog consisted of meta data. A graphical user interface using visual references such as maps will allow for spatial browse and query of the catalog. Access to IDBMS data by non-Naval Oceanographic Office persons will require explicit DoD approval.

Leonard Walstad

Personal-

Name: Leonard Walstad

Title: Dr.

Affiliation: College of Oceanic and Atmospheric Science, Oregon State U.

Address: Ocean Admin Bldg 104, Corvallis, OR 97331-5503

email: lwalstad@oce.orst.edu l.walstad/omnet

phone: (503) 737-2070

fax: (503) 737-2064

Data Management Activities Summary-

I am currently chairman of the GLOBEC data management committee. The GLOBEC steering committee has endorsed the objective of choosing a data management framework which makes use of a distributed data management system. Furthermore, we believe that the system must encourage interaction, rather than hinder the exchange of data.

As an ocean modeler, my data sets are generally stored as flat files in machine format floating point and integer numbers. However, I generally make extensive use of the output of these numerical models for purposes beyond the usual graphical display. Typical uses are forcing of biological systems, analysis of float tracks, and vorticity and energy dynamics. Because I make extensive use of the output and use several numerical models, an analysis system would be of great benefit. A key concern is the ability of databases to deal with single entities which involve hundreds of MB of data (i.e. the output of a single snapshot of a biophysical model).

I use data assimilation to provide initial and boundary conditions, and also to update the physical fields within numerical models. This component of my research would be substantially simplified if an efficient interface to data was provided.

I believe that oceanographers need communication tools more than traditional database tools. Several key components of a valuable system are:

1. extensibility
2. explorability
3. modularity
4. efficiency
5. abstraction

Warren B. White

Personal-

Name: Warren B. White

Title: Physical Oceanographer

Affiliation: Scripps Institution of Oceanography

Address: SIO/UCSD, La Jolla, CA 92093-0230

email: wbwhite@ucsd.edu

phone: 619-534-4826

fax: 619-534-8041

Data System-

Data system name: JEDA Center

Discipline: quality-controled temperature-depth observations over the globe; monthly mean global upper ocean temperature gridded fields (1979-1993)

Data Managed:

Type of Data: temperature-depth observations

Inventory Meta Data [Y/N]: yes, in the GTSPP format

Digital Data, Data Products [Y/N]: yes, in the NetCDF format

Number of Data Granules: 1-2 million

Total volume of Data [Megabytes]: 1.5-2.0 GBytes

Data Management Activities Summary-

We are a WOCE DAC for upper ocean temperature profiles collected for the period 1990-1995. Presently, we have conducted quality control on the historical file for the 11-year period 1979-1989. We are presently conducted QC on the delayed-mode data collated by NODC for the period 1990. One of the steps in quality control is the gridding of upper ocean temperature anomalies. We have accomplished this for the period 1979- 1989 at 11 standard levels in the upper 400 m. The standard grid is 2o latitude by 5o longitude by month over as much of the ocean as the observations will allow. Interpolation errors also accompany these gridded estimates. Our climatological reference is computed on this same spatial grid each month for the 10-year period from 1979-1988. We plan to publish to the oceanographic community (over the Internet) the QC profiles, the gridded anomaly products, and the climatological reference.

J. R. Wilson

Personal-

Name: J. R. Wilson

Title: Director

Affiliation: Marine Environmental Data Service

Address: Department of Fisheries and Oceans, 200 Kent Street, Ottawa, Ontario, K1A 0E6, Canada

email: R.Wilson.MEDS (Omnet)

phone: (613)990-3009

fax: (613)990-5510

Data System-

Data system name: MEDS Oceanographic Database

Discipline: Physical Oceanography

Data Managed:

Type of Data: Water column phy and chem properties
Inventory Meta Data [Y/N]: Y
Digital Data, Data Products [Y/N]: Y
Number of Data Granules: 600,000 multi-variable stations
Total volume of Data [Megabytes]: 1,000

Type of Data: Drifting buoy data
Inventory Meta Data [Y/N]: Y
Digital Data, Data Products [Y/N]: Y
Number of Data Granules: 5,000,000 observations
Total volume of Data [Megabytes]: 2,000

Type of Data: Wave measured and hindcast data
Inventory Meta Data [Y/N]: Y
Digital Data, Data Products [Y/N]: Y
Number of Data Granules: 4000 station years

Total volume of Data (Megabytes]: 30,000

Type of Data: Inland and coastal water levels

Inventory Meta Data [Y/N]: Y

Digital Data, Data Products [Y/N]: Y

Number of Data Granules: 7,500 station years

Total volume of Data [Megabytes]: 1,000

Data system name: DFO Chemical Contaminants Data Directory

Discipline: Chemical Oceanography

Data Managed:

Type of Data: National Directory of Data Sources in DFO

Inventory Meta Data [Y/N]: Y

Digital Data, Data Products [Y/N]: Y

Number of Data Granules: 7 regional inventories

Total volume of Data [Megabytes]: .05

Data Management Activities Summary-

MEDS is involved in the acquisition, processing, quality control, archival, and dissemination by broadcast and on request of ocean station, wave, and water level data. These activities support national and international research, engineering, regulation, and management activities.

Typical applications of our data include, real time water level data for use in setting flow rates along the St. Lawrence Seaway, real time wave data for forecasts and warnings, engineering quality wave data for design of fixed and floating structures and ships, data in support of national research activities, and regulation and management of fisheries. MEDS also provides data management and dissemination services in support of international data exchange programs in general and the WOCE Surface Velocity and Upper Ocean Thermal Programs.

Data systems used include sequential tape systems for the voluminous wave and water level time series data, on-line VMS ISAMs for the drifting buoy, ocean station, water level, and wave spectral data, and Oracle relational databases for inventories and the contaminants data directory.

Over the past two years work has been carried out on the development of distributed client/server database capabilities using commercial software such as Oracle and SQLnet, and on a more general system utilizing the exchange of standard request, menu, raster, graphic, text, and data objects in a style that accommodates dissimilar database technologies including home built ones.

Design criteria for the system are as follows:

1. Client/Server Architecture (minimize network traffic).
2. Software in the field (client) not data or product specific. Host or server services can be changed and enhanced without modifying client software.
3. Not necessary for user to have a personal account and password on the server. User does not login on the server in the traditional fashion.
4. System allows a user to browse an inventory, identify data, and copy data back to the client from an associated database.
5. The client will be a workstation operating in a "windows" mode usually with various tools available for such things as GIS, spreadsheet, database.

Implementation of such a system was found to involve the development of a number of standard objects including the list given above. Objects are passed in both directions between client and server. Both client and server have installed processing and communications modules. The communications modules consist of network "listeners" which detect the arrival of standard objects in some fashion. In the prototype system in DEC VMS the "listener" consisted of a DCL procedure which woke up every 10 seconds and tested for the presence of a file with a certain name. Once the file was detected a program was invoked that opened the file and discovered the type of data, the type of product required, space-time ranges, parameter identifications, etc. The arriving object also carried the node address to which the reply object would be sent. Menu and request objects carried further information such as programs to be run to process the request or reply.

Such a system was found to be relatively generic and simple. It relieved the need for someone to login to the server in the traditional sense and the associated security risks. It also relieved the need for large number of user licenses on the server as the listener/processor operated as a single user. Since the standardization was in the exchange objects and their format was cast in concrete, software could be coded on the server to deal with any type of database including home built ones such as the ISAMs used in MEDS.

One purpose of the system and the design was to facilitate the development of ad hoc GIS databases. Point or polygon data could be returned to the client with raster, text, or vector graphic objects attached to it. The point and polygon data would be loaded up into the GIS. The user could click on a point or polygon on the map, see what raster, text, or vector objects were attached to it, select one, and have a window popped with a photograph, vector graphic, or textual information displayed. Similarly the user could have retrieved an object from the server formatted for uploading into a spreadsheet or Dbase application on the client.

This development and the pilot project were done over the past two years. We have not yet managed to find the resources to code and implement a production system.