N94- 33827

User Interface Development and Metadata Considerations for the Atmospheric Radiation Measurement (ARM) Archive

P. T. Singley, J. D. Bell, P. F. Daugherty, C. A. Hubbs, and J. G. Tuggle

Environmental Sciences Division
Oak Ridge National Laboratory
Bldg. 0907, Mail Stop 6490
P.O. Box 2008
Oak Ridge, Tennessee 37831-6490
Phone: (615) 574-7817
Fax: (615) 574-4665
sin@ornl.gov

Based on work performed at the Oak Ridge National Laboratory Oak Ridge, Tennessee 37831 managed for the U.S. Department of Energy under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.

Introduction

This paper will discuss user interface development and the structure and use of metadata for the Atmospheric Radiation Measurement (ARM) Archive. The ARM Archive, located at Oak Ridge National Laboratory (ORNL) in Oak Ridge, Tennessee, is the data repository for the U.S. Department of Energy's (DOE's) ARM Project. After a short description of the ARM Project and the ARM Archive's role, we will consider the philosophy and goals, constraints, and prototype implementation of the user interface for the archive. We will also describe the metadata that are stored at the archive and support the user interface.

ARM is a part of the global research effort directed toward understanding weather and climate change. The current generation of climate simulations, general circulation models (GCMs), cannot treat the physics of radiative transport and cloud behavior at the relevant distance scales. DOE has initiated ARM to characterize the physical and dynamical structure of the atmospheric column well enough to significantly improve the modeling of the radiative flux of the earth. This entails measuring radiative fluxes and a wide range of atmospheric conditions at five highly instrumented sites worldwide. The ARM sites constitute the Cloud and Radiation Testbed (CART). Each site will collect data from all its instruments for transmission to the ARM Archive. The first site, in Lamont, Oklahoma, has been operational since June 1992. Other sites will come on-line over the next few years.

The ARM Archive stores ARM data and will provide the scientific community with data taken from the sites, along with data developed from the merger of site data with data from external sources, information describing the quality assurance (QA) checks, and contextual information. The archive will eventually use a mass storage system architecture based on the National Storage Laboratory (NSL) architecture to manage and store these data. This system uses a relatively small computer that controls a group of mass storage devices linked by a high-speed data network.

Philosophy and Goals of the ARM User Interface

ARM Archive users—students, government-funded researchers, and policymakers—span a fairly large range of interests and capabilities. Initially, the user interface will be primarily designed to support professional researchers in atmospheric science and related disciplines. As more information about the data is available at the ARM Archive and more summary and

C-6

4

value-added data products are created, the focus of the user interface will expand to support a broader user community.

To support users, both initially and over the long term, the principal goal of the user interface is to provide enough information about data and products that are housed in the ARM Archive so the users can select exactly the data they need. To make the necessary information as accessible as possible, the user interface is designed to address the users in terms familiar to them. For instance, the user interface offers complete instrument names rather than the cryptic abbreviations that instruments are known by within the CART Data System. In addition to providing the users with familiar terminology, the user interface hides the details of how data are managed at the ARM Archive. The users do not need to know about file names, data base structures, or how to develop a data base query to get access to archive data. To successfully retrieve data from the archive, the users indicate the instruments of interest, date ranges, and other criteria (such as data processing level or QA level) that will narrow their selection. Then the users request that the data be retrieved.

Another goal of the user interface is to make sure that it is inexpensive enough so that it can be given away to anyone. Furthermore, every effort will be made to port the interface to a wide variety of computer hardware. Input from users working with the initial interface will help refine this system. This should ensure that the interface will continue to provide easy access to the ARM Archive.

Constraints on the User Interface

Constraints as well as goals shape all systems. In addition to the usual limits on money, time, and resources, the ARM Archive has several technical constraints, some unique to the ARM Archive and some that affect any large scientific data archive.

One of the leading constraints for a large scientific data archive is simply its size. Although not as large as several of the NASA data centers, the ARM Archive will hold as much as 100 terabytes of data and metadata by the time the ARM Project ends. With that amount of data and metadata, maintaining all or even a significant portion of it on spinning disk in a data base management system is not feasible, either technically or economically. Almost all of the data, and a good fraction of the metadata, will be maintained only as files in a tape-based mass storage system. The user interface will be based primarily on the metadata that are kept in a relational data base management system (RDBMS). In addition, some metadata will be managed and accessed in a Wide Area Information Server (WAIS), a search/retrieval system for computer networks. The smallest data granule given to the user will be a file. Because of the volume of information, users will not be able to directly browse the data. Summary or value-added products may be created so users can browse.

Another aspect of keeping the data files in a mass storage system is that the user can only request that data be retrieved from the system and not examine these data using the user interface. The user must later retrieve the requested files via "ftp" (electronic transfer) or wait for surface mail to deliver the physical media containing the requested data to the user's computer. Direct interactive exploration of the data in the ARM Archive is not available.

Additional difficulties are imposed by the fact that not all of the data or metadata arrive at the archive produced in the same format, written with the same degree of formality, or subject to the same level of quality control. This diversity and how it is dealt with by the ARM Archive system are discussed later in the paper. Briefly, most of the data and the formal metadata, such as instrument, location, and current calibration readings, arrive packaged in highly structured NetCDF files generated by the Site Data System (SDS). Some data arrive in the file format of the instrument that produced them, with little associated metadata. Finally, there are logs describing conditions of instruments and other information about the site that affect instrument operation. These are human generated and fairly informal, and there is little or no quality control of the entries. All of this variability makes presenting the users with the necessary metadata in a uniform fashion a very challenging problem.

Implementation of the User Interface

We have chosen to build the user interface on a client/server model shown schematically in Fig. 1. The user interacts with screens provided by a client application. This client sends requests to, and receives data from, a server system, which includes the archive's RDBMS and file-retrieval system. Currently, TELNET is used to access the archive's host computer and the user interface. In the future, the client will reside on the user's local computer while the server runs on the archive's host computer. Our prototype interface is based on the X-windows protocol using the Motif window manager.

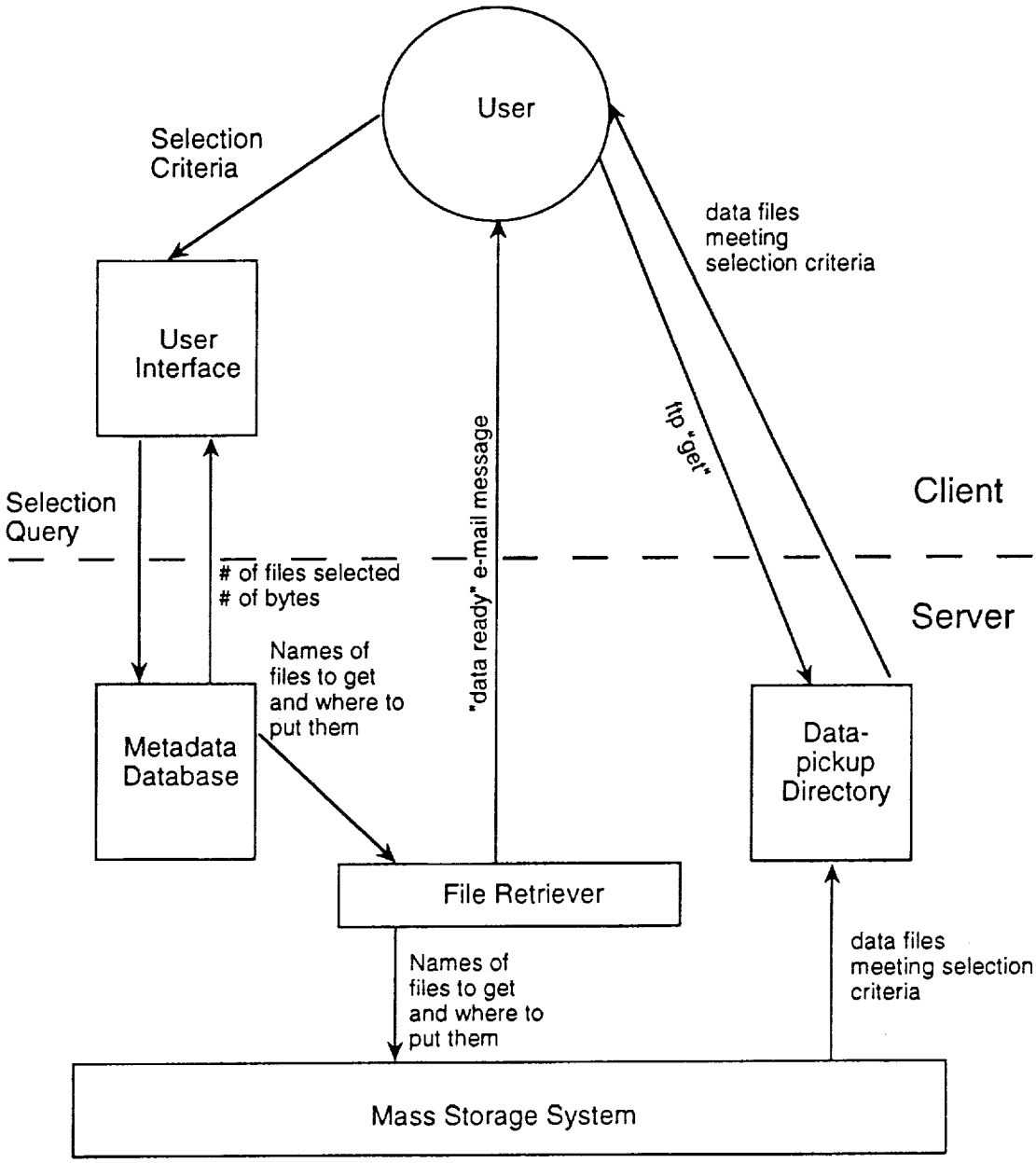
The initial screen is shown in Fig. 2. Here the user is asked for a user ID and an easily remembered password. On the second screen of each session (Fig. 3), the current identifying data about this user are displayed for verification: name, phone number, electronic-mail address, and surface-mail address. New users will be prompted for these data when they first log on to the archive.

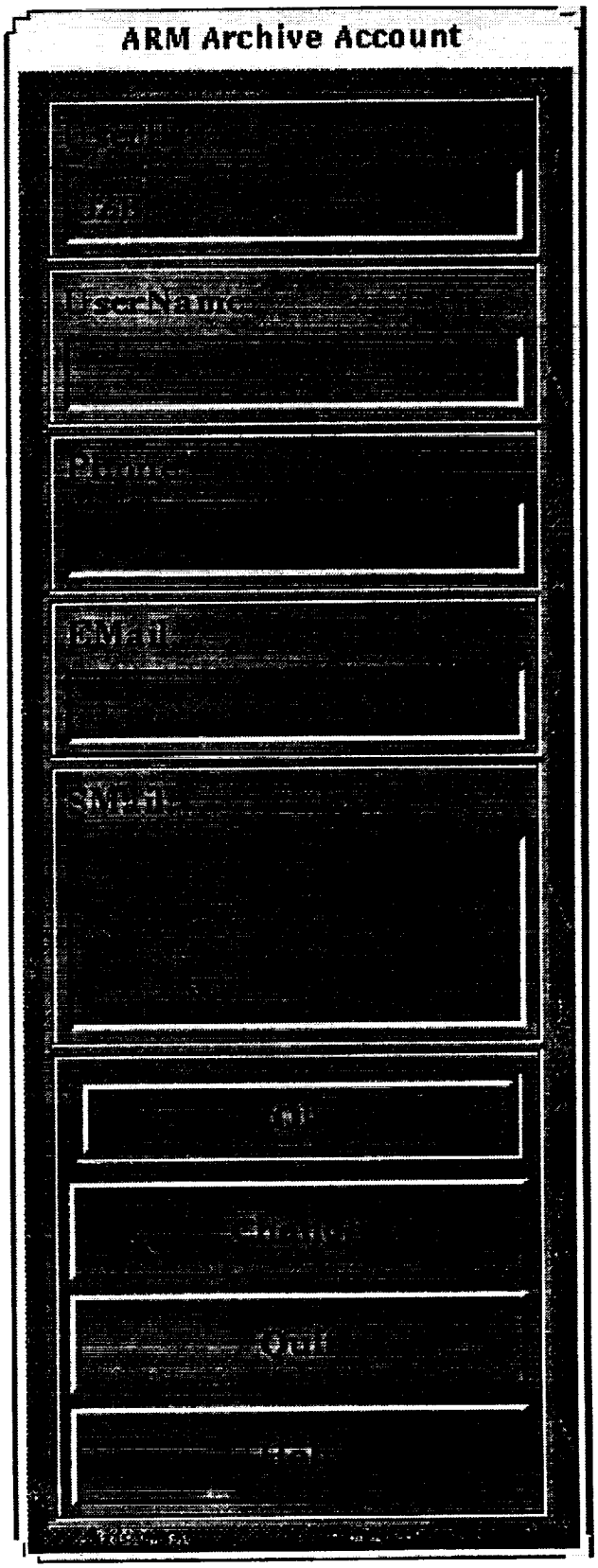
The SDS packages data into files labeled by platform and date. The Data Criteria screen (Fig. 4) is hence composed of sections for selecting a set of platforms and a range of dates. The user can select platforms based on either the SDS platform name or its equivalent "plain English" instrument name (and accompanying location information). This prototype assumes that the user already knows which named data streams are contained in the data files for a single platform or instrument; an RDBMS table and interface screens for this mapping will be added later. The user can type or use a scroll bar to specify the start and end of the date interval for which he or she wishes to get data. The system will return all those data files that include at least one data point within the interval specified.

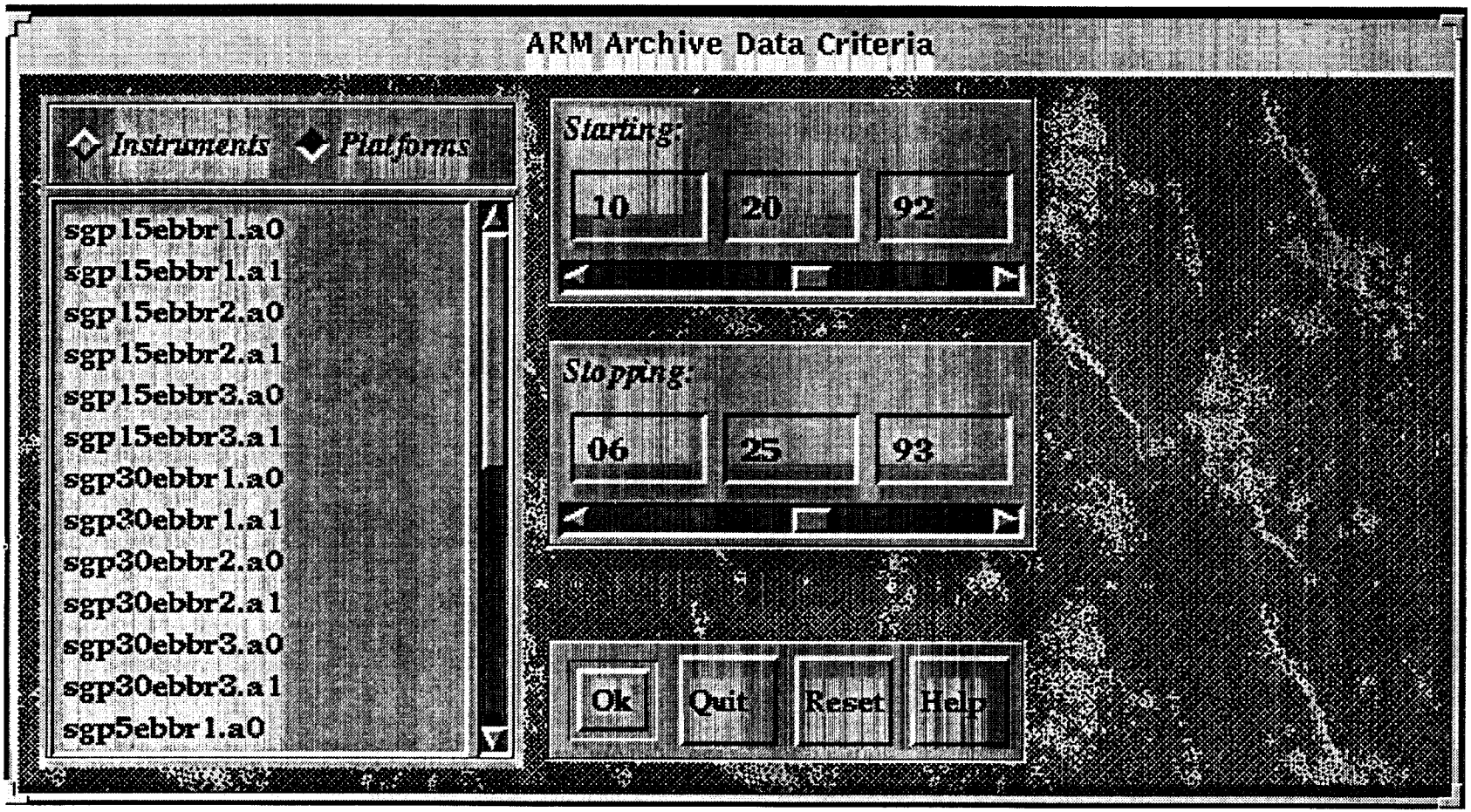
The restricted set of queries that users can presently make (based on the small number of search parameters) is stored as string templates in the client C-program. They are filled in on the basis of the user's choices from the Data Criteria screen, then executed against the metadata data base. When that query is executed, a Transfer Confirmation screen (Fig. 5) displays the number of files (and the total number of bytes of data contained therein) that meet the selection criteria. The user may then choose Initiate or Cancel at this screen. "Initiate" causes the client program to request that the files meeting the selection criteria be retrieved from mass storage and delivered to a pickup directory on the server computer system; "Cancel" returns the user to the Data Criteria screen directly for a new query without requesting data retrieval. Back at the Data Criteria screen, the user may refine the existing query by altering a few values or start afresh (Reset) by clearing all of the settings chosen for the previous query.

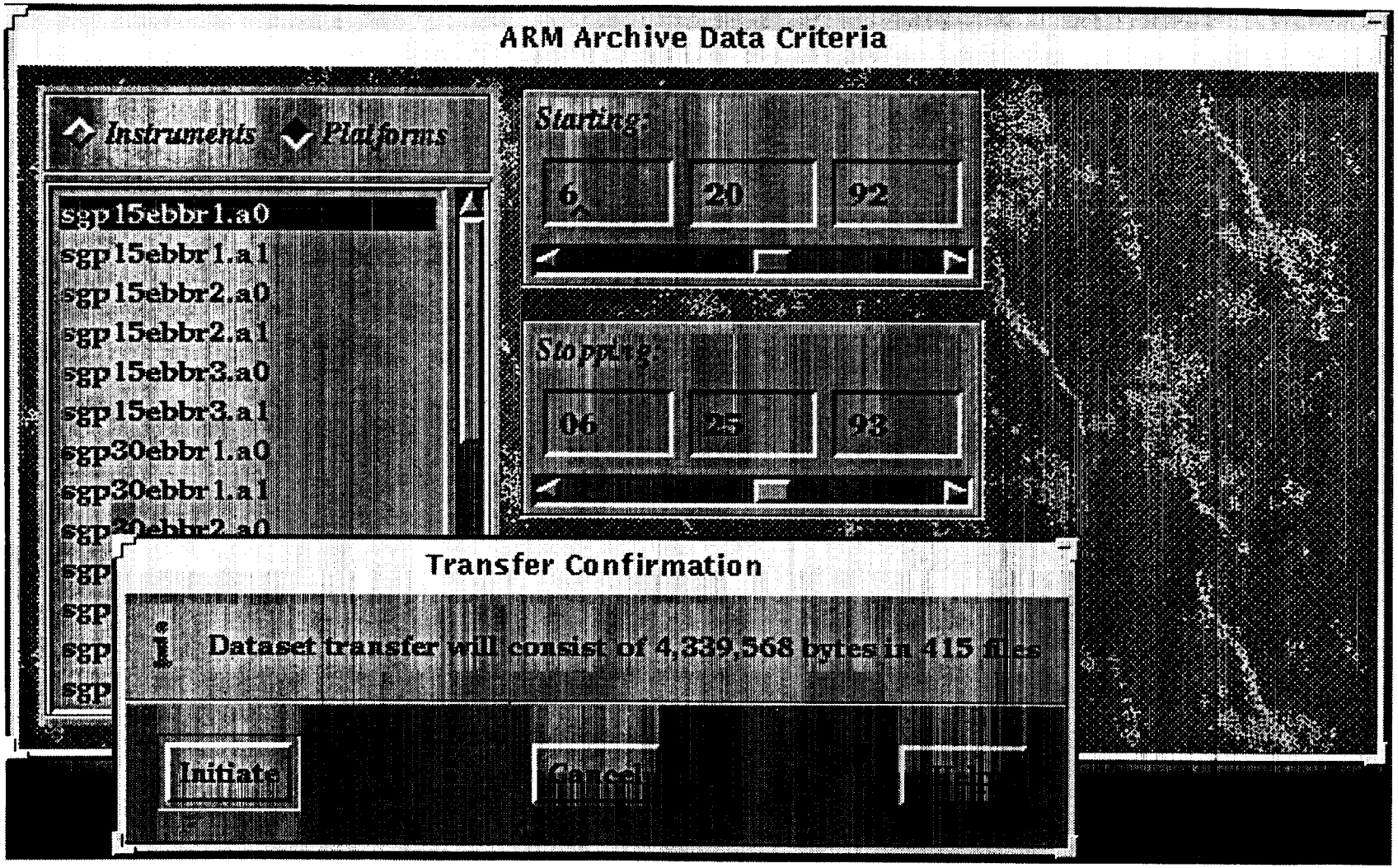
The server portion of our user interface consists of the metadata data base and the file-retrieval and user-notification system. The metadata data base is implemented as a set of tables in the Sybase RDBMS, which are queried by the client program through Sybase-provided library functions. The metadata that pertain to data files stored at the archive are described below; information about user addresses and tables for processing requests for file retrieval are also stored in this data base. When the user initiates a request for data from the client, the file names requested are written into a Sybase table, which is then read by a process that submits a retrieval request to our mass storage system (currently, a Storage Tek Silo managed by an IBM 3090 computer). The files are retrieved via ftp and then put into a holding directory on the host computer to await pickup by the user.

When all of the files requested by a user in a single session have been retrieved from the mass storage system, an electronic-mail message is sent to the user, informing him or her of the availability of the requested data . The user can then copy those data files by ftp to his or her local computer. After an interval of a few days, the files will be deleted from the pickup directory.


User
Selection Criteria
data files meeting selection criteria
User Interface
ftp "get"
"data ready" e-mail message
Client
Selection Query
# of files selected
# of bytes
Server
Names of files to get and where to put them
Metadata Database
Data-pickup Directory
File Retriever
Names of files to get and where to put them
data files meeting selection criteria
Mass Storage System




ARM Archive Data Criteria
Instruments
Platforms
sgp15ebbr1.a0
sgp15ebbr1.a1
sgp15ebbr2.a0
sgp15ebbr2.a1
sgp15ebbr3.a0
sgp15ebbr3.a1
sgp30ebbr1.a0
sgp30ebbr1.a1
sgp30ebbr2.a0
sgp30ebbr2.a1
sgp30ebbr3.a0
sgp30ebbr3.a1
sgp5ebbr1.a0
Starting:
10
20
92
Stopping:
06
25
93
Ok
Quit
Reset
Help


ARM Archive Data Criteria
Instruments
Platforms
sgp15ebbr1.a0
sgp15ebbr1.a1
sgp15ebbr2.a0
sgp15ebbr2.a1
sgp15ebbr3.a0
sgp15ebbr3.a1
sgp30ebbr1.a0
sgp30ebbr1.a1
Starting:
6
20
92
Stopping:
06
25
93
Transfer Confirmation
Dataset transfer will consist of 4,339,568 bytes in 415 files
Initiate
Cancel

Metadata for the User Interface and the Archive

As noted above, the primary unit of data given to the user is a single data file, which contains data recorded over a specified time interval from a specified set of sensors (an instrument or platform). In order to allow the user to select the desired files, retrieve them, and use the data properly, we must deal with three classes of metadata: (1) data extracted directly from the individual data files, (2) other data about individual files (e.g., file size and storage location), and (3) site operations logs and other documentary information that are not keyed to a specific data file.

To extract needed metadata from the data files themselves, we have a copy of the suite of programs that produces the NetCDF files for the SDS. We use this code to extract the data-start and data-end dates and times, and the number of samples for each variable from each NetCDF data file, for entry into our data base. As files arrive at the archive, we record their file name, arrival date/time, and file size in our Sybase data base. Further information about storage locations and dates is collected as the files are sent to our mass storage system and as a permanent, vault-archived copy of the files is written by that system. Other metadata include the Site Operations Log, platform and data dictionaries for instruments, and other textual information that may affect the correctness or usability of data files but are not directly linked to them. The Site Ops Log is being stored in a table in the metadata data base (based on date of entry), as well as being archived as entries arrive. The other text files will be managed (and accessed by the users) through the WAIS system, which is specifically designed to allow browsing of large free-text data files for keywords.

The current stock of metadata may be significantly expanded as the ARM Project continues. As discussed above, all of the header information in the NetCDF files from the SDS could be incorporated into the metadata data base. We are also starting to explore schemes to allow users or persons responsible for specific instruments to comment on data files, perhaps at the single data point level. (In this case, we intend for users requesting data with existing comments to receive the comments along with the data files.) We further expect that some metadata will arrive in a form that is not computer-readable, and we are pondering our response.

Future Improvements to the User Interface

To make the ARM Archive user interface as helpful as possible, the capabilities that have been discussed in this paper must be extended to provide the users with more information about the available data. In addition to making more information available about the holdings of the ARM Archive in general, more selection criteria need to be available for the users to refine their requests for data.

Additional selection criteria will be derived from the formal metadata transmitted with the data from the CART sites. As with the current selection criteria of platform and date, this data will be managed with the use of the metadata RDBMS. In order to implement additional selection criteria, we need to work with the user community to identify the useful selection criteria for each platform and to extract that metadata from the data files and place it in the metadata data base. We also envision a desire to select data files for one platform on the basis of the availability of data from another platform for the same time interval: "Give me the BSRN1 data for June 1993 where EBBR9 data exist," for example. The user interface is designed so that new selection criteria can be easily added to the user interface screens.

Most of the information about the ARM Archive that explains the contents to users is in textual form. If ARM data are to be accessible to relatively naive users, this textual information must be made available on-line. The current plans are to manage textual information about the ARM Project with a WAIS server. A WAIS client will become part of the client portion of the user interface to make all ARM text available for perusal and downloading through the user interface.

Several potential metadata sources, such as operators' descriptions of instrument status in the site operations log, are textual with little formal structure. This type of information can be critical to the user in deciding if a particular data file is desired or not. In order to make text information part of the selection parameters, links need to be developed between the metadata kept in the RDBMS and those kept in the WAIS system. We are developing a design to provide this connection using a common identifier in the RDBMS records and the text record that will allow textual information as part of the selection criteria for requesting data files.

As a final assistance to users in selecting data, we are exploring the possibility of logically linking textual comments to data files. The proposed implementation would allow the users to see brief comments on the data files that they are about to request. On the basis of those comments, they might elect to remove some data files from their request. The comments would deal with data quality and use issues that were not captured in other parts of the metadata system. One proposal is that some of these comments might be from previous users of the data.

Conclusions

Scientific data are useful only when they are producing scientific or policy results. The ARM Archive user interface is designed to make the ARM data quickly and easily available to the user community. To accomplish this goal, the user interface will provide the users with information in the terms of the atmospheric science discipline. Over time, it will also provide users with extensive documentation about the condition of the data, why and how the data were collected, and other information to make data selection and use easier.

Metadata that provide clear, accurate, and precise information about the context of the basic data are necessary to support this type of user interface. The ARM metadata are being organized with the use of an RDBMS for data that are very formally organized and lend themselves to management in tabular format. For those data that are textual and do not easily fit the row/column format, a WAIS system will be used for data management and access. In the future the ARM Archive will explore ways to link the metadata in the RDBMS and WAIS systems to provide users with both a rich set of selection parameters and concise descriptions of the data they are requesting.