CONF-970478--1

Inside an Environmental Data Archive WWW Site*

Jon W. Grubb, Sarah V. Jennings, Teresa G. Yow, Ph.D., Anthony W. Smith

Abstract

The Oak Ridge National Laboratory (ORNL) Distributed Active Archive Center (DAAC), which is associated with NASA's Earth Observing System Data and Information System (EOSDIS), provides access to tabular and imagery datasets used in ecological and environmental research. Because of its large and diverse data holdings, twin challenges for the ORNL DAAC are to help users find data of interest from the hundreds of thousands of files available at the DAAC without overwhelming them and to manage such a large collection of data. Therefore, the ORNL DAAC has developed a number of World Wide Web (WWW) tools such as the Biogeochemical Information Ordering Management Environment (BIOME), a WWW search and order system, as well as WWW-based data management and configuration control tools. This paper describes the specialized attributes incorporated into these systems that allow for easy access to and management of the data.

Category of Submission

This technical paper describes the operation of the ORNL DAAC public WWW environmental data archive site funded by NASA.

1. Introduction

1.1 The ORNL DAAC

The Oak Ridge National Laboratory (ORNL) Distributed Active Archive Center (DAAC) is one of nine data archive and distribution centers belonging to NASA's Earth Observing System Data and Information System (EOSDIS). Both the Earth Observing System (EOS) and EOSDIS are components of NASA's contribution to the U.S. Global Change Research Program through their Mission to Planet Earth Program.

The ORNL DAAC specializes in archiving and distributing data and data products relating to the Earth's biogeochemical dynamics. These data often come from NASA-sponsored ground-based field investigations, and include tabular or point data and imagery from satellite and aircraft sensors. Non-NASA biogeochemical data and value-added products relevant to global change research are also included.

1.2 Creation of the ORNL DAAC WWW Site

In 1994 the ORNL DAAC created a World Wide Web (WWW) site using a National Center for Supercomputing Applications (NCSA) httpd 1.4 server and the Unix operating system on a Silicon Graphics workstation. The rationale for this development effort was that the WWW is the

premiere resource access mechanism available on the Internet today, and will remain so for the foreseeable future.

The first ORNL DAAC home page was a relatively simple text description of the DAAC, its holdings, and contact information for obtaining the data. In addition, detailed descriptions of each dataset were available. These text pages have been improved and updated, and they continue to serve as an option for users who wish to obtain data by that avenue. They are located at http://www-eosdis.ornl.gov.

As the number of datasets and users grew and the DAAC's knowledge of WWW processes increased, we saw the need for a search and order system to guide users to data. In response to this need we developed the Biogeochemical Information Ordering Management Environment (BIOME) search and order system in 1995. BIOME is located at http://www-eosdis.ornl.gov/BIOME/biome.html, and can be accessed directly from the DAAC home page by clicking on BIOME under Options for Ordering Data.

Also, as the complexity of site management increased, we saw the need for tools to automate site management. In 1996, the DAAC created WWW-based database management and configuration management (CM) tools. These tools are not accessible by the public, but will be described in following sections.

1.3 Overview of Unique Features of the ORNL DAAC WWW site

The ORNL DAAC WWW site uses many of the generic Web features that are in common use today, as well as advanced features that help users locate data quickly and easily and help site personnel manage the data. Some of those features are briefly described below; the subsections that follow describe each of these features in detail.

* Browser-aware linking and dynamic paging. The ORNL DAAC WWW site categorizes browsers based on their capabilities. The pages are modified according to the ability of the user's browser to display them. High-end browsers such as Netscape 3.0 can get pages with frames, tables, and Java applets in addition to the information available to the character-only pages available to character-based browsers such as Lynx.

* User search methodology. The site does not dictate the user's search starting point. By design, there are several points from which a user can start depending upon what the user already knows about his/her search goals.

* Metadata-based rapid access capabilities. Managing large amounts of data requires the use of metadata. The metadata is stored in a relational database management system (RDBMS) to allow for efficient searching of hundreds of thousands of metadata records.

* User-selected dynamic product packaging and delivery. Users can choose from multiple methods of data delivery (e.g., download, FTP, zipped, Mac format), with data formatted on-the-fly and automated as much as possible.

* On-the-fly graphs of tabular data. BIOME allows users to plot selected variables to view a graphical representation of tabular data.

* WWW-based metadata tools. A WWW interface allows the site's database manager to update, move, and manage data quickly and easily.

* WWW-based Configuration Management (CM) tools. A WWW interface allows the site's webmaster to update, move, and manage HTML files quickly and easily.

2. Browser-Aware Linking And Dynamic Paging

On-the-fly browser customization allows the ORNL DAAC WWW site to take advantage of the most innovative WWW features while still maintaining backwards compatibility with older browsers and text-based browsers. Because many of our users are scientific researchers working in remote areas, we try to balance their needs with those of users who have access to the latest technology.

The WWW interface developed at the ORNL DAAC allows any user with a Web browser and Internet access to search and order/download ORNL DAAC data directly to his/her machine. Because we have no way of knowing which browser a user might have and do not wish to exclude any potential users, we designed the interface to work with any browser capable of supporting forms, including non-GUI browsers (e.g., Lynx).

2.1 Browser Identification and Classification

A Bourne shell script (/cgi-bin/browser.sh) is executed as the home page is being constructed prior to being sent to the user's browser. The script examines the Unix environment variable HTTP_USER_AGENT, which contains the descriptive identification of the user's browser. The script launches a C program that parses the browser description and categorizes it into one of four classes of browsers (additional classes can be added in the future):

1. incompatibles that cannot handle forms (Cello, etc.),
2. character based that can still handle forms (Lynx, etc.),
3. Mosaic-compatible and variants (various Mosaics, HotJava, etc.), and
4. Netscape-compatible and variants.

As new previously unknown browsers are seen by the system, the user is requested to call or email ORNL DAAC User Services with a description of the browser. The browser is then entered into one of the classes listed above. In the one year period from August, 1995 to August, 1996, 1152 unique browser/platform combinations accessed the ORNL DAAC site.

2.2 Constructing Dynamic Pages

The user's browser class is stored in a file with the name based on the IP address of the user. This file is read by any other process that needs to know the browser class. This is how the pages are dynamically altered.

The script /cgi-bin/is_gui.sh uses the browser class to determine if the user's browser is capable of displaying images. If so, the section of the page that allows the user to select information based on images is included in the page. If the browser class does not allow images, this portion of the page is not included for display. Thus, the page has been dynamically altered based on the user's browser capabilities.

By design, the links that exist on a page are there only because the user's browser can display the pages linked to. Links to pages requiring image capabilities are all in page components shown only to browsers that can display images. In this way, the site exhibits dynamic linking.

While the advantages of browser-awareness for the user are apparent, there are some disadvantages for the system maintenance personnel. The design incorporates include statements that pull in the appropriate "modules" for each browser. Thus, there are a few additional files to be created and maintained; however, careful design of the interface has minimized these maintenance problems.

3. User Search Methodology Awareness

In general, users who are searching for information already have some initial information that they are trying to complete or expand. As explained previously, this web site contains a vast amount of information that can be difficult to search without a convenient starting point. In addition to the text pages, the BIOME user interface supplies the user with several starting points based upon what the user already knows (see Figure 1).

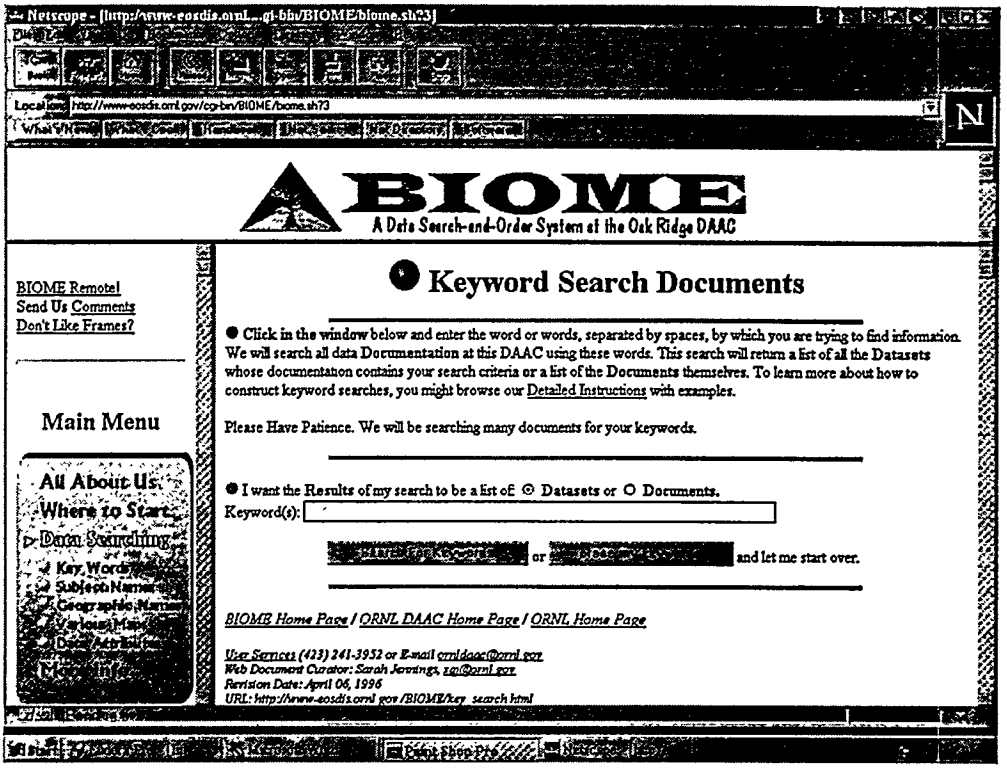

Figure 1: BIOME Search Options

The basic unit of storage at this web site is the "dataset." A dataset contains the results of one particular research effort. Each dataset can contain one or more data files. There are hundreds of datasets and hundreds of thousands of files. So how does a user get started? There are five starting points within BIOME from which a user may choose:

1. attribute search,
2. keyword search,
3. subject search,
4. region search, and
5. map search.

A discussion in the WWW Access to Earth Observation/Geo-referenced Data (http://www5conf.inria.fr/fich_html/workshops/Enloe.html) held May 6-10, 1996, in Paris, France, as part of the Fifth International World Wide Web Conference confirmed that these starting points are among those most commonly used by users searching for earth observation/geo-referenced data on the WWW.

3.1. Attribute Searching

Attribute searching, as shown in Figure 2, allows the user to search the DAAC metadata database by some attribute(s) of a dataset. BIOME displays a list of data attributes from which to select the search criteria (i.e., dataset name, investigator, source, sensor, or geophysical parameter).

This selection option provides the user with a continuously shrinking results set until the user has the desired dataset(s). A Sybase database (described in detail in Section 4) is searched for matches to a user's input, and the search results are put into a results form and returned to the user's browser.

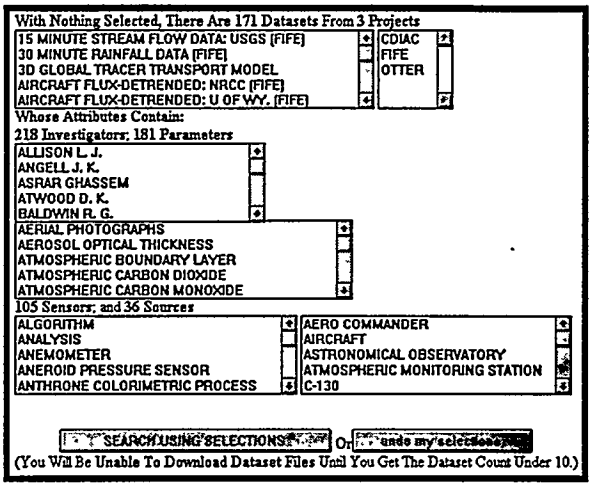

Figure 2: Attribute Search

3.2 Keyword Searching

For keyword searching the user enters a word, phrase, or Boolean combination of words and phrases. This selection criteria is then passed to a text search engine based on the Unix "grep" command. This command uses the user's input to search through all the data documentation to arrive at a list of datasets that meet the selection criteria. The search engine searches documents describing the datasets to create a list of documents that meet the search criteria. These documents are then mapped to their corresponding datasets.

3.3. Subject Searching

The data at this DAAC covers a variety of different subjects, e.g., meteorology, hydrology, atmospheric chemistry, etc. However, these wide subject categories may not be specifically mentioned in the documentation or the metadata. To assist the user in searching for information about a general subject, we have categorized datasets into various subject areas. The user can select one or more subjects from a subject list that is linked to a page containing dataset titles and links to data and documents.

3.4. Region Searching

Because the ORNL DAAC archives data from field investigations, many datasets are associated with specific geographic regions. This search method allows the user to select one or more geographic regions. Examples of geographic regions include the Indian Ocean, Oregon USA, and Mongolia. The latitudes and longitudes of the defined region are retrieved from the database;

those values are then compared to the latitude and longitude for each data file as listed in the metadata. All datasets that contain files with latitude and longitude values within the specified ranges are identified, and the datasets to which the files belong are identified. BIOME then returns a list of all the datasets containing data about the selected region(s).

3.5. Map Searching

BIOME allows users to select a point from various geographic maps, and the system returns all the datasets that contain data about the selected point. There is a world map and several continental maps. The maps are clickable images where the pixel location corresponds to a map position. The latitudes and longitudes of the defined map position are retrieved from the database; then those values are compared to the latitude and longitude for each data file as listed in the metadata. All datasets that contain files with the selected latitude and longitude within their latitude and longitude ranges are identified, as are the datasets to which the files belong. BIOME then returns a list of all the datasets containing data about the selected location.

3.6 Search Results

In all cases described above, the initial search results are shown on a screen similar to the attribute search screen. The metadata search loop may be repeated as many times as necessary to narrow a search down to just exactly what a user wants. A flow diagram that illustrates this process is shown in Figure 3.

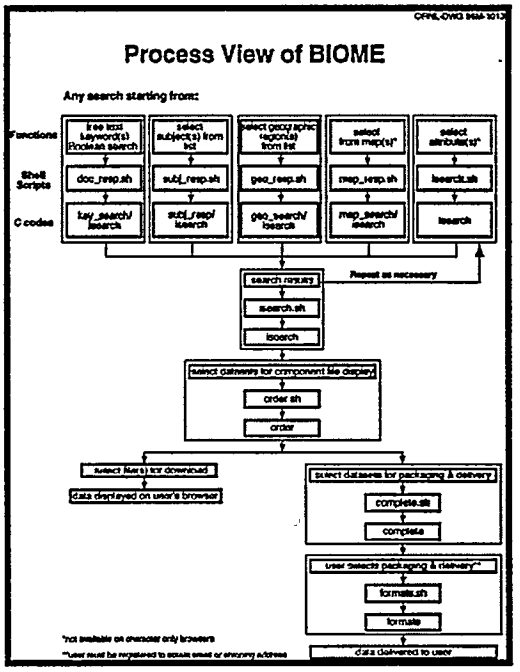

Figure 3: Process View of BIOME

4. Metadata-Based Rapid Access Capabilities

The metadata at this web site is contained in a Sybase database. There are two major components to this database: the dataset tables and the file or "inventory" tables. The dataset tables contain the broad attributes that apply to a complete dataset. The inventory tables contain those attributes that apply only to a particular file belonging to a dataset. Together, these tables constitute the metadata database. The actual scientific data is stored separately from the metadata. The location of the data files is included in the metadata database.

Because Sybase has a very clean and efficient interface to the C programming language, we made the decision to create small, specific C processes that perform very specific retrieval and page assembly functions for the web server. This means that a page will contain different components depending upon the results of a query.

The database was originally designed to be used for retrieval of metadata for display to an X-Windows display system. In order to keep from having to support two sets of tables, those same tables were used for the retrieval of web-based display data. Extra indexes and cross mapping tables were added when necessary to allow the needs of the web pages to be met while maintaining the existing database design. The result is a database that responds with completed pages in seconds and supports both the X and web interfaces. Usage has shown that in most cases the network is still the slowest part of any retrieval.

Any page that requires components to be built from a data query are called from a cgi-bin script. The script is always a Bourne shell script that handles any setup and environmental issues as well as providing the browser with the first few items of page output from cgi-bin scripts. The script always calls C code modules to do the actual query and finish the page output.

All forms pages are GET posted. Some forms also pass arguments in the page link call. The C code modules get the forms information from the GET input but also retrieve special information from the link call. Some pages are NOT forms but pass information through their link calls to other pages which ARE forms.

For example, the dataset results from a previous page are passed to a subsequent page to be used as part of the query so that the user gets more refined results with each query. The user can then order a complete dataset from a page that is not a form but was created as the result of a previous query. The link always calls the same page but passes as part of the link call the dataset name to be ordered. The called page reacts to the dataset argument by including information about that dataset in the custom-generated instructions on how to specify delivery.

5. User-Selected Dynamic Packaging and Delivery

Once the user has selected down to the dataset(s) of interest, he/she has the option of downloading data files directly to the browser or selecting complete datasets of files to be collected and either made available by FTP or recorded onto some media and shipped to the user.

If hard media such as CD-ROM or diskette is requested, the web site insists that the user "register" by providing an email and a shipping address for delivery.

Hard media requests require User Services intervention. A hard media request causes an email containing the user's data request and shipping information to go to User Services personnel. The requested data is placed into a special area for processing by User Services. User Services merely loads the requested media, records the data, and packages and ships the media.

FTP delivery requests are as automated as possible. To be considerate to others on the network, all requests are compressed into a single file but are in the platform format requested by the user. Again, the user must register to provide an email address for contact instructions.

The web site maintains Unix (tar, compress, gzip), PC (PK-zip), and Mac (Stuffit) packaging software that is invoked whenever the user requests FTP delivery, which is the most popular type of data delivery. The user is emailed an order confirmation when the order is placed and then emailed FTP instructions when the order is ready, usually within a few minutes. This is all done without any human intervention although every action is logged for human review at a later date.

6. On-the-fly Tabular Graphing

BIOME allows users to see a graph of selected data. Tabular data are parsed according to arbitrary classifications describing the configuration of the data. For each tabular data set there is a file describing the configuration of the tabular data and classifying the data. BIOME then parses the file based on the file classification and uses the GD1 library to generate a plot of the data. The user's browser is sent a .gif with the selected labeled columns plotted in color (see Figure 4 for a sample plot). This technique allows one graphic engine to display all the different layouts of tabular data.

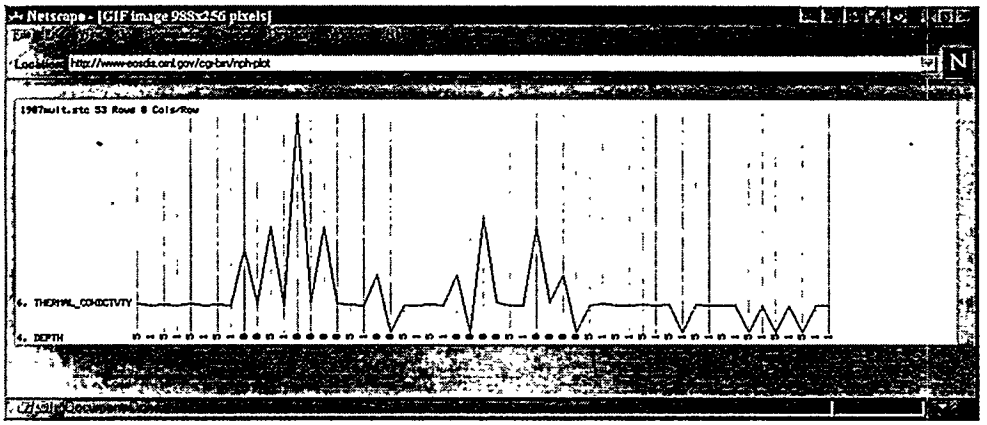

Figure 4: Sample Plot

The user can select which column of tabular data is to be plotted on the X-axis. The Y-values plot is autoarranged. Any column can be plotted on the X-axis, however only numerical columns can be y values.

The tabular grapher relies on the GD1 libraries available at http://www.boutell.com . The site also provides excellent documentation and examples.

7. WWW-based Metadata Tools

Metadata for the data archived at the ORNL DAAC is stored in several Sybase databases. As the complexity of the data holdings has increased, the task of maintaining the databases has become increasingly difficult and time-consuming. Fortunately, the Web-based database administrator (DBA) maintenance tool provides options that make the task of the database administrator less difficult. This tool is a GUI interface that uses HTML 2.0 cgi scripts (with some 3.0 capabilities) and C processes executed by the cgi scripts to access the databases and perform database functions using Sybases's DBLibrary. What makes this interface so useful and robust is the design options that have been custom-built and implemented.

For example, the DBA tool handles the ingest of new metadata by providing on-the-fly templates of database tables generated dynamically from Sybase's system tables. New data can be typed onto the templates, eliminating the need for manually construction Sybase bulk copy files using "vi" or a similar editor, a task that is tedious and error prone. In addition, the DBA Maintenance Tool easily handles updates to existing metadata. The tool offers such options as global updates to the databases; changes can be made to all tables in a database that contain a particular field as well as to other databases containing the same table and field. The tool also easily handles single updates to a database.

Other options include automated bulk copies out of the database and the printing of the current structure for each table. The DBA Tool also automatically generates a transaction log that provides a record of all DBA actions on the databases. Future enhancements will include automated database dumps, table creation options, and the granting of user privileges.

8. WWW-based Configuration Management Tools

One of the many software management issues the ORNL DAAC has addressed is how to allow multiple Web developers to work on a common set of HTML documents. The Revision Control System (RCS) was chosen to be the backbone for archiving and managing these documents.

The Directory Management System (DMS) as shown in Figure 5 provides a Web interface to RCS. By using a Web interface, this system is easily usable by those who are not familiar with RCS and its many UNIX commands. Furthermore, the information provided by RCS is arranged in a clear and concise layout on a single HTML page as opposed to performing a series of UNIX and RCS commands at the prompt. Examples of the information and the RCS functionality DMS provides the user are: (1) capability to quickly view all of the files which are and are not maintained in the RCS archives; (2) check-in and check-out files from the software archive; and (3) view which users have checked out which documents.

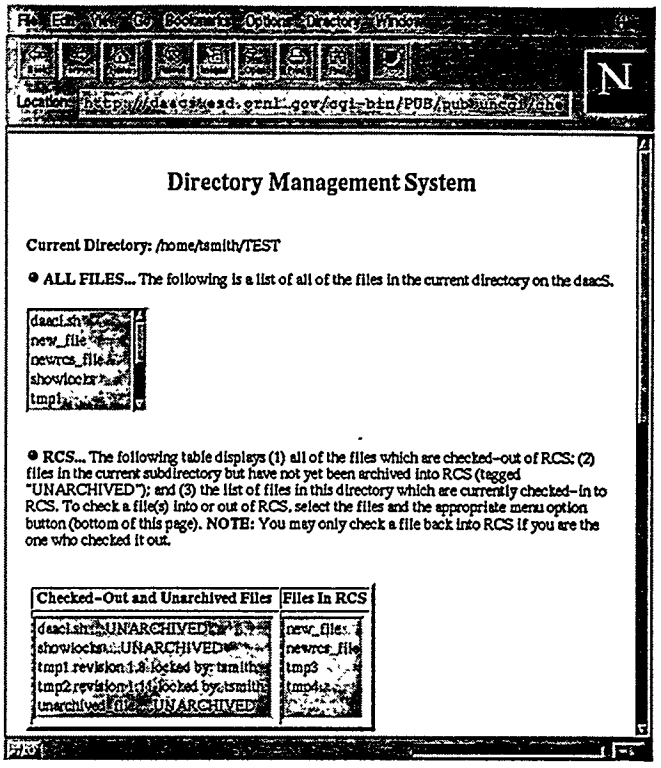

Figure 5: Directory Management System

Another function of the DMS is to copy documents from one UNIX machine to another. This is important to the DAAC because all Web development occurs on one machine, and when it is ready, is moved to another machine where it is then accessible to the world. What the DMS provides are mechanisms for determining which files on the development server are new and which ones have been modified compared to the operational server. Furthermore, DMS has the capability of tagging these new or modified files and uploading them to the operational environment. This system has greatly reduced the confusion of which files in the development area have been modified and have not yet upgraded to operational status. A sample copy status screen is shown in Figure 6.

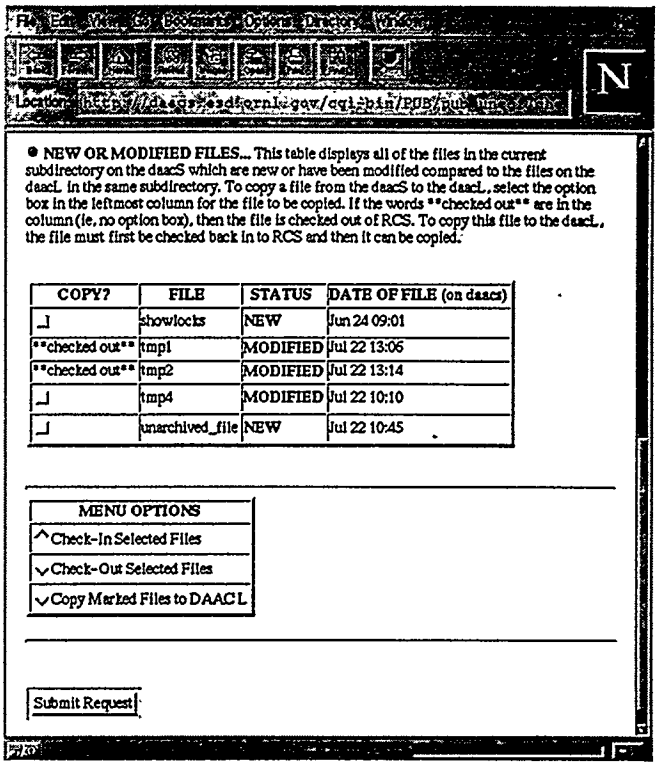

Figure 6: DMS Remote Copy Utility

9. Future Plans

Over the next few years the DAAC will be incorporating vast numbers of new datasets. Within two years it is expected that our number of files could easily triple, quadruple, or more. At that point the search mechanism could be sorting through a quarter of a million files. BIOME was designed to accommodate additional files and search mechanisms as needed.

The ORNL DAAC is making plans to add search capabilities for near-line imagery files that will be stored in a mass-storage system. A browse capability will allow users to view images by generating a thumbnail .GIF image of a larger imagery file. The programs necessary to search and retrieve near-line data are currently being tested.

The web developers are also planning to create a database table of small .gif images that are reductions of the real images. These can be quickly sent to image-displayable browsers for viewing by the user. The user can then select the real image from the sample image. The real image can then be downloaded, if small, or packaged for FTP or hard media shipment if the file is large.

The browser-aware capability of BIOME will be utilized to determine if the user's browser can execute JAVA applets. Several applications of this technology are envisioned. Users will be able to outline a region on a map, rather than specifying a point. Or the BIOME server can create a JAVA applet on-the-fly which when downloaded and activated will graph the user's just-

downloaded tabular data. If the user's browser is VMRL capable, BIOME will create a 3-D graph that can be viewed from any angle.

10. Conclusion: Lessons Learned

In summation, BIOME provides WWW access to a large number of tabular and imagery datasets relating to ecological and environmental information. The challenge is to help users find data of interest to them from the hundreds of thousands of available data files without overwhelming them. BIOME allows individuals to easily search an otherwise bewildering array of data products and retrieve/order the data online in a simple and efficient manner.

In accomplishing this task the ORNL DAAC learned a number of lessons that are applicable to other search and order systems on the WWW. Foremost among these is that management of a database this size requires metadata. Metadata can be extremely human-intensive, especially where various types of data from disparate sources are involved. Managing the metadata requires a RDBMS, and dealing with large amounts of data requires a robust server.

Documentation is also very important. A search and order system is of little use if it can't provide at least an abstract describing the data to a potential user. It is also important that users retrieve companion files to the data, (e.g., documentation, software, etc.) to help them better understand and manipulate the information.

Good system design allows a server to optimize network response times, although it is more work in the design phase. Running a large-scale system on a server also requires full-time server administration. The System Administrator(s) should expect to be on 24-hour call and have pagers that notify them of any serious server problems. Provisions must also be made for orders that are too big to download. These options include additional FTP protocols, tape, and CD-ROM.

Most important, providing data to a world-wide community requires support for a wide range of skill and knowledge levels, as well as various browsers and platforms - including Unix, PC, and Mac. To accomplish this, the DAAC developed a browser-aware search-and-order system that allows users to begin their search at any one of five starting points. Users have the freedom to search for data in the way that best meets their needs using a system that is tailored to take advantage of the capabilities of the particular browser being used.

As in many fields, the challenge is not merely to accomplish the task, but to make it look easy. In this BIOME excels. In the words of a BIOME user, "I have been very impressed with the way in which you have made data accessible for retrieval. The query system is very user friendly and unbelievably easy to use."

11. References

1. Lemay, Laura, Teach Yourself Web Publishing with HTML in a Week, Sam's Publishing, 1995.

2. Lemay, Laura, Teach Yourself More Web Publishing with HTML in a Week, Sam's Publishing, 1995.

12. Authors

Jon W. Grubb
Energy, Environment, and Resources Center
University of Tennessee
10521 Research Drive, Suite 100
Knoxville, TN 37996, USA
grubb@gandalf.rmt.utk.edu
Jon W. Grubb is a Research Associate at the University of Tennessee Energy, Environment, and Resources' Pellissippi Research Institute, and serves as the WWW Programming Specialist for the ORNL DAAC.

Sarah V. Jennings
Transportation Center
University of Tennessee
10521 Research Drive, Suite 200
Knoxville, TN 37996, USA
xqj@ornl.gov
Sarah V. Jennings is a Research Associate at the University of Tennessee Transportation Center's Pellissippi Research Institute and serves as WWW Curator and Documentation Specialist for the ORNL DAAC.

Teresa G. Yow, Ph.D.
Computational Physics and Engineering Division
Oak Ridge National Laboratory
P.O. Box 2008, M.S. 6407
Oak Ridge, TN 37831, USA
tgy@ornl.gov
Dr. Teresa Yow is a Systems Analyst in the Computational Physics and Engineering Division of ORNL. She is database designer and database administrator for the ORNL DAAC.

Anthony W. Smith
Computational Physics and Engineering Division
Oak Ridge National Laboratory
P.O. Box 2008, M.S. 6407
Oak Ridge, TN 37831, USA
axg@ornl.gov
Anthony Smith is a Systems Analyst in the Computational Physics and Engineering Division of ORNL. He serves as the User Interface Specialist for the ORNL DAAC.

* Research sponsored by NASA under Interagency Agreement DOE No. 2013-F044-A1 under Lockheed Martin Energy Research Corp., contract DE-AC05-96OR22464 with the U.S. Department of Energy.

"The submitted manuscript has been authorized by a contractor of the U.S. Government under contract No. DE-AC05-96OR22464. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes."

DISCLAIMER