Data Warehousing, Metadata, and the World Wide Web*

T. G. Yow, Ph.D.[1], and S. V. Jennings[2], A.W. Smith[1], J.W. Grubb[2], and P.F. Daugherty[1]
[1]Oak Ridge National Laboratory, [2]University of Tennessee
Post Office Box 2008, Mail Stop 6407, Oak Ridge, TN 37831, USA
telephone (423) 241-3952, fax (423) 574-4665, e-mail tgy@ornl.gov, xqj@ornl.gov, axg@ornl.gov, grubb@gandalf.rmt.utk.edu, and pvd@ornl.gov

Abstract

The connection between data warehousing and the metadata used to catalog and locate warehouse data is obvious, but what is the connection between data warehousing, metadata, and the World Wide Web (WWW)? Specifically, the WWW can be used to allow users to search metadata (data about the data) and retrieve data from a warehouse database. In addition, the Internet/Intranet can be used to manage the metadata in archive databases and to streamline the database administration functions of a large archive center.

The Oak Ridge National Laboratory's (ORNL's) Distributed Active Archive Center (DAAC) is a data archive and distribution center for the National Air and Space Administration's (NASA's) Earth Observing System Data and Information System (EOSDIS); the ORNL DAAC provides access to tabular and imagery datasets used in ecological and environmental research. To support this effort, we have taken advantage of the rather unique and user-friendly features of the WWW to (1) allow users to search for and download the data we archive and (2) provide DAAC developers with effective metadata and data management tools.

In particular, the ORNL DAAC has developed the Biogeochemical Information Ordering Management Environment (BIOME), a WWW search-and-order system, as well as a WWW-based database administrator's (DBA's) tool suite designed to assist the site's DBA in the management of archive metadata and databases and several other DBA functions that are essential to site management. This paper is a case study of how the ORNL DAAC uses the WWW to both manage data and allow access to its data warehouse.

Keywords

Database, metadata, data management, configuration control, data archive

* 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.

DISCLAIMER

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 the National Air and Space Administration's (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 its Mission to Planet Earth Program. The ORNL DAAC archives and distributes data relating to the Earth's biogeochemical dynamics. These data come from NASA-sponsored ground-based field investigations and include tabular data and imagery from satellite and aircraft sensors. Non-NASA biogeochemical data relevant to global change research are also included.

1.2 Setting Up the ORNL DAAC WWW Site

In 1994 the ORNL DAAC created a WWW site using a National Center for Supercomputing Applications (NCSA) httpd server and the Unix operating system on a Silicon Graphics workstation. 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 continue to serve as an access option. They are located at http://www-eosdis.ornl.gov.

As the number of datasets and users grew, we saw the need for a more sophisticated search-and-order system. 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 from the DAAC home page. Also, as the complexity of the archive increased, we saw the need for tools to automate site management. In 1996 we created WWW-based database management tool suite and several other customized utilities designed to help with the DBA functions at the DAAC. All DAAC utilities can be accessed via an Intranet DAAC Utilities Menu that is not accessible by the public.

The ORNL DAAC WWW site's customized search-and-order system and its customized management utilities use many generic Web features as well as advanced features that help users locate data quickly and easily and help site personnel manage the archive's data and metadata. The following subsections describe some of these features.

2 The Warehouse Described

As a data repository for NASA's field investigations, the ORNL DAAC catalogues, archives, and distributes data to users all over the world. The data from each project

archived at the DAAC is organized into "datasets," or groups of related experimental results. Each data file in a dataset is called a "granule." At the ORNL DAAC there are currently over 200 datasets, some with as many as 9000 granules each for a total of over 60,000 granules. New datasets continue to arrive daily. In order to efficiently archive and distribute this data, the ORNL DAAC generates metadata to describe the data and stores the metadata in our Sybase DataBase Management System (DBMS) tables.

2.1 The Metadata Database

The database underlying BIOME and our X-based search-and-order systems is actually a series of metadata databases. The data itself (especially the satellite imagery data) is too large to fit into a database. Instead, the databases consist of metadata describing the data and its location.

Metadata for the data archived at the ORNL DAAC is stored in several Sybase databases. The databases are identical in structure but differ in content. There is an ingest database for initial entry of data, a developmental database that undergoes rigorous testing, a database used for a local operational search-and-order system, and a database used by a larger operational search-and-order system. Metadata progresses through the system of databases as it is ready.

2.2 The Data Archive

The data itself is stored on-line, off-line, and near-line. Small tabular datasets are stored on-line on spinning disk. CD-ROMs, tapes, and proprietary data are stored off-line. Larger datasets, i.e., satellite imagery, is stored near-line in a mass storage system consisting of tapes on a Storage-Tek silo using Unitree software and Application Programmer Interfaces (APIs).

The location of the data is transparent to the user in that the user does not know where the data is stored. The near-line data storage system uses smart staging so the access time varies from seconds to minutes depending upon when that data was last accessed. It can be negligible enough that the data appears to be on-line. For off-line data the user is given a different set of delivery options. From the user's perspective, the data is all orderable on-line, with the majority of it also being deliverable on-line. With the exceptions of hard media (e.g., CD-ROMs) all data delivery is automated.

3 BIOME

The ORNL DAAC's search-and-order system, BIOME, uses a customized WWW interface that overlays the site's Sybase metadata databases. The customized WWW interface provides a user-friendly Graphical User Interface (GUI). Users can browse data holdings by scrolling pick lists generated from the metadata databases; users may then specify selection criteria including spatial or temporal coverage, geophysical parameters,

or dataset attributes like dataset ID or principal investigator name. We currently use httpd version 1.4 on our operational server and version 1.5 on our developmental server. We use Sybase version 10.0.1, IRIX version 6.2, and Perl version 5.0.

BIOME has several customized features that make data retrieval fast and efficient:

- 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 can get pages with frames, tables, and Java applets in addition to the information available to character-based browsers such as Lynx.
- User search methodology. The site does not dictate the starting point of the user's search. 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 amount of data requires metadata, which is stored in a relational database management system (RDBMS).
- 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.
- Supports viewing images.

4 WWW Tools

The data holdings of the ORNL DAAC have grown very rapidly over the past few years; we quickly realized that we needed help managing and distributing the data as well as help performing archive DBA functions. In response to our needs, DAAC programmers have developed several interfaces and utilities that have turned us into a cutting-edge WWW site.

These include database management tools, a directory management system, and an ingest status board. All the tools are behind a firewall, protected by both IP address and password. By using IP address protection of the computer firewall and httpd verification, access is allowed only to those who are logging in from a computer that has an IP address (e.g., 204.120.68.9 or fake.ns1.gwi.net) that has been preapproved for access by the system administrator. Once access to the computer has been granted, access to the main menu is password protected. Even after the main menu has been reached, various tools (e.g., the database management tool) have their own passwords, which are tightly controlled. Thus, only persons who have the need and capability to use the tools are allowed access to those tools. However, those persons with the need, the capability, and the correct passwords and IP address approval have the luxury of accessing the system from virtually anywhere.

4.1 WWW-Based Database Management Tools

4.1.1 DBA Needs

As the complexity of the data holdings has increased, the task of maintaining the databases has become increasingly difficult and time-consuming. We needed a customized tool that would perform several key functions for the DBA. First, we needed a tool that would ease the repetition of making the same changes to multiple databases (e.g., development and operational databases). We had to consider the possibility of global changes in all tables, in all databases, as well as the little changes that might affect tables in some databases but not others.

Fortunately, the Web-based DBA maintenance tool suite provides options that make the task of the database administrator less difficult. This tool is a GUI interface that uses HTML 3.2, cgi and perl scripts, and C processes executed by the cgi scripts to access the databases and perform database functions using Sybase's DBLibrary. The tool is accessed via the DAAC Utilities Menu.

Because Sybase offers little in the way of interfaces, the DBA tool suite provides a GUI interface that "visually connects" the DBA and the databases. We chose to create the GUI on the WWW for several reasons. First, a GUI WWW browser would provide a friendly interface that could reduce the tedium of database maintenance to the ease of clicking a button. Second, using the WWW would allow DBA functionality from any platform running a GUI Web browser and having access to the Internet and appropriate permissions.

4.1.2 DBA Tool Suite Options

What makes this interface so useful and robust is the design options that have been custom-built and implemented. For example, the DBA tool suite 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 manual construction of Sybase bulk copy files using a text 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.

Another important and useful feature of the DBA Tool is promotion of metadata from one database to another. Metadata is first ingested into a staging database and is then promoted from the staging database to the development database and later to the DAAC's operational database. The DBA tool allows the DBA to promote data from the staging database to any other database with only a few mouse clicks. The tool uses Sybase system

tables to determine the names of tables in the databases of interest and promotes from one database to another by appending records from a table in one database to a comparable table in the next or by replacing the contents of one table with the contents of another. Appends or replaces can be done table by table or to all tables at once.

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 backups, table creation options, and the granting of user privileges.

4.2 WWW-Based Archive Management Utilities

Managing a large data warehouse requires a specialized set of tools to manage the files, directories, and software that provides the archive's infrastructure. Configuration management and other DBA functions can be a nightmare without these tools. These utilities, which are accessed from the DAAC Utilities Menu via the Intranet, are described below.

4.2.1 The Directory Management System

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. We chose the Revision Control System (RCS) as the tool for archiving and managing these documents.

Our Web interface to RCS is the Directory Management System (DMS). This WWW interface makes this system easy to use by those who are not familiar with RCS and its UNIX commands. Furthermore, the information provided by RCS is arranged in a clear and concise layout on a single HTML page rather than appearing as a series of UNIX and RCS commands at the prompt. Examples of the information and the RCS functionality DMS provides the user are the capability to (1) quickly view all of the files that 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.

Another function of the DMS is to copy documents from one UNIX machine to another. This is important to the DAAC because all WWW development occurs on the development machine; then when the software is ready, it is moved to the operational machine. The DMS provides the mechanism for determining which files on the development server are new and which ones have been modified compared to the files on 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 confusion over which files in the development area have been modified but not yet upgraded to operational status.

4.2.2 The Ingest Status Board

One of the newest and most useful of our warehouse utilities is an electronic whiteboard that automates keeping track of ORNL DAAC datasets as they are being processed through the system. Replacing the old-style hallway whiteboard with its magic markers and hand-drawn status boxes, the electronic whiteboard offers a WWW GUI interface that allows team members to add the names of new datasets that are being ingested and to make updates to the status of datasets that are moving through the system.

Users first access our Intranet utilities menu screen and select the status board utility; they then enter their passwords. Once inside the utility, they select datasets of interest and update the status of datasets (for example, indicating that metadata processing for a dataset was completed on a certain date or that data was successfully ingested into the development database on a particular date). This tool allows all DAAC systems people to see at a glance the current status of all in-house work.

5 Conclusion

The ORNL DAAC provides WWW access to a large number of tabular and imagery datasets relating to ecological and environmental information. The ORNL DAAC has accomplished this task by designing and offering a customized WWW search and order system that allows efficient and rapid data search and retrieval. To manage the metadata and documentation that supports this data archive, DAAC team members have also developed several WWW data management and configuration control utilities.

By developing customized WWW tools to manage global ecological and environmental data, the ORNL DAAC has made an important contribution to NASA's Mission to Planet Earth Program. By staying on the cutting edge of WWW technology, the ORNL DAAC remains an important player in this program.

6 Biography

Dr. Teresa G. 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.

Mrs. Sarah V. Jennings is a Research Associate at the University of Tennessee's Pellissippi Research Institute and serves as WWW Curator and Documentation Specialist for the ORNL DAAC.

Mr. Anthony W. Smith is a Systems Analyst in the Computational Physics and Engineering Division of ORNL. He is a computer programmer for the ORNL DAAC.

Mr. Jon W. Grubb is a Research Associate at the University of Tennessee's Pellissippi Research Institute and serves as WWW Programmer for the ORNL DAAC.

Ms. Patricia F. Daugherty is a Environmental Sciences Division of ORNL. She serves as the Systems Engineer for the ORNL DAAC.