Services for Distributed Online Data Repositories

Helen Conover, Ken Keiser, Matt Smith, Marilyn Drewry, Sara Graves
Information Technology and Systems Center
University of Alabama in Huntsville
Huntsville, AL 35899

Abstract- Data services for distributed online data repositories are being provided for users of the DISCOVER Data Pool. Technologies that facilitate a centralized data ordering system, but use services at distributed repositories to fulfill orders and provide the requested products from the Data Pool will be described. Services provided include subsetting, visualization, order packaging and tracking. The repository components of the Data Pool system are being hardened with the intention of being deployable at additional sites to provide online data services.

Introduction

Distributed **I**nformation **S**ervices for **C**limate and **O**cean Products and **V**isualizations for **E**arth **R**esearch (DISCOVER) is a REASoN (Research, Education and Applications Solution Network) project, being lead by Dr. Frank Wentz of Remote Sensing Systems (RSS), with the University of Alabama in Huntsville (UAH) performing the information technology research components. The core mission of DISCOVER is to provide highly accurate, long-term climate data records and near-real-time ocean products suitable for the most demanding Earth research applications via easy-to-use display and data access tools. These products are derived from a large network of satellite microwave sensors going back to 1979. Most of the products are produced in near real-time (3-12 hours) on a 24x7 basis and hence are also suitable for some weather applications. The products include sea-surface temperature and wind, air temperature, atmospheric water vapor, cloud water, and rain rate.

The Information Technology and Systems Center (ITSC) and the Global Hydrology Resource Center (GHRC) at the University of Alabama in Huntsville (UAH), are conducting information technology research efforts for DISCOVER to explore new technologies and approaches for managing distributed online repositories of scientific data. Specific information technology goals of DISCOVER include:

- On-line services for data access and visualization
- Interoperability technologies for improved usability
- Flexible architecture to adapt to changing user requirements
- Exploring new technologies
- Integrating them into the DISCOVER information system
- Hardening selected tools and making them available to the wider community

Experience and capabilities from the DISCOVER IT effort has subsequently been employed in a similar effort for the Southeastern Universities Research Association (SURA) Coastal Ocean Observing and Prediction (SCOOP) program. Activities for the SCOOP Data Catalog are discussed in this paper as an example of how the technology developed for this REASoN can be extended to other projects and domains.

Data Management Objectives

The technology research area targeted is a distributed service architecture for custom data processing and will include both modular software components, and the basic semantic representations of these services necessary to chain them together to perform user-specified tasks. These technologies will allow the DISCOVER team to “mix and match” services in a variety of specialized user applications that will minimize the barriers to access and use of DISCOVER data and information products. This effort demonstrates the open, distributed, heterogeneous data system envisioned in NASA’s Earth Science Data Systems Working Group (ES DSWG) and contributes to the evolution of distributed national systems for on-line data distribution.

Services Approach

Online data repositories are rapidly becoming available for larger amounts of scientific data. Through the use of web services it is possible to automate the management and distribution of these data holdings, making them more easily accessible by users and distributed applications. The GHRC Data Pool is providing multiple interfaces to the same data repositories through web services and applications. For instance, the same data is available through regular HTTP/FTP access, OPeNDAP/DODS, Open Geospatial Consortium (OGC) compliant Web Map Services (WMS), online search/order, and manual user services requests.

Online availability of large volumes of DISCOVER products, along with specialized tools and services, has encouraged users to retrieve data electronically. However, because of data volumes required, bandwidth limitations, or for other reasons, some users have continued to request data shipment on electronic media such as tape or CD. Orders from both the automated online delivery and the manual user services requests have been integrated through an automated Data Order Tracking System (DOTS) using web service interfaces that allow for common tracking functionality across applications. DOTS metrics are combined with FTP and web statistics to provide a full picture of data distribution and access activities across the DISCOVER project.

The use of a web services architecture has allowed the Data Pool to be implemented across heterogeneous computational platforms at distributed locations, and to be more fault tolerant to interruptions and failures of system components.

GHRC Data Pool

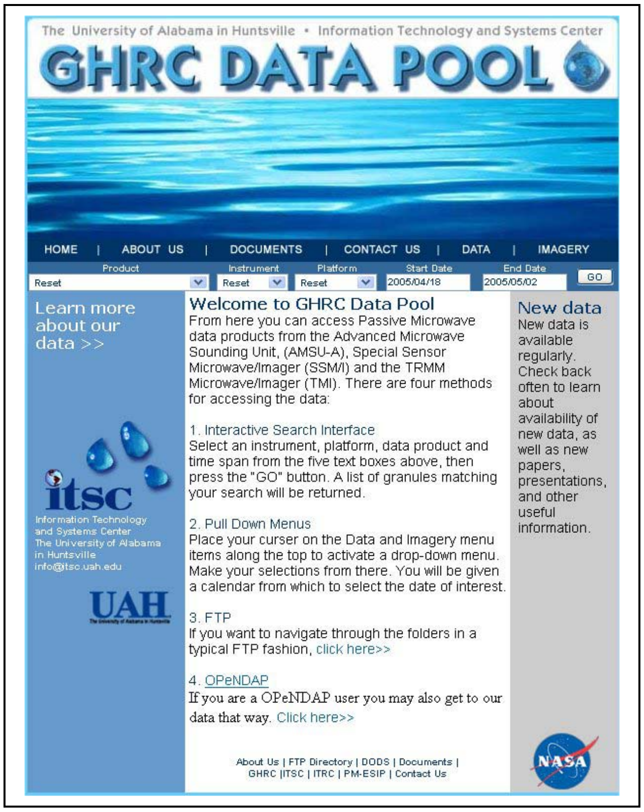
The GHRC Data Pool grew out of the need to provide more direct access to a growing amount of passive microwave data. As disk storage prices have decreased, online storage has become not only practical, but in some ways preferred. The availability of increased online data volumes naturally led to implementation of applications and services that could automate the search and order operations for this data.

A. Data Ingest through Catalog Services

As new data files are populated on the distributed repositories, ingest scripts are triggered to record the appropriate metadata about the new files in the GHRC data catalog. A layer of Catalog Services has been provided that supports remote update of the inventory information in the catalog. This approach is scaleable to support ingest functionality on multiple distributed repositories into a central data catalog.

B. Search and Order Interfaces

A search and order interface was developed (http://datapool.nsstc.nasa.gov) that allows users an interactive environment for searching, subsetting and packaging of available data products (see Fig. 1). For practical reasons the search and order interface was implemented on systems that had secure connections to the catalog database within the NASA network system at the GHRC.


The University of Alabama in Huntsville • Information Technology and Systems Center
GHRC DATA POOL
HOME | ABOUT US | DOCUMENTS | CONTACT US | DATA | IMAGERY
Product
Instrument
Platform
Start Date
End Date
GO
Learn more about our data >>
itsc
UAH
The University of Alabama in Huntsville
NASA
About Us | FTP Directory | DODS | Documents | GHRC |ITSC | ITRC | PM-ESIP | Contact Us

Product: Reset
Instrument: Reset
Platform: Reset
Start Date: 2005/04/18
End Date: 2005/05/02

Welcome to GHRC Data Pool

From here you can access Passive Microwave data products from the Advanced Microwave Sounding Unit, (AMSU-A), Special Sensor Microwave/Imager (SSM/I) and the TRMM Microwave/Imager (TMI). There are four methods for accessing the data:

1. Interactive Search Interface
Select an instrument, platform, data product and time span from the five text boxes above, then press the "GO" button. A list of granules matching your search will be returned.

2. Pull Down Menus
Place your curser on the Data and Imagery menu items along the top to activate a drop-down menu. Make your selections from there. You will be given a calendar from which to select the date of interest.

3. FTP
If you want to navigate through the folders in a typical FTP fashion, click here>>

4. OPeNDAP
If you are a OPeNDAP user you may also get to our data that way. Click here>>

Information Technology and Systems Center
The University of Alabama in Huntsville
info@itsc.uah.edu

New data
New data is available regularly. Check back often to learn about availability of new data, as well as new papers, presentations, and other useful information.

Fig. 1: Data Pool Search and Order User Interface

C. Order Tracking Services

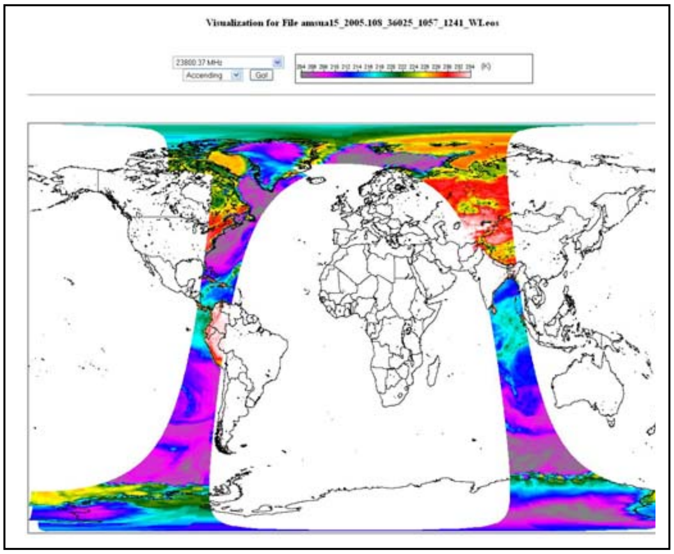
The data order tracking system (DOTS) is also on this same NASA network so web service interfaces were developed to provide distributed access to the system, from the various Data Pool components, for the purpose of creating new orders, requesting order information and updating order status.

D. Online Repository

The online data repository resides at distributed locations on the UAH campus and eventually other locations, such as RSS. In addition to fulfilling automated data orders, the repositories support additional data access protocols such as HTTP, FTP, and OPeNDAP (DODS). The repositories also maintain THREDDS (Thematic Realtime Environmental Distributed Data Services) catalogs of the data holdings in support of THREDDS-enabled applications such as the Integrated Data Viewer (IDV) available from Unidata.

E. Visualization Through Web Mapping Services

Interactive visualization of data files is being provided for users through the use of OGC-compliant web mapping services (WMS). Using file location information from the GHRC data catalog, access and display of selected data files is fast and provides an effective mechanism for users to review the coverage of a particular file (see Fig. 2) prior to completing an order and receiving the data. WMS functionality is available for each of the swath data collections currently available in the Data Pool. Support for gridded data products will be added as resources are available. An additional benefit is that the web mapping services are also publicly available for other applications and users to incorporate into other OGC-compliant applications.



Visualization for File amsua15_2005.108_36025_1057_1241_WI.eos

23800.37 MHz

Ascending

Fig. 2: Visualization of Data File with Web Mapping Service

F. Order Broker

Currently the concept of an Order Broker is tightly integrated with the Data Pool application, but in the general sense this functionality could be served by other remote components that are uncoupled from any particular application. Once the Data Pool application creates a new order in DOTS, that order is available to be claimed by a broker able to negotiate order fulfillment with an appropriate packaging component. As an order is claimed by the broker the order status is updated to indicate the packaging of the order is in progress. A broker, in this sense can be an automated process, as in the case of most Data Pool orders, or can be a human-in-the-loop using a separate user interface application to claim and fill an order. Manual monitoring of the tracking system is triggered by an appropriate status update when problems are encountered during the course of filling an order.

G. Packaging Services

Packaging of automated orders is most efficiently performed on a resource close to the repository's disk storage, in other words, with direct disk access. An order broker places waiting orders in a queue and a scheduled job periodically checks the queue for the next order waiting to be filled. Upon successful completion of the packaging job, the broker receives notification, sends an email confirmation to the user who placed the order and updates the order status in the order tracking system.

SCOOP Data Catalog and Archive System

The SURA Coastal Ocean Observing and Prediction (SCOOP) program is developing data management concepts for the NOAA coastal modeling community, in specific, and for the larger NOAA oceanographic community in general. ITSC is leading the SCOOP data management effort and has extended many of the lessons-learned from DISCOVER to that project by refining data transport approaches and incorporating, early on, the use of distributed services for catalog and inventory interactions.

A. Data Transport to Users and Archives

SCOOP is directly supporting the near real-time execution of coastal models so the data management system is required to move data around to the appropriate partners quickly and efficiently. The data transport mechanism for the initial version of the SCOOP system is built around Unidata's Local Data Manager (LDM). The LDM system includes network client and server programs and their shared protocols. Coastal modelers require input data as soon as it is available, so the "push" philosophy of LDM fits well as the data is transported as soon as the data producer makes it available to the product queue. Since all of the SCOOP data is being transported via LDM and the archives are remote to the modelers, the decision was to add the distributed data archives as listeners to these streams and they receive the data to be archived at the same time as the other "down-stream" users receive the data. The archived data is important for retrospective studies where modelers recreate specific runs, most likely with slight modifications for comparison.

B. Catalog Services

SCOOP is deploying a centralized data catalog that contains metadata about data collections and archive inventories. The SCOOP data producers (modelers) and the archives are all distributed participants so access and manipulation of the data catalog is facilitated through a layer of catalog services. These services support creation of new metadata records, input of ingest inventory information and queries on collections and inventories. The standardization of these services supports the development of custom applications by other SCOOP participants, as well as allowing query access from outside of the SCOOP community.

Conclusions

The use of services to support distributed online repositories is proving to be an effective data management approach. Web services insulate customers and applications from the implementation details of the underlying system, including programming languages and hardware resources. Changes to the service implementations will not ripple to the calling applications and users as long as the previously defined interface remains constant. Standardization and publication of service interfaces allows systems to scale easily to include multiple distributed resources.

Future Directions

The DISCOVER project will seek to incorporate additional repositories into the GHRC Data Pool system and provide additional services, such as advanced data analysis. As interest arises from other projects, the UAH research team would like to explore the adaptation of these approaches to other systems and domains.