DISCOVERing **Online Data and Services**

Sara Graves, Helen Conover, Ken Keiser
Information Technology and Systems Center
The University of Alabama in Huntsville
Huntsville, AL 35899

Abstract – **The DISCOVER project provides online services for data access, ordering and visualization of the project's data products and information via a Data Pool distributed across multiple institutions. Some key aspects of the Data Pool design include distributed and heterogeneous data servers, centralized databases for metadata management and order tracking, and attention to interoperability standards. The project also utilizes several important technologies such as data mining for meteorological events, the Earth Science Markup Language for data/service interoperability, the OPeNDAP data transport mechanism, and OpenGIS Web Services.**

I INTRODUCTION

DISCOVER (Distributed Information Services for Climate and Ocean Products and Visualizations for Earth Research) is a NASA Earth Science REASoN (Research, Education and Applications Solutions Network) project. The primary objective of the DISCOVER project is to provide highly accurate, long-term ocean and climate products suitable for the most demanding Earth research applications via easy-to-use display and data access tools. The products include brightness temperatures from a variety of passive microwave satellite instruments, as well as such derived products as sea-surface temperature and wind, air temperature, atmospheric water vapor, cloud water and rain rate. A key element of DISCOVER is the merging of multiple sensors from multiple platforms into geophysical data sets consistent in both space and time. DISCOVER is a collaboration of Remote Sensing Systems (RSS), NASA/Marshall Space Flight Center and the University of Alabama in Huntsville (UAH).

The information technology focus of DISCOVER is on providing online services for data access, ordering and visualization of the project's data products and information. The DISCOVER Data Pool approach automates the ordering, visualization, packaging and delivery of scientific data from multiple online repositories. The Data Pool offers a variety of data access and interoperability technologies for improved usability. The distributed and heterogeneous computing environments of the project team combined with the distributed services technology approach provide an excellent arena in which to explore emerging data delivery and interoperability paradigms.

II DISTRIBUTED DATA POOL

A key component of the DISCOVER infrastructure is its distributed Data Pool. This collaborative project has online repositories for public data access at multiple sites, but location is transparent to users. In addition to the Data Pool on the DISCOVER web site, these data sets are listed in the Global Change Master Directory, EOS Data Gateway, Federation Interactive Network for Discovery, and Unidata THREDDS (Thematic Real-time Environmental Distributed Data Services) data registries.

A. *ESML and Standard Data Formats*

Standard, self-describing data formats provide the benefits of common data access methods and preprocessing tools, interoperability with common analysis packages, integrated data and documentation, and facilitation and integration of diverse data products. However, different user communities often prefer different formats, and many find simple ASCII and binary data structures easy to use and understand. Space efficiencies gained through the use of such simple formats can also be important, especially for users working with time series. DISCOVER is addressing these conflicting user needs, as well as providing interoperability solutions, by offering the data in a primary format preferred by the majority of the DISCOVER user community, and providing tools to translate the data into any of several additional standard formats. This flexible approach is based on the Earth Science Markup Language (ESML), developed for NASA by UAH [http://esml.itsc.uah.edu]. ESML description files contain standardized descriptions of the content, structures, and semantics of a particular set of data files [1, 2]. DISCOVER will provide ESML descriptions of the primary data format for each of the different data products. ESML-based "transcoder" services will be developed to supply services such as format conversion and subsetting. These transcoders will be integrated into the various DISCOVER data distribution mechanisms, so that users' data format preferences can be supported easily.

B. *Integrated Data Services*

The Data Pool's user interface integrates data services into the data selection and retrieval process. Because all DISCOVER data sets are derived from polar orbiting satellites with essentially global coverage, search fields include keywords (geophysical parameter, instrument, and satellite) and temporal constraints, but not spatial constraints. However, geographic subset selection [3] is an integral component of the data selection workflow. After users select search and subsetting constraints, they are provided with a resulting list of data files, grouped temporally. Each file name is a link to the data file, so users may download the data directly. Alternatively, they may select individual files, temporal groups, or the entire results set to be (optionally) subsetted and packaged for more efficient delivery. Current packaging options include various compression and bundling utilities such as "tar" and "gzip." The DISCOVER team will plans to integrate additional subsetting, reformatting, and other transcoder services into the Data Pool, providing for greater usability and a variety of data delivery options. In addition, electronic data subscriptions will be offered, that push data to a user's machine as soon as it is available.

C. *Information Products*

The DISCOVER web site [http://discover-earth.org] provides access to a 5-year record of global atmospheric temperatures from the AMSU-A instrument, together with 20-year average temperatures, record highs, and record lows. This feature allows users to visually compare current temperature trends with the historical record. Another popular feature provides current information on tropical cyclones and hurricanes, also derived from AMSU-A data [4]. The data are mined in near real-time, as part of the product generation processing stream. The technique is based upon the close relationship between the cyclone warm core in the middle and upper troposphere and surface low pressure (and thus surface wind speeds). Where cyclone warm cores are identified, maximum sustained wind speeds are calculated, and information on these detected storms is cross-correlated with National Hurricane Center information. The interactive web site provides a synopsis of the characteristics of each cyclone detected, annotated images, and storm track information. In addition, the storm location and parameters are captured in a database for further analysis.

III Data Access And Visualization Using ESML-supported Services

The DISCOVER Data Pool provides multiple methods to access and visualize the available data products. These include conventional HTTP and FTP access, as well as services that provide for enhanced usability and interoperability, such as GridFTP, OPeNDAP, and OpenGIS-compliant web mapping and coverage services. GridFTP is an enhanced FTP capability, developed for the evolving Grid community, that provides improved data transfer rates. DISCOVER will provide GridFTP capabilities in support of users and applications from ongoing Grid research projects such as the Information Power Grid (IPG), Linked Environments for Atmospheric Discovery (LEAD) and others [5].

The DISCOVER team is incorporating ESML technology with some of these conventional services to build more flexible applications that can support the distributed and heterogeneous nature of the DISCOVER data products. As these ESML-based tools are hardened for general use, they will be made available to the scientific community.

A. *ESML-OPeNDAP Data Server*

DISCOVER online repositories not only provide a large amount of data for users to retrieve, but also facilitate the use of applications capable of accessing data dynamically across the network. A prime example is OPeNDAP (Open Source Project for a Network Data Access Protocol), a software framework that allows local data to be made accessible to remote locations independent of local storage format [6]. DISCOVER is providing OPeNDAP access to the various datasets in its Data Pool. Typical OPeNDAP servers use data set-specific modules to provide client applications access to distributed data sets, as well as let users browse the OPeNDAP-provided information across the web. UAH has recently prototyped an OPeNDAP server that uses ESML technology to serve heterogeneous data sets through a single OPeNDAP server module. This functionality is currently in beta testing. This is a practical integration of technologies, with ESML providing access to multiple data formats and OPeNDAP providing distributed access and delivery. During the first year of the DISCOVER project, we will further refine and test this functionality. DISCOVER will continue to update the ESML-OPeNDAP services as new functionality becomes available, and will extend the services to new data products resulting from ongoing research.

B. *ESML-OGC Data and Information Servers*

A new feature planned for the DISCOVER web site is a set of tools that allows users to visually overlay microwave data and other information products as merged images, like that shown in Fig. 1. Users will be able to select data layers, geographic areas, and time periods for display. Data layers will include the DISCOVER data products described in Section I, as well as mined information and events such as cyclones discussed in Section II.C. These tools will use OpenGIS Web Services (OWS) protocols [7] to request and display the data layers. OpenGIS is an international consortium of industry, government agencies and educational institutions striving to standardize interfaces and protocols for using geospatial data. DISCOVER's use of these standard protocols will assure that the data layers conform to national data access standards, and can be merged with images from other sources. This interoperability will facilitate the flow of DISCOVER products to the research and applications communities for modeling, decision support, and educational development. In concert with the ESIP Federation, the DISCOVER team has done preliminary work with OWS

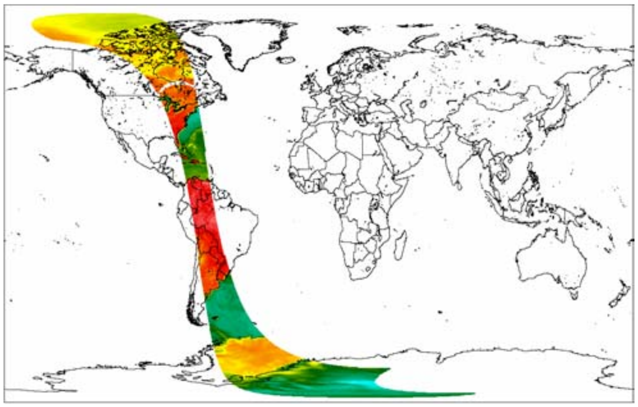

Figure 1: DISCOVER Overlaid Web Mapping Images for SSM/I Swath Data and World Boundaries

Web Map Services, which return the data images that can be overlaid or merged with other data layers to create composite maps.

Web Coverage Services differ from mapping services in that they return actual data in response to a standardized request protocol, rather than just an image of the data. The web coverage service (WCS) specification is still evolving within the OpenGIS community and the DISCOVER team, through UAH's OpenGIS Associate Membership, will participate in further development and refinement of the WCS specification. The DISCOVER team is developing WMS and WCS services layered on ESML technology that will allow distributed data repositories to provide these services for heterogeneous data sets with limited special processing. Web Feature Services (WFS), which can be used to provide detailed information on a specific feature within an image or data set, are also being explored. For example, if WMS is used to overlay cyclone locations over a sea surface temperature image, WFS can be used to provide maximum sustained wind speed information for the cyclone.

IV Managing Data and Orders across Distributed Repositories

In order to provide seamless access to data at multiple sites, as well as to integrate these data products into national systems for data search and online distribution, such as the EOS data and service catalogs and Geo-Spatial One-Stop initiative, DISCOVER provides catalog services for all data products at the various data server locations. Under development is an automated metadata generation tool that crawls the online data repositories regularly to dynamically update the DISCOVER database with information about newly generated data files. This tool uses knowledge about the directory structures and file naming conventions to derive basic metadata such as product type, date, and location. The database maintained in this fashion can support data searches across the DISCOVER project, and provide interoperability with EOSDIS and other distributed data search services. The DISCOVER Data Pool provides two complementary methods of data discovery: users may select data products from the catalog with a keyword search, or browse the file system and access data or images via a calendar interface.

For efficient handling of data orders across distributed repositories, DISCOVER implements distributed data processing services on the file servers where the data resides. These distributed services are managed through centralized functionality for catalog services and order tracking. Centralized and automated order tracking services allow for the use of automated order processing while at the same time tracking information for metrics and possible user service monitoring in the case of system malfunctions. DISCOVER is providing centralized catalog and order tracking services that are accessible from data processing services distributed at the data repositories. Figure 2 illustrates this association between processing at distributed repositories and a centralized Data Pool. Services at the distributed repositories, such as ingest, web mapping, subsetting and packaging are not required to be homogeneous implementations as long as they communicate processing information back to the catalog and order tracking services at the Data Pool. This allows a repository to have specialized processing for a particular type of data but communication with the Data Pool through the standardized service interfaces provides transparency across the heterogeneous repositories.

V Software Reuse

DISCOVER's distributed architecture will accommodate the introduction of plug-and-play components that can be enhanced independently through upgrades or replacement. Selected Earth science software components will be available to others in the science community. In particular, ESML and the HDF-EOS Subsetting Engine (HSE) are already being reused by other Earth science software system development

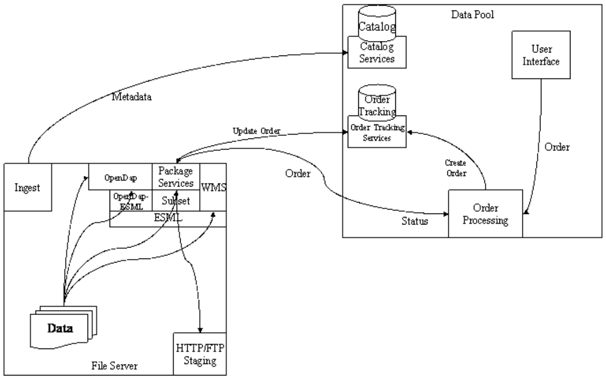

Figure 2: Data Pool Architecture

teams. HSE, a dataset-independent subsetting service for HDF-EOS data, provides robust, operational subsetting software, available for use within EOSDIS and by the science community [http://www.subset.org/downloads/sds-hse.html]. In addition to its use in DISCOVER, HSE has been incorporated into the EOSDIS Core System (ECS) for use in DAAC order filling operations, the AMSR-E Science Investigator-led Processing System (SIPS), the ESDIS Data Pools, and as a standalone service by the MODIS Land team at NASA Goddard. The ESML software library, used by applications to decode data files accompanied by ESML descriptions, is especially well suited for open-source access since it is intended for use and acceptance across the Earth science community and extension through community participation [http://cvs.sourceforge.net/viewcvs.py/esml/]. Additionally, ESML promotes the goal of interoperability in a federated science environment.

ACKNOWLEDGMENT

The authors would like to thank our DISCOVER project partners, Frank Wentz and colleagues at RSS, Michael Goodman of NASA/MSFC, and Roy Spencer of UAH.

REFERENCES

[1] R. Ramachandran, M. Alshayeb, B. Beaumont, H. Conover, S. Graves, X. Li, S. Movva, A. McDowell, M. Smith, "Earth Science Markup Language: a solution for generic access to heterogeneous data sets", NASA Earth Science Technology Conference 2001, 28 August 2001.

[2] R. Ramachandran, H. Conover, S. Movva, S. Graves, "Using ESML in a Semantic Web approach for improved Earth science data usability", SCISW2003 Workshop, Sanibel Island, Florida, Oct. 20, 2003.

[3] S. Graves, B. Beaumont, M. Smith, "ECS-HSA: the HEW subsetting appliance", HDF-EOS Workshop VII, Silver Spring, MD, Sep. 23 - 25, 2003.

[4] R. W. Spencer, W. D. Braswell, "Atlantic Tropical cyclone monitoring with AMSU-A: estimation of maximum sustained wind speeds," *Mon. Wea. Rev.*, vol. 129, pp. 1518-1532, 2001.

[5] K. Droegemeier, V. Chandrasekar, R. Clark, D. Gannon, S. Graves, E. Joseph, M. Ramamurthy, R. Wilhelmson, K. Brewster, B. Domenico, T. Leyton, V. Morris, D. Murray, B. Plale, R. Ramachandran, D. Reed, J. Rushing, D. Weber, A. Wilson, M. Xue, and S. Yalda, Linked environments for atmospheric discovery (LEAD): A cyberinfrastructure for mesocyclone meteorology research and education, AMS Interactive Information and Processing Systems (IIPS), Seattle, WA, 2004.

[6] P. Cornillon, J. Gallagher, T. Sgouros, “OPeNDAP: accessing data in a distributed, heterogeneous environment,” *Data Science Journal*, vol. 2, pp. 164-175, 5 November 2003.

[7] OGC Web Map Services specification version 1.1.1, http://www.opengis.org/docs/01-068r2.pdf.