ECHO Responds to NASA's Earth Science User Community

Robin Pfister*, Richard Ullman*, and Keith Wichmann†

*Code 423, NASA/GSFC, Greenbelt, MD 20771
†Global Science & Technology, Inc., 6411 Ivy Lane, Suite 300, Greenbelt, MD 20770

ABSTRACT

Objective.

Over the past decade NASA has designed, built, evolved and operated the Earth Observing System Data and Information System (EOSDIS) Information Management System (IMS) in order to provide user access to NASA's Earth Science data holdings. During this time revolutionary advances in technology have driven changes in NASA's approach to providing an IMS service. This paper will describe NASA's strategic planning and approach to build and evolve the EOSDIS IMS and to serve the evolving needs of NASA's Earth Science community. It discusses the original strategic plan and how lessons learned help to form a new plan, a new approach and a new system. It discusses the original technologies and how they have evolved to today.

Significance.

NASA's initial approach of characterizing the user community as a whole rather than recognizing the existence of distinct sub-communities led to issues in the IMS design. This paper talks about a new strategy that is based on realities of diverse sub-community needs and limited resources.

The original and current user interface (EOS Data Gateway-EDG), based on the old strategy, has evolved to adequately support many users. However, the diverse nature of scientists, instrument specialists and the public, drives the need for different methods of finding data. NASA has adopted a new approach of externalizing EOSDIS metadata into a clearinghouse and providing a foundation for the interoperability of externally developed user interfaces and data services. The resulting system, called EOSDIS ClearingHOuse (ECHO), is based on providing a common, well-defined message level interface to the metadata. With ECHO's focus on supporting various searches of the metadata and brokering the subsequent orders for data, individual communities can tailor the user interface design to their own needs and access methods. This approach enables the community to participate in defining their user interfaces without focusing on the details of the underlying infrastructure. By allowing a variety of protocols to be used, ECHO allows existing systems to continue to work with only the addition of a protocol translation layer. New systems that are customized to the needs of a community can be developed to use either the native or any of several other protocols to communicate with ECHO's repository of current information.

Methods.

NASA's approach is to build a service broker and metadata clearinghouse on a flexible infrastructure that enables various facets of the community to use their own resources to develop functions that can interoperate. The infrastructure is based on XML and Java Beans and supports heritage as well as XML protocols for context sharing among distributed functions and systems.

ECHO, acting as a broker, has three classes of customers. The first customer is the set of providers that are willing to participate by providing metadata to ECHO. Here the goals are to make participation as painless as possible for the providers and to allow flexibility in their level of participation. The second customer is the set of client interface developers that are willing to participate by hosting "browsers" to ECHO metadata. The system provides a single place to gain access to a diverse collection of provider's data systems with a minimum of hassle. The simple message-based transaction system allows for relatively simple adapters to be built to connect existing client interfaces. The third customer is all of the people who use the client interfaces. The system will strive to satisfy these customers by keeping the data as current as possible (within the realms of what the provider's supply) and providing quick, stable access on a continuous basis.

Results.

The first phase of this system will be operational in December 2000. At that time NASA will gain better recognition

and understanding of results using this new approach. These findings will be discussed in the paper.

1. INTRODUCTION

The Earth Observing System Data and Information System (EOSDIS) is part of NASA's Earth Science Enterprise which is NASA's contribution to the U.S. Global Change Research Program. Global Change Research is an ambitious program that includes a large suite of long term measurements from a variety of sources including spacecraft, aircraft, and surface observations. Data collected are held in geographically distributed archives. The purpose of EOSDIS is to provide easy access to data and services so scientists can focus their energy on research. Over the past decade, NASA has been developing, operating and improving components of an Information Management System (IMS) to provide user access to these data. There have been many changes since the beginning of this effort. User needs have changed and information technology has been revolutionized. These changes have driven the evolution of the basic architecture and design of the IMS.

2. INITIAL INFORMATION MANAGEMENT SUBSYSTEM (IMS)

2.1. Initial User Characterization

As a multidisciplinary Earth science program, EOS brings experts in all fields of Earth science together to study global issues. Early in the EOSDIS project, assumptions about the nature of the user community were based on speculation rather than practice because no system of this nature previously existed. Our primary user community (NASA Principal Investigators) was generally understood to require similar services from IMS components. Sub-community specific needs were thought to be mostly limited to help-desk support targeted for particular data collections. NASA identified heritage archives that had expertise in the various sub-disciplines and designed a distributed information system around them. Local archives interacting directly with their users would meet specific needs.

2.2 Present System

Figure 1 illustrates the overall functional architecture of the EOSDIS in 2001. EOSDIS is an end-to-end ground data processing system that includes flight operations, data acquisition, data capture, initial processing and backup archival. Data are routed to the Distributed Active Archive Centers (DAACs) or Science Investigator-led Processing Systems (SIPSs) where data are further processed into higher level products. The processed data are archived in the DAACs and distributed to the end user community.

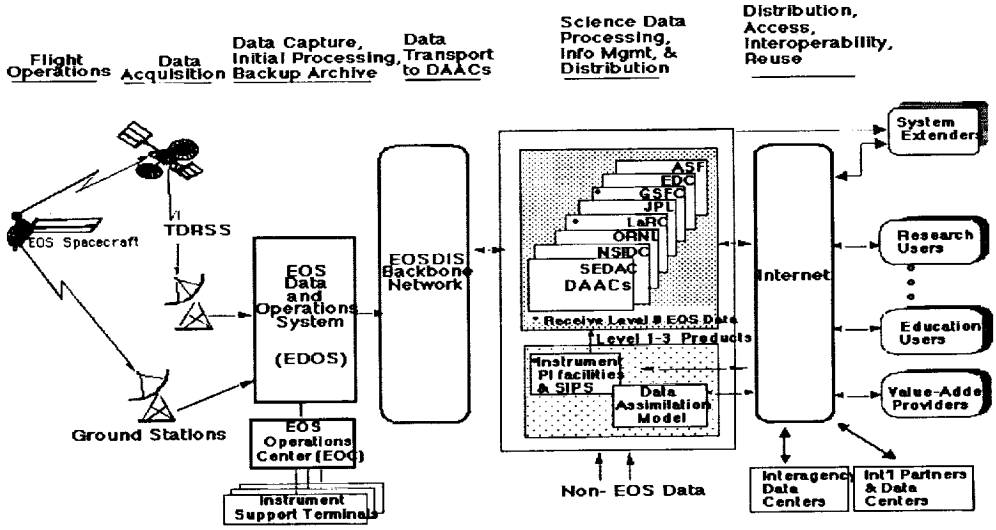

Figure 1. EOSDIS Functional Components.

The Information Management Subsystem (IMS) subsystem is in the area of "Access and Interoperability" on the right in figure 1. The IMS components that provide search and order of data among the distributed archives are based on a prototype that was initiated in 1991. The Version-0 (V0) prototype was a proof-of-concept for interoperability among distributed archives. The infrastructure was designed to provide data search, browse (viewing a sample image), and order functions to scientists in all Earth Science sub-disciplines. Over a 4-year effort, V0 was developed on top of the existing data and services at each DAAC. The V0 prototype succeeded in providing single-point access, through a common interface, across distributed data at the archives (Figure 2). During the development of V0, information technology was revolutionized. As technology evolved, the user community expected the user interface to be presented in the latest technology on their desktop. In the beginning, character-based user interfaces that would run on VT100 terminals were the standard. After a year into prototyping, X-Windows-based graphical user interfaces took over the market. At the operational release of the X-Windows-based GUI, html interfaces emerged and quickly took over the market. Just as users were getting used to the cumbersome click-and-wait interaction of html, Java emerged as a more interactive option. Our development team could barely develop basic functionality before it all needed to be redone with a new user interface technology.

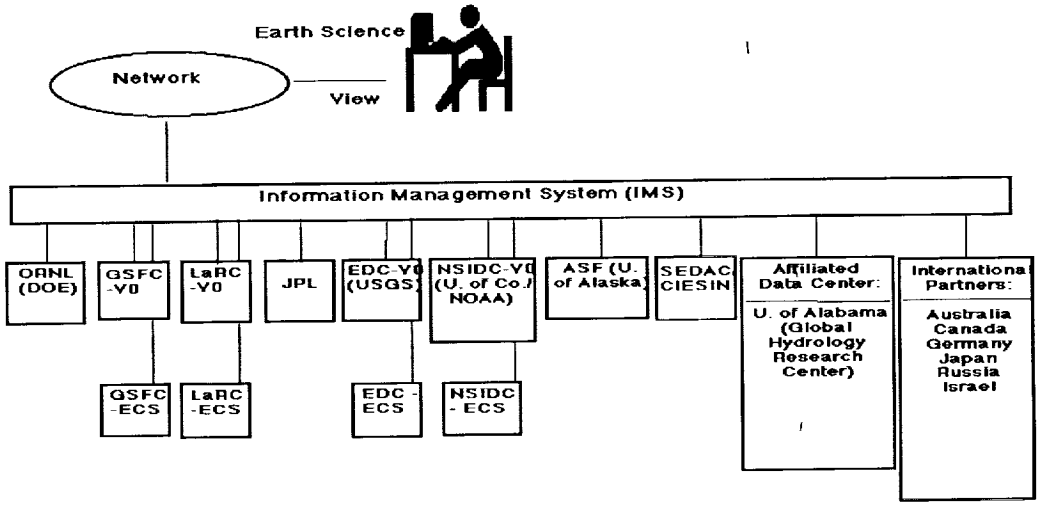

Figure 2. Version 0/EDG Interoperability.

Operational since August of 1994, the prototype, now called the EOS Data Gateway (EDG), has grown in capability, user base, and data provider base. In addition to the archives in the U.S., EOSDIS has interagency and international partners who supply data through interoperable links and collaborate on other EOSDIS activities. Today, EDG provides access to over 1000 datasets at 19 heritage, international and ECS data archives. The URL for EDG is http://eos.nasa.gov/imswelcome.

3. NEW ARCHITECTURE SUPPORTS VARIED DATA ACCESS PARADIGMS

3.1 Changes in User Needs and Community Characterization

Early in the EOSDIS specification process, scientists had limited experience with data search and retrieval systems so it was difficult for them to know and communicate their needs for data access. With the information technology revolution, information systems are pervasive in everyday life. Our user community has gained experience with search and order systems as Internet navigation and paradigms of e-commerce have become routine. Users have also become more familiar with the EDG interface to EOSDIS. Our community is now better able to communicate their refined preferences for data access. We have learned that different science disciplines sometimes desire very different ways to access data. The "one-size-fits-all" approach with the V0/EDG IMS cannot continue to satisfy our community's expectations.

To address these changing expectations, NASA is re-architecting the IMS and developing a component called ECHO (EOS ClearingHOuse) that will support varied paradigms needed by our users. Through the use of the eXtensible Markup Language (XML) and e-commerce concepts, ECHO will provide more flexibility for end users and strengthen the ability of discipline specific groups and data providers to serve particular sub-communities.

3.2 The EDG Architecture

The V0/EDG architecture (figure 3) is designed to support an iterative search and retrieval paradigm for data access. The system defines two tiers of metadata that describe data holdings. The most general level is a catalog of datasets; a small core of metadata used to describe each collection. Key attributes are dataset name, topic area, coverage extent and archive location. The more detailed tier is the inventory of granules (the smallest unit of data independently managed in the inventory) for every dataset. In the heritage and present EOSDIS, the granule is also the smallest unit of data that can be retrieved by the end-user. The EDG architecture reinforces the logical hierarchy of collection and inventory metadata by physical separation of the classes and their hierarchical application to the search process. Using catalog metadata, the EDG guides the user in construction of a search query. After it is constructed, the query is submitted by EDG in parallel to each of the interoperable data centers that are identified by the catalog as having granules that may result in a hit. The two-tier system is intended to reduce the occurrences of nonsensical queries. Even so, it is still possible to generate a query that results in too many or too few data items.

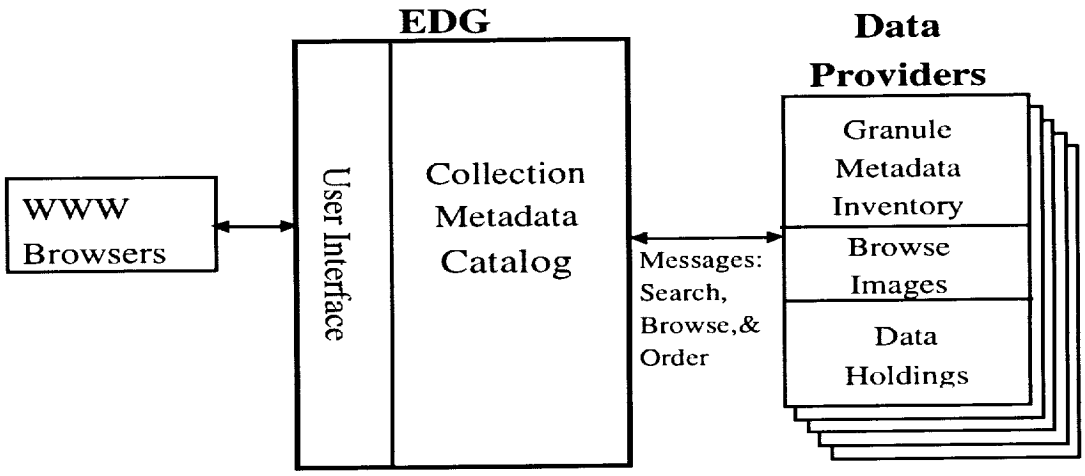

Figure 3. Present EDG Context Diagram.

Other shortcomings with the architecture have been identified. The query is performed in parallel among the distributed archive systems, the final result is not known until all systems have responded. The performance of the system is bound to the slowest member. If network traffic or systems problems affect one member archive, all queries are affected. If a member archive experiences an outage, access to that respective inventory is unavailable. The tight coupling of the EDG web server to the collection catalog middleware is another source of difficulty. The additions of new data collections to the catalog or new client/server functionality are unacceptably intertwined and it is difficult to provide alternate views through the single server. Equally important is the emergence of new ways of searching. The EDG concept is a first generation office automation system. The construction of queries based on catalog metadata is a familiar and a natural extension of traditional library catalog searches translated to a computer environment. Other data access paradigms made possible with computer technology rely on much richer metadata than is available at the catalog level. Many of our users prefer the paradigm of navigation and discovery. Rather than performing iterative queries, these users want the system to present the available data so they can simply navigate the options (presented by metadata) and discover desired data. Prototyping efforts have revealed that the navigation and discovery paradigm cannot be supported with the present architecture because network separation of the collection-level metadata from the inventory metadata makes such an application impractical.

3.3 The ECHO Architecture

The new ECHO architecture makes several important changes from the heritage EDG architecture. Two are most significant. First, the metadata holdings are no longer fragmented by the catalog and inventory hierarchy, all

metadata is brought into the clearinghouse. The second is that the strong coupling between the metadata holdings and the user interface is severed (see figures 3 and 4 for comparison). In its place is an abstraction layer consisting of an XML based message protocol.

Bringing all the metadata into a single clearinghouse solves the problems of network reliability. System performance for data access is dependent only on the ECHO system itself. Only after the user identifies the granules of choice, is the link to the data providers necessary. Even then, if the link is unavailable for any reason, ECHO will store the order and forward it when the data provider system is accessible. The integration of catalog and inventory metadata with the inclusion of browse imagery enables the new search paradigms discussed above. The client API layer permits varied clients to interact with the clearinghouse. The EDG will be retooled to take advantage of this interface to provide the traditional iterative query interface. New clients will provide navigation- and discovery-based views of the combined data holdings. NASA will encourage other parties to deploy clients tailored to the preferences and desires of specific sub-communities.

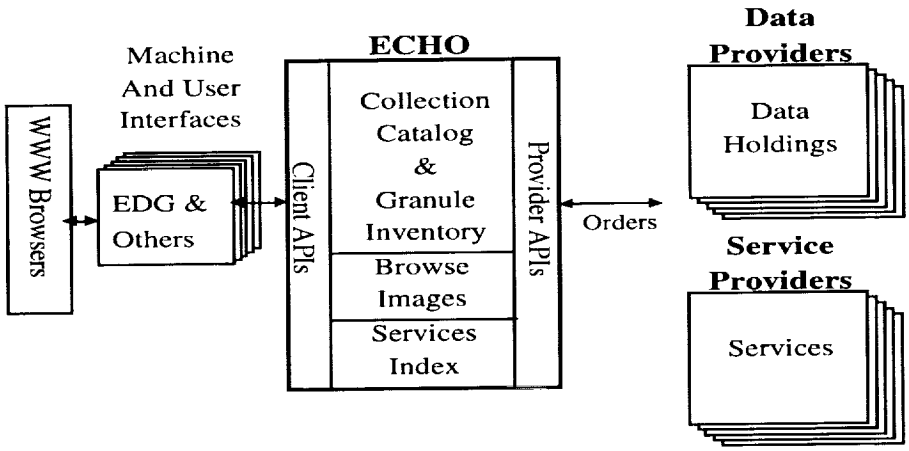

Figure 4. ECHO Context Diagram.

This new architecture offers other opportunities to streamline user processes. Currently, we see several data services such as subsetting and on-demand processing being developed independently and remotely from the data providers. Although, these services are intended to operate with the data in the provider archives, there is no mechanism in the current system to connect data to these value added services. The burden is on the user to find the data, to separately find the service, make the necessary connections to get the data transported to the service location, and finally apply the service to the data. ECHO will broker data services between data and service providers so the user is never aware of the data transport and service application details.

4 CONCLUSIONS

ECHO provides an infrastructure that allows diverse communities to share tools, services and metadata. As a service broker, ECHO decentralizes end user functionality and supports interoperability of distributed functions. As a metadata clearinghouse, it supports old iterative query data access and new navigation and discovery data access paradigm that serves to eliminate zero-hit and mega-hit results sets. A well documented, message based interface is provided instead of an integrated web server. This approach allows varied search providers to build their own user interfaces so they are not limited by the data search and order system provided by NASA. Users can search and find data regardless of provider down time. If the provider is still down when the user submits an order, ECHO will continue to attempt to submit the order on behalf of a user. For data providers, ECHO off-loads system resources required for searching. This new approach to the IMS offers providers the flexibility to support community-specific services and functionality that their users need.