REUSABLE SOFTWARE SERVICES FOR SCIENCE DATA SYSTEMS

Dan Crichton
Dan.Crichton@jpl.nasa.gov

Tom Fouser
Tom.Fouser@jpl.nasa.gov

Dana Freeborn
Dana.Freeborn@jpl.nasa.gov

Sean Hardman
Sean.Hardman@jpl.nasa.gov

Julie Wang
Julie.Wang@jpl.nasa.gov

Jet Propulsion Laboratory - Pasadena, California 91109-8099

ABSTRACT

A common science data processing software framework yields the benefits of reuse while remaining adaptable to address requirements that are unique to the mission. The Earth Science Data Systems section at the Jet Propulsion Laboratory (JPL) has been developing infrastructure services that allow for reuse and rapid development of new science data systems at relatively low cost, increasing the reliability of the deployed software system. The software framework, named Enterprise Data Management (EDM) Services, implements a set of services for data archiving, database hosting, application hosting, middleware and metadata management. These services enable projects to reuse services that are already provided by the institution in the construction of their applications. In addition to operating the framework using common information services, the EDM service components can be delivered to and operated by the mission to allow for scalability and tailoring to the needs of the mission. The SeaWinds Processing and Analysis Center (SeaPAC), for example, integrated the EDM Catalog and Archive Service with the SeaPAC ground data system to enable the process management function of the system. The Catalog and Archive Service was delivered to the SeaWinds project through a collaboration established with EDM and SeaWinds providing a database and system independent implementation that scales and adapts to meet the needs of the SeaWinds mission.

As the EDM framework matures, new services to support the construction of science data systems will be added to support small and medium size missions and experiments. Simple PI-led experiments will then be able to use science data system services without having to implement and run their own systems. This will alleviate scientists and mission specialists from having to reconstruct new data systems for each science mission while providing a common interface and operating environment for running science data system applications and services.

The EDM framework is an extension of the research and development work performed under the Object Oriented Data Technology (OODT) task at JPL. OODT has successfully integrated its software through collaborative efforts with the National Institutes of Health and NASA's Planetary Data System.

1. EDM SERVICES

The EDM framework of services focuses on filling the gap between systems by leveraging and providing services and interfaces necessary to interconnect data systems for creating an information architecture. EDM focuses on providing an application framework based on standard interfaces in an effort to breed high quality, interoperable enterprise applications and data systems. Enterprise data systems are commonly heterogeneous and in many cases provide little or no interoperability to share data. A key to linking enterprise data systems and databases is to provide a common metadata model. The services and interfaces illustrated below (Figure 1), exploit relationships and access between data systems in order to support hosting, location, access, retrieval, distribution, and exchange of data products within an organization's information enterprise.

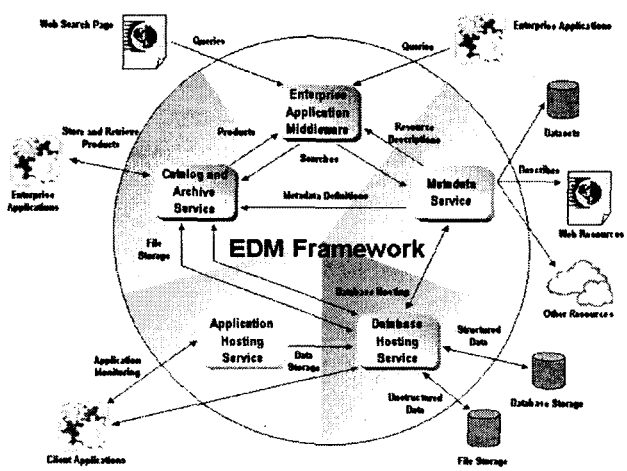

Figure 1. EDM Framework of Services

The EDM framework allows for the building of an information architecture that integrates data and information systems that span multiple disciplines and organizations. EDM provides a framework of services that concentrates on achieving the following goals:

- Provide cross-disciplinary data management support and services
- Increase interoperability of enterprise applications
- Enable data, information and knowledge management
- Increase software reuse
- Integrate data management services with an institutional information architecture
- Integrate data management services with an institutional security architecture

The framework consists of five services. The Enterprise Application Middleware, Metadata Service and Catalog and Archive Service are software-based services which can be operated at the JPL institutional level, or can be delivered to internal or external customers to be integrated in their environment. The Application Hosting and Database Hosting Services, on the other hand, are JPL institutional services and are only available to internal JPL customers.

1.1 Enterprise Application Middleware

This service will form the hub of information systems implementations. It will provide the interface between the application user interfaces (i.e. Portals) and the data systems and archives. This service is designed with the following features:

- Provide common application messaging infrastructure
- Provide access to heterogeneous information repositories
- Provide application integration
- Highly scalable and highly reliable

This service enables seamless access to distributed resources allowing location and exchange of geographically distributed data [1]. These capabilities are a direct result of research and development work performed under the Object Oriented Data Technology (OODT) task at JPL [2].

1.2 Metadata Service

This service provides the ability for information systems to store information related to data products managed and available within the enterprise. The service provides a series of registries that allow for registration, capture, navigation, and reuse of:

- Defined metadata schemas or data dictionaries
- Defined common and specified data elements
- Data resource descriptions
- Relationships between metadata schemas

This service applies industry metadata standards to promote the efficient sharing of information. Common data elements, from the Dublin Core Metadata Element Set [3], are utilized to describe data systems, data sets and data products in a standardized manner. Data elements, themselves, are described using the ISO/IEC 11179 [4] standard.

1.3 Catalog and Archive Service

The purpose of this service is to provide an active product storage and retrieval capability for missions and projects. The service is designed with the following features:

- Product policy definition based on project needs
- Product type flexibility (archiving any blob)
- Cataloging based on metadata (from Metadata Service)
- On-line, near-line, and off-line storage of products
- Secure access to products
- Event-based task execution (specialized for instantiations based on project need)
- Distributed access utilizing an Application Program Interface (API)

Along with providing a data product repository this service enables rule-based processing of those products. The service provides a pluggable interface for executing tasks and rules. Tasks can be written to execute processing algorithms whereas rules can be written to evaluate a task's readiness for execution. Both of these capabilities are managed by a queuing mechanism allowing for automated process control.

1.4 Application Hosting Service

The purpose of this service is to provide an institutionally supported and stable, managed system environment for project or departmental applications. This service is designed with the following features:

- Common application and storage architecture
- Reliable and scalable data storage
- Secure, well maintained facility
- Assistance on hardware evaluation, COTS selection, installation and configuration
- Provide common infrastructure for application management
- Support on system management, installation, upgrade, integration, monitoring, backup, and restore
- 24x7 or 9x5 support

1.5 Database Hosting Service

The purpose of this service is to provide an institutionally supported and stable, managed database environment. This service is designed with the following features:

- Secure, well maintained servers and facility

- Support for data management on structured and non-structured data
- Support for small and large-scale databases
- Support on database administration, backup, monitor for consistent data reliability and performance
- Scalable and reliable data storage
- 24x7 or 9x5 support

2. EDM/SEAWINDS COLLABORATION

The EDM and SeaWinds projects at JPL collaborated to create a platform and database independent service for managing files and tasks. The result of this collaboration was the EDM Catalog and Archive Service.

The goals of this collaboration were to:

- Build new process and data management components that are reusable, reliable and scalable
- Replace unreliable SeaWinds components with EDM reliable components
- Implement a rule-based task execution capability that will ensure "lights out" SeaWinds operations

The benefits derived from this collaboration were:

- Reduced cost and schedule for SeaWinds project
- Integrated development with existing SeaWinds design
- Developed a framework for future science data systems that is reliable, reusable, scalable, platform-independent and database-independent
- Implementation supported two separate ground data systems (SeaWinds on ADEOS II and SeaWinds on QuikSCAT)

Illustrated below is the SeaWinds Processing and Analysis Center architecture (Figure 2). In subsequent illustrations the Data Management and Automatic Process Control portion of the architecture is shown in more detail.

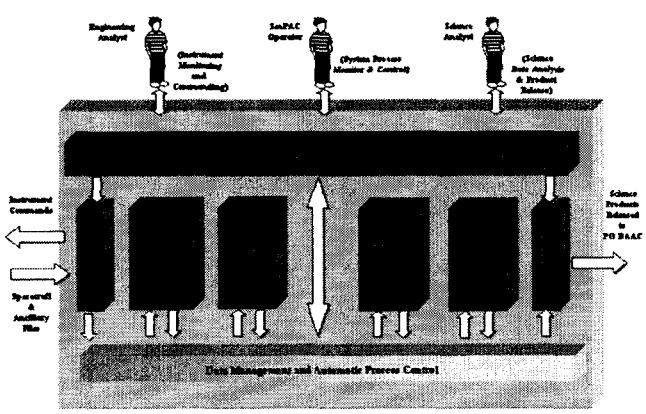

Figure 2. SeaWinds Processing and Analysis Center Architecture

The first detailed illustration below (Figure 3), depicts how the original architecture was heavily reliant on vendor specific products. The result was an architecture that was database dependent, platform dependent and operator intensive.

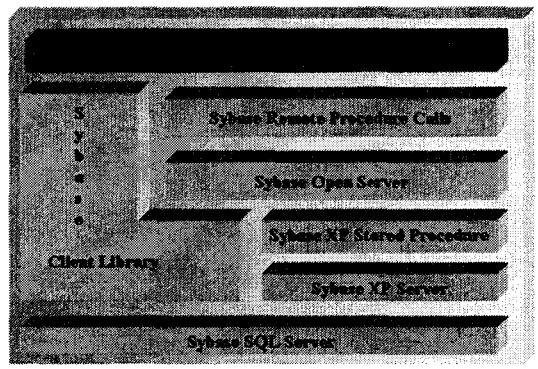

Figure 3. Original Data Management Architecture

The second detailed illustration below (Figure 4), depicts how the improved architecture was simplified while providing new functionality. The result was an architecture that was database independent, platform independent and provided "lights out" operation.

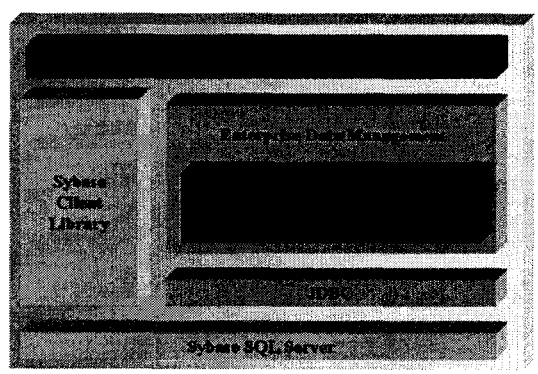

Figure 4. Improved Data Management Architecture

3. GENERIC FRAMEWORK

The end result of the SeaWinds collaboration coupled with other efforts within the EDM project is a shrink wrapped version of a science data system with the following characteristics:

- The framework accommodates all sized missions (scalable & reusable)
- The data system hardware configuration is not driven by the framework (hardware-independent)

- The selection/purchase of a database is not driven by the framework (database-independent)
- Mission-specific features can be "plugged in" to the framework (adaptable)
- Provide interoperability with mission operations and science analysis (one integrated system)
- Allow data sharing from multiple missions at distributed repositories for science analysis, data modeling, and knowledge discovery

The illustration below (Figure 5) depicts a generic architecture for a science data system.

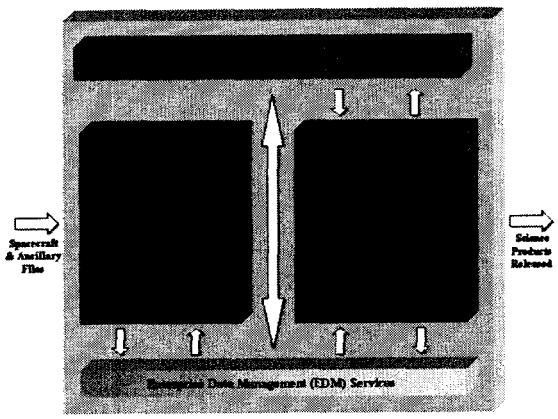

Figure 5. Generic Science Data System Architecture

As depicted in the generic architecture, mission specific processor applications and analysis tools are still necessary for a particular mission or project. EDM provides the management of products and metadata, and the interfaces necessary to access that data from mission specific applications. Specific contributions from the EDM services for the generic science data system are as follows:

- Enterprise Application Middleware
  - Provides method to access other mission resources
  - Enables data sharing across mission and science disciplines
  - Enables collaboration with partners
- Metadata Service
  - Captures mission data models
  - Manages mission data dictionary
  - Identifies mission resources and their location
- Catalog and Archive Service
  - Stores and catalogs mission files
  - Validates product metadata
  - Provides autonomous rule-based process management
- Application Hosting Service
  - Hosts mission specific applications
  - Provides 24x7 system monitoring and notification
  - Provides system administration
- Database Hosting Service
  - Hosts mission databases
  - Provides storage for mission files
  - Provides database administration

4. CONCLUDING REMARKS

The EDM framework of services, where possible, promotes the use of open standards and open source software. The underlying architecture is designed in a flexible manner to accommodate new technology and evolving standards. The services were developed with the Java programming language using XML for service interfaces. The messaging layer currently supports CORBA and Java Remote Method Invocation (RMI).

The solution presented, although applied to science data systems, is not limited to that discipline. In fact, the framework is adaptable based on the metadata definitions that are defined. This allows for the solution to then be applied to other disciplines including planetary, astrophysics, healthcare, business, etc.

5. REFERENCES

[1] Crichton, D.J., Hughes J.S., Hyon J.J., Kelly, S.C., Science Search and Retrieval using XML, The Second National Conference on Scientific and Technical Data, U.S. National Committee for CODATA, National Research Council, March 13-14, 2000. [http://oodt.jpl.nasa.gov/papers/codata.pdf]

[2] Object Oriented Data Technology (OODT). [http://oodt.jpl.nasa.gov]

[3] Dublin Core Element Set 1.1, "Dublin Core Metadata Element Set, Version 1.1: Reference Description", 2 July 1999. [http://www.dublincore.org/]

[4] ISO/IEC 11179 "Information Technology -- Specification and standardization of data elements," 2000. [http://www.iso.ch]