

# **A MODEL DRIVEN ARCHITECTURE FOR HIGHLY DISTRIBUTED, DATA INTENSIVE SYSTEMS**

## ***WHITE PAPER***

Daniel J. Crichton  
Paul M. Ramirez  
Chris A. Mattmann  
J. Steven Hughes

*Jet Propulsion Laboratory  
California Institute of Technology*

**May 22, 2008**

### **Technical Areas Addressed**

- (i) Automated metadata extraction from unstructured and semi-structure information
- (ii) Scalable, secure, federated search
- (iii) Resolution of object identifiers into the address(s) of the most available copy
- (iv) Reliable, persistent identification of unique digital objects and variants
- (v) Provenance and tracking of objects

### **Point of Contact**

Daniel Crichton  
[Daniel.Crichton@jpl.nasa.gov](mailto:Daniel.Crichton@jpl.nasa.gov)  
4800 Oak Grove Drive, MS 169-415  
Pasadena, California 91109  
Phone: 818.354.9155  
Fax: 818.393.1370  
Workshop Attendance: Yes

## 1. Executive Summary

The Jet Propulsion Laboratory (JPL) has been researching and building data intensive systems for highly distributed scientific environments for many years. Due to the dynamic and changing mission environment for both solar system and earth exploration, these systems have a number of critical architectural principles that have been paramount to defining an architecture that can evolve with exploration and technological changes. As a result, JPL has defined a data and computational grid architecture which enables capture, processing, discovery, access, and transformation of digital data objects across highly distributed environments based on rich metadata descriptions of the objects. This architectural framework, called the Object Oriented Data Technology (OODT) framework [2,5,8], was selected as runner up for NASA Software of the Year in 2003 and has been extensively used on planetary, earth, astrophysics and biomedical projects in order to support archiving, processing and distribution of data in highly distributed environments. In addition, JPL has been chairing the Information Architecture Working Group as part of the Consultative Committee on Space Data Systems (CCSDS) in order to define an international reference architecture for space information management based on the management of information objects which consist of both metadata and data objects.

## 2. Conceptual Architecture

One of the critical characteristics of the architecture has been leveraging architectural patterns across very different science environments. The software architectural patterns for digital object data management have been leveraged across multiple environments despite the differences in the domain models and the disciplines themselves. The architecture focuses on definition of both the information and software architecture portions of a system.

As part of defining the architecture for any science-oriented data system, we identified particular functions as having common architectural patterns that would allow us to implement a set of information services. The services include data capture, discovery, access, retrieval, processing, and distribution. These services allow for distributed, independent deployment yet still work in concert with one another allowing virtual systems that span organizational boundaries to be constructed.

In addition to the above descriptions of functional services within the architecture, there is also the *information* architecture that is critical to forming the domain implementation. As part of designing the information architecture for any domain, we have been actively involved in developing a standard information model for the representation of information associated with data objects managed within the different scientific domains. This includes models for critical data objects that are acquired as part of planetary science or cancer research that will be described later. The data objects that are captured, managed and exchanged as part of the architecture are described by a “metadata object” which provides a set of attributes for the object as described in the domain information model.

The OODT framework provides a set of core services that implement the above functions which themselves are driven by the domain model (e.g., the cancer biomarker information model, the planetary science information model, etc). The loose coupling

between each service and its associated domain model allows for the services to support multiple domains. Each of the OODT services can be deployed independently and then integrated using XML-based interfaces over a distributed, grid architecture. This service independence and insulation makes it possible to query multiple organizational repositories (either local or distributed) concurrently, compiling the results into a unified view, and making them available for analysis. The OODT service framework is based on the software architectural notion of components. Each component has well known interfaces that enable them to be plugged together in a distributed manner. The components themselves sit on top of off-the-shelf middleware so that they can be deployed into an enterprise topology.

The *Catalog and Archive Service* component provides the ability to catalog, process, and store information objects in a distributed environment. The *Profile Service* component provides a registry of information about managed information objects necessary to discover them. Multiple profile services can be distributed and integrated into a directed graph topology in order to crawl the registries and locate critical information objects. The *Product Service* provides a mechanism for access, retrieval, and transformation of science data products and information from remote repositories. The *Query Service* provides an interface to distributed components so that they work together.

Each of our domain implementations is working to build specific applications on top of this common services framework. Several applications have built portals that provide the own specific extensions for searching and accessing data via the framework. For example, the NASA Planetary Data System (PDS) used a Lucene-based search engine that integrated with OODT to provide sub-second searching across highly distributed databases using a text-based search interface. The benefit of the framework is that it has substantially helped to both build new data systems as well as integrate existing data systems in a virtual data system while controlling software development costs through reuse and standard interfaces.

### **3. Distributed Object Storage and Retrieval Vision Revealed**

The OODT framework has enabled parts of the distribution object storage and retrieval vision to be realized. Initial goals of the framework were driven from the perspective that access to information should be transparent to end consumers. In order to materialize a framework to support this simple concept several intermediary concepts were born, namely annotation, distribution, identification, and consistency of information objects. On top of these concepts areas of concern such as security, scalability, and maintenance were laid. The following will discuss the extent to which portions of the vision that the framework has addressed. Focus will be placed on areas which the OODT framework has its most significant contributions.

Automated metadata extraction from unstructured and semi-structured information has always been at the forefront of the OODT framework. The model behind the framework is based on information objects and their annotations. These annotations are a collection of metadata either provided or derived. Derived metadata is facilitated through pluggable extensions of the *Profile Service* and is the basis for extracting metadata from information objects. Every information object is annotated with a minimal set of metadata, originally prescribed by Dublin Core but later refined, to enable the services provided within the framework. The minimal set of metadata can be automatically from

both unstructured and semi-structured information objects. Additional metadata is extracted by aforementioned extensions. The interface for these extensions is loosely defined as taking in an information object and producing a collection of metadata. It is here where future work can be done the process of building said extensions.

Scalable, secure, federated search is key to any distributed object storage system and was not lost on the founders of the OODT framework. Initially, the search was implemented as network query engine in the form of the *Query Service*, later this evolved to include indexing to increase scalability and performance. At this point, the integration of security is not very granular. Upon inception of the framework much of the security was handled by who was allowed to access the data grid, current work has taken a more granular approach to security by demonstrating restricted access down to the information object level. This is an area where the framework can be flushed out to provide consistent access and authorization control across a data grid. The plan to accomplish this goal is through a unified identity service back by a technology such as LDAP.

Resolution of object identifiers into the addresses of the most available copy provides the backbone for location independence in a data grid framework. The *Query Service* is the workhorse in resolving where data resides given unique object identifier. Resolution of addresses is done by polling the *Profile Service*, which provides an interface to a set *Profile Servers* possibly arranged in a hierarchy for scalability, for location information on a given object identifier and then decides on a location and passes it up the consumer. How this decision is made can be an extension point in the framework to allow users to modify behavior and not be stuck with the idea that it is just the closest object as this may or may not be the requirement for a given instantiation of the framework. Furthermore, investigation as to whether or not this information should be pushed upstream has been under recent consideration to enhance scalability. Technologies such as bit torrent have pushed forth the idea that access to an object can come from multiple sources and is a definite area for future study.

Provenance and tracking of objects has recently made its way into the OODT framework. While at this time it is limited to within the *Catalog and Archive Service* component it is envisioned as a first class service. The work was originally scoped to this service as a proof of concept and to facilitate the flushing out of requirements on such a service. The vital information provided by this service will enhance existing services and allow work within the data grid framework to occur in a simplified manner. For instance, this type of service will be able to publish information to the *Query Service* about where a particular information object has gone, thereby allowing it to work more efficiently.

Reliable, persistent identification of unique digital objects and variants has been a concern that the OODT framework has had to deal with due to the nature of the *Product Service*; which provides transformations of information objects. Each information object has associated metadata that provides not only unique identification but also any transformations it may have undergone. These transformations occur as decorators on an objects identity and thus do not change it. The inherent issue with this approach is that derived products are not assigned a new identity; however this issue will be flushed out in the provenance and tracking service.

#### **4. Case Study: Planetary Data System**

The NASA Planetary Data System (PDS) was born out of a recommendation from the National Research Council (NRC) and has become NASA's official data system managing scientific results from solar system exploration for the planetary science research community. It is a distributed system sponsored by the NASA Planetary Science Division and is composed of eight teams, called nodes. The node structure is organized primarily by sub-discipline within planetary science to focus expertise with acquiring and curating data. The core science discipline nodes include Atmospheres, Geosciences, Planetary Plasma Interactions, Rings, and Small Bodies Nodes. Each works with an advisory group that provides guidance on priorities to that sub-discipline area.

The Jet Propulsion Laboratory provides system engineering leadership and software development to the Planetary Data System handling global aspects that include architecture, standards, and software development. JPL has been successful at leveraging the OODT framework to enable distributed access to very repositories that are maintained at the scientific discipline nodes of the PDS [5,7]. In addition, PDS has been recently working with other space agencies to enable sharing of digital data objects from planetary science repositories at other space agencies.

The PDS has an information architecture that has been in existence for over twenty years. While the technologies underpinning the PDS have changed, the information architecture has been very stable. At the lowest level, the information architecture of PDS dictates that all data objects returned from missions be annotated using a set of metadata that is either attached or detached from the data object itself. The metadata is governed by a rich data dictionary which itself falls out of the PDS science data model and identifies data elements and permissible values for documenting the planetary science data results. The PDS also provides standards for both the structure (grammar) of the metadata as well as the applicable metadata for any given data object. This allows PDS to validate data results against the information architecture.

The technical architecture of PDS is able to use the information architecture such that changes in the model and data dictionary drive changes in the software. The OODT framework allows for distributed discovery and access of data based on the PDS information architecture. In fact, in 2002, the OODT architecture was inserted into PDS connecting distributed nodes into a virtual grid without requiring changes to the repositories or information architecture within PDS [7]. One of the core principles has been to ensure that the information and technical architectures are separated such that the model can be updated without requiring changes to software. This is absolutely essential when dealing with robotic exploration missions since the scientific instruments tend to be "one of a kind" often resulting in model changes from mission to another.

## **5. Case Study: Early Detection Research Network**

The Early Detection Research Network (EDRN) is an infrastructure funded through the National Cancer Institute (NCI) for supporting collaborative research on molecular, genetic and other biomarkers in early cancer detection and risk assessment. The EDRN model is comprised of 3 major units: Biomarker Development Laboratories (BDL), Biomarker Reference Laboratories (BRL), and Clinical Epidemiology and Validation Centers (CEVC). The hub of the network is the Data Management and Coordinating Center (DMCC). In total, there are over forty institutions that participate within the EDRN.

The Jet Propulsion Laboratory has served as the “informatics center” for the EDRN defining the overall architecture and standards, and developing software that enables the EDRN to both capture, discover and share digital data objects from cancer biomarker research. As a result, JPL has led definition and development of the EDRN Knowledge System. The EDRN Knowledge System encompasses several components which aid in the management and distribution of information acquired during the biomarker discovery process including biomarkers, studies, specimens, science data and publications [1]. Over the past few years, EDRN has become a leader in developing and promoting informatics technologies to enable data sharing across the EDRN enterprise and has served as a model for data sharing for the National Cancer Institute. This has included the deployment of “ERNE” [4,6], the EDRN Resource Network Exchange, which has enabled researchers to discover information about specimens that exist at research institutions within the EDRN. EDRN is currently connecting over twelve institutions into a virtual data grid, allowing researchers to access information from very diverse systems that exist a geographically distributed cancer centers around the United States.

Similar to planetary science, JPL was able to leverage the OODT framework to connect distributed cancer centers together [5]. However, rather than use the information architecture for planetary science, JPL was able to define an information architecture for cancer research which allows for the capture, discovery and exchange of data objects for cancer research [3]. The OODT software itself, which provides basic functions for managing and accessing data objects, did not change between planetary and cancer research. This was a major goal that was based on the premise that there are standard patterns for distributed data management that several systems share. It also validated the need to separate the information architecture from the technical architecture since the information architecture needs to be discipline-specific. In the case of the EDRN, JPL was able to work with the scientists to define an underlying core ontology for cancer research. The ontology effort has been critical in order to identify the core objects and their attributes, as well as understand how to share those data objects between distributed applications and systems.

## **6. Conclusion**

JPL has a number of lessons that it has learned over the years which have been critical to building systems and an infrastructure to support management of digital objects. A well-defined architecture has been critical to allowing JPL to build a reusable framework that has supported a number of projects including science, engineering, and business environments. In addition, separation of the technical and information architectures have been a core principle to enable disciplines to define objects that meet their specific requirements. JPL has published a number of papers and has presented extensively on its architecture and framework which is also being published as part of the *Reference Model of Space Information Management* within the Consultative Committee on Space Data Systems. It’s work with the Planetary Data System has served as a model for space agencies around the world as well as other scientific disciplines including cancer research and pediatrics.

## REFERENCES

- [1] D. Crichton, S. Kelly, C. Mattmann, Q. Xiao, J. S. Hughes, J. Oh, M. Thornquist, D. Johnsey, S. Srivastava, L. Esserman, and B. Bigbee. “A Distributed Information Services Architecture to Support Biomarker Discovery in Early Detection of Cancer”. Accepted for publication at the *2nd IEEE International Conference on e-Science and Grid Computing*, Amsterdam, the Netherlands, December 4th-6th, 2006.
- [2] C. Mattmann, D. Crichton, N. Medvidovic and S. Hughes. “A Software Architecture-Based Framework for Highly Distributed and Data Intensive Scientific Applications”. In *Proceedings of the 28th International Conference on Software Engineering (ICSE06)*, pp. 721-730, Shanghai, China, May 20th-28th, 2006.
- [3] J. S. Hughes, D. Crichton, S. Kelly, C. Mattmann, T. Tran. “Intelligent Resource Discovery using Ontology-based Resource Profiles”. *Data Science Journal*, Vol. 4, pp. 171-188, December 2005.
- [4] D. Crichton, H. Kincaid, J.S. Hughes, S. Kelly, S. Srivastava, D. Johnsey. “Creating a National Virtual Knowledge Environment for Proteomics and Information Management”. In *Informatics and Proteomics*. Marcel Dekker Publishers. December 2004.
- [5] D. Crichton, J.S. Hughes and S. Kelly. “A Science Data System Architecture for Information Retrieval”. In *Clustering and Information Retrieval*. Kluwer Academic Publishers. December 2003.
- [6] D. Crichton, H. Kincaid, S. Kelly, S. Srivastava, D. Johnsey. “A National Data Grid Infrastructure for Sharing Biospecimens in Early Cancer Detection”. In *Proceedings of the Digital Biology: the Emerging Paradigm*, Bethesda, MD. November 2003.
- [7] D. Crichton, J.S. Hughes and S. Kelly. “A Distributed Data Architecture for 2001 Mars Odyssey Data Distribution”. In *Proceedings of the Space Mission Challenges for Information Technology*, Pasadena, California. July 2003. <http://oodt.jpl.nasa.gov/papers/dda.pdf>
- [8] D. Crichton, J.S. Hughes, S. Kelly and J. Hyon. “Science Search and Retrieval using XML”. In *Proceedings of the 2nd National Conference on Scientific and Technical Data*, Washington D.C., National Academy of Sciences. March 2000. <http://oodt.jpl.nasa.gov/doc/papers/codata/paper.pdf>