A Planetary Data System for the 2006 Mars Reconnaissance Orbiter Era

PV-2004, "Ensuring the Long -Term Preservation and Adding Value to the Scientific and Technical Data"

5-7 October2004, ESA/ESRIN, Frascati, Italy

J. Steven Hughes[1], Daniel Crichton[1], Gerald Crichton[1],
Ronald Joyner[1], Sean Kelly[1], Chris Mattmann[1], Joel Wilf [1]

[1] Jet Propulsion Laboratory Pasadena
4800 Oak Grove Drive
Pasadena, CA 91109 USA
Email: {steve.hughes , dan.crichton, jerry.crichton, ron.joyner,
sean.kelly, chris.mattmann, joel,wilf}@jpl.nasa.gov

INTRODUCTION

The Planetary Data System (PDS) is the official science data archive for NASA's planetary science community and currently contains about 10 terabytes of data collected from over thirty years of solar system exploration missions. The PDS is an online archive that consists of several geographically distributed science discipline nodes. These nodes support the production of archive quality data products, curate the products, and provide supporting science and technical expertise to the planetary science community.

The success of the PDS is primarily due to the early development and use of a data architecture and data standards for organizing and describing the data in the archive. After being validated against the standards and peer-reviewed by discipline scientists, the data was distributed on physical media to all subscribed users. These "self-contained" archive quality volumes contained not only scientifically useful data products but supporting ancillary data files and informative metadata. The ability of individual scientists to build their own local data library has resulted in high expectations with regard to data availability.

The advent of the Internet combined with sky rocketing costs for distributing physical distribution media forced a review of the distribution process and resulted in the successful development of a middleware-based framework [2] for the on-line distribution of the 2001 Mars Odyssey mission data products. As a result, all missions slated to archive their data are now expecting to use the new distribution capabilities. In addition, it has been recognized that the framework can be used to integrate and simplify the loosely coupled data product production "pipeline" currently used to produce science data products.

The key development challenges now involve software reuse and system deployment. As missions produce larger volumes of data and more product types, the tendency now is to create temporary data nodes at the instrument team institutions to produce and curate the data during the mission. This trend suggests that customized data production and distribution subsystems be configured for specific product types and deployed as needed. The mandate remains however that all data repositories conform to a single data architectural standard and that are capable of being viewed as an integrated whole by external users.

Scalability also poses to be a key challenge for future planetary missions. The science data archived from all planetary exploration missions prior to the 2001 Mars Odyssey mission totals approximately 5 terabytes. The currently active Mars Odyssey mission is expected to double this volume and the 2006 Mars Reconnaissance Orbiter (MRO) is expected to increase the resulting volume by a factor of 10. Future missions such as the 2012 Jupiter Icy Moons Orbiter with its relatively unlimited power supply are predicting even far more data.

This paper will describe several efforts now proposed to address this challenge while providing reuseable, customized, scalable, and remotely deployable software packages that will meet the data intensive requirements of the MRO era and beyond.

DATA SYSTEM REQUIREMENTS

The conversion of the PDS to an online data system necessitated the development of an updated set of system requirements. This was accomplished by a PDS-wide gathering and definition effort that included lessons learned from implementing the 2001 Mars Odyssey distribution system and needs from the project and the science user communities. This input was drafted into a set of requirements by a PDS working group and then edited and documented by the central node systems engineering team.

The resulting Software Requirements Document (SRD) [7] includes a system description, functional requirements, non-functional requirements (system characteristics and constraints), priorities, and an approach for changing the requirements. Subsequent review also partitioned the requirements into those for a "core" system and an "extended" system. The core system includes capabilities for data ingest, validate, track, notify, search, display, retrieve, transfer, and administer. These capabilities are fundamental and were already largely well-defined. The extended system extends the capabilities of the core system with correlative search, advanced display and retrieval, data mining, geometric processing, etc.

DATA DISTRIBUTION – SEARCH, DISPLAY, RETRIEVE, AND NOTIFY

Enabling a user to find, view, and get data and related resources are addressed by the following key functional requirements. A user shall be able to 1) find data based on any indexed attribute, 2) view a visual representations of data through a web interface, 3) download selected products along with associated products, 3) download a subset of selected products, and subscribe to be notified when new data become available.

As previously mentioned, the PDS developed a middleware-based framework for the on-line distribution of the 2001 Mars Odyssey mission data products, called PDS-D (see Fig. 1). The deployed data subsystem met the data distribution requirements for the Mars Odyssey mission, is now successfully distributing almost all the legacy data in the archive, and is expected to meet the distribution requirements for missions planned through the MRO era.

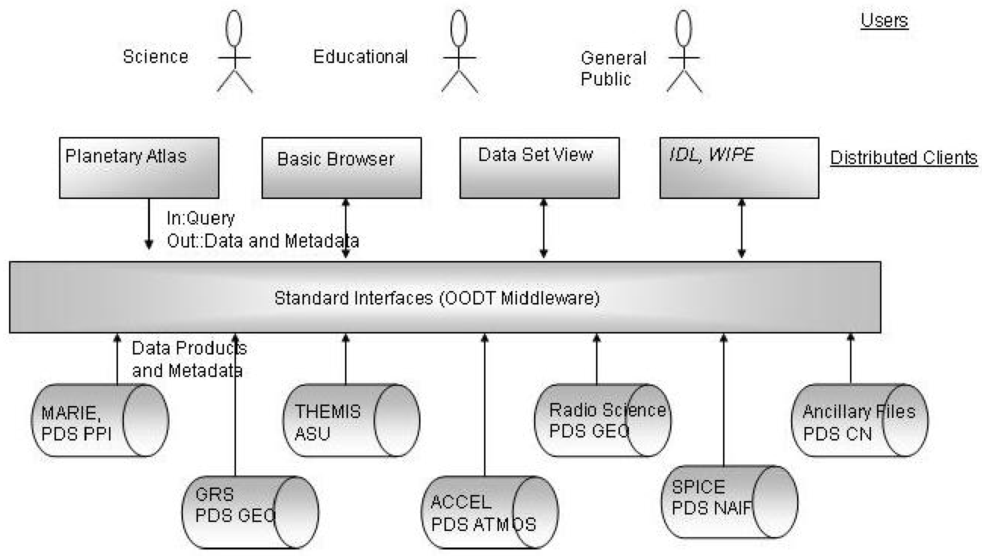

Figure 1 - PDS-D Configuration for Mars Odyssey

PDS-D is a based on a four-tiered information architecture that consists of application, web, service, and storage tiers. The multi-tiered architecture partitions the system into separately managed components and also provided a means by which the system can be deployed with minimal impact on existing resources. For example, the Planetary Atlas developed by the PDS imaging node is the primary application for searching and retrieving imaging products from the archive and represented a significant investment of developmental resources. The Planetary Atlas was integrated into the client tier of the architecture as can be seen in Fig. 1. The service (middleware) and storage (repositories) tiers are also illustrated in the figure.

In the storage tier, the data archive remains geographically distributed and locally managed by the science domain experts. The development of this tier leveraged heavily on the data architecture previously created by the PDS for the archive. Organized into archive volumes on optical media, the data with its accompanying metadata and ancillary data was copied from the archival volumes onto mass storage devices at a local node to make the collection accessible through the service tier.

The service tier includes the middleware as well as service components that support the search, retrieval, and transformation of data from the geographically distributed components of the storage tier. Key service components include product servers that provide a common interface for the retrieval and transformation of data products from a data repository. Profile servers provide a common interface for product search. A query service manages queries entering the system by broadcasting them to the appropriate product or profile servers and compiling the results for transmission back to the user. The service tier and its implementation using the Object -Oriented Data Technology (OODT) platform is described in more detail in [1,2,3].

Application developers at the client tier can interface directly with the service tier through JAVA APIs. However the web tier includes a web server to allow http level access. In fact the majority of applications developed for the infrastructure have used http level access for data search and retrieval.

A key design principle that has evolved out of the PDS experience is the importance of separating the data architecture from the technology architecture [4]. This principal naturally arises by observing the differences in the life cycles of data architectures versus technology architectures. Data architectures typically model a specific domain and evolve as the domain evolves, typically very slowly. However the technologies use to implement systems evolve much more rapidly and are often out-of-date soon after an implementation has finally been deployed. The independence of the two architectures allows both to evolve as needed to meet the needs of the enterprise. The OODT framework uses XML to manage the metadata both for resource descriptions and messaging and the ISO/IEC 11179 [6] specification as a standard for vocabularies. A novel XML document called a resource profile is use to provide a single uniform view of all resource descriptions for both querying and returned results [1].

The data architecture and data standards together act as important leverage for the development and implementation of the technology architecture. For example, the existence of and adherence to a data architecture not only made the configuration of the PDS storage tier very easy, but the wealth of metadata in the archive also makes the indexing of products for catalog search relatively easy. In addition the metadata also provides a rich source of information for data mining and correlative search applications.

DATA INGESTION – INGEST, VALIDATE, AND TRACK

Governing the incorporation of mission archive data into the system is addressed by the following key functional requirements. The system shall a) receive data , b) catalog all metadata or "data descriptions", c) validate the data delivery against a specification to ensure that it is compliant with the data standard and d) keep track of the status of every data delivery. These core functional requirements are not much effected by specific requirements from the upcoming MRO and Cassini missions. However MRO-specific requirements do raise data volume issues and Cassini-specific requirements raise model complexity issues associated with the large number of data types expected from the mission. Two key issues being addressed are 1) the validation functional component needs to scale to meet the combined load of the two missions and 2) even though the data will be available online, the requested distribution volume to the planetary community will be very difficult to meet using commonly available web technologies.

The PDS is currently in the process of developing the data ingestion, validation, and tracking subsystem to meet the given requirements. The multi-tiered architecture based on the OODT middleware that was introduced for the PDS-D implementation will again provide the base platform for development. This platform has matured and is now “open source” and a “web service” interface has been enabled. The PDS data repository structure introduced for PDS-D will provide the data storage component.

Key to the validation requirement is the existence of a specification against which all archived data are validated. This is embodied in the PDS data standards as documented in the PDS standards reference. This reference document is being edited to make it more suitable as a specification. After validation the results are used as a criteria for accepting a delivery for ingestion. Levels of validation include syntactic or checking that the documents are “well formed”, semantic, or checking that the data is appropriately described using domain knowledge, and quality checks that ensure the delivery is complete, consistent, and useable.

The production of archive quality data products requires a well-defined process and “work flow” that manages the movement of individual data product through a production “pipeline”. Key functions include controlling product versioning, state changes, and the application of either automated or manual process. An Ingest Workflow Manager is planned, based on the successful ground data system built for the SeaWinds project. This system, built from an existing component of the OODT framework called the Catalog and Archive Service (CAS), will now be validated in a multi-mission environment as shown in Fig. 2. The CAS provides a rule-based infrastructure focused on tying together a set of rules used to construct both a catalog and a repository into the workflow. The rules are implemented as set of Java-based processes that run as part of the process of ingesting a data product. The ingestion process is part of a larger transaction which means that products that fail to be properly ingested will cause the CAS to rollback. The multi-mission configuration is supported by enabling highly diverse processing scenarios. These include scenarios from simple validation to complex processing and creation of new data products. The CAS will extend the distributed PDS by enabling access to the catalog and repository using the OODT information architecture which provides common definitions for data queries and product resources.

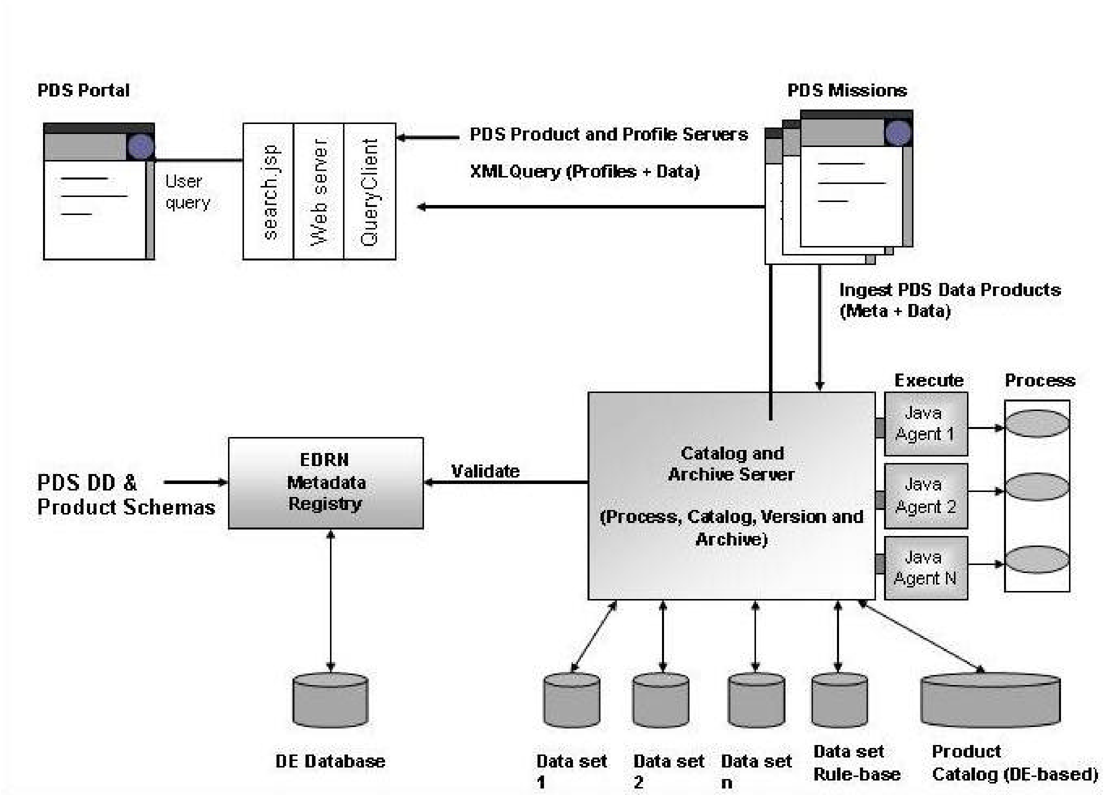

Figure 2 - PDS Distributed Ingest Architecture

Similarly a prototype Data Tracker developed for the Cassini mission will be generalized to meet multi-mission data tracking requirements. It should be noted that the validation requirements just discussed focus on the data production pipeline. PDS policy still requires passing a science peer review before a data collection can be considered archived. Fig. 3 illustrates a mission data flow into the PDS for distribution and archive. It abstracts the multi-tiered distribution architecture of Fig. 1 into a single component for a single node and integrates the ingestion, validation, and tracking elements into an additional component.

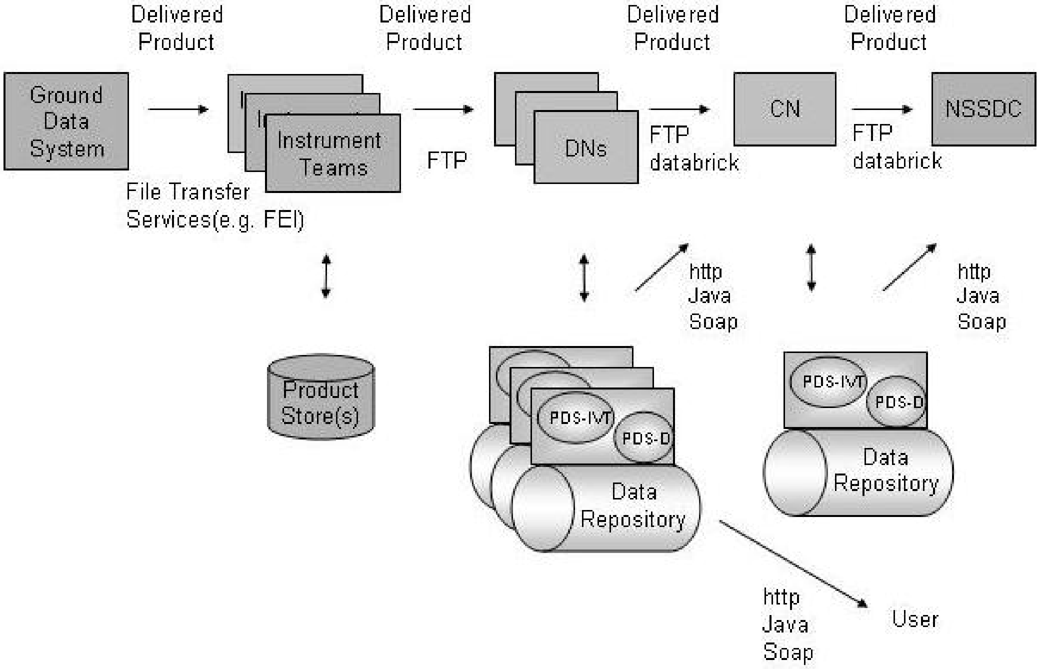

Figure 3 - Product Data Flow with Ingestion, Validation, and Distribution

CONCLUSION

The Planetary Data System archives data for the planetary science community. Although the total data volume in the archive is not large relative to other science domains such as Earth Science, the planetary science domain is much more complex than most other space science domains since they involve dynamic systems that include orbiting target bodies, moving instrument platforms, complex instruments, a plethora of frame of references, changing sources of light, etc. The development of a data architecture and data standards in the early years of the PDS enabled the creation of a data archive consistent in its structure, meaning, and organization as well as rich in descriptive information. The advent of the web has provided the technologies needed to make the planetary science archive available to a wider range of customers in increasingly more useful and sophisticated ways.

Movement towards standard mechanisms to catalog, validate, process and distribute data via distributed software interfaces is enabling the PDS to handle the increase in volume and complexity for future missions. Key to this advancement is the formulation of an information architecture that provides for independent data and technology architectures and the development of a multi-tiered infrastructure that allows the integration of heterogeneous distributed data repositories into single system with common system interfaces. [1,2,3] The planned implementation of the ingest/validate/track functional components into the existing infrastructure will provide a highly-automated end-to-end data production and distribution system for the planetary science community. This is a long awaited goal that few other science domains of this breadth have accomplished.

Future development includes addressing the desires of some MRO scientists to distribute entire collections of highly derived data to thousands of users at the end of the mission. Calculations based on current estimates predict this would necessitate the distribution of petabytes of data in the period of a few months. A proposal has been written to research available technologies such as grid services [5] and to deploy a prototype infrastructure across the planetary science community.

REFERENCES

[1] Crichton D, Hughes JS, Hyon J, Kelly S., "Science Search and Retreval using XML". Proceedings of the 2nd National Conference on Scientific and Technical Data, National Academy of Science, Washington D.C, 2000.

[2] Kelly S, Crichton D, Hughes JS., "Deploying Object Oriented Data Technology to the Planetary Data System", Proceedings of the 34th Lunar and Planetary Science Conference 1607, 2003.

[3] Mattmann C, Ramirez P, Crichton D, Hughes, JS, "Packaging Data Products using Data Grid Middleware for Deep Space Mission Systems", Proceedings of the 8th International Conference on Space Operations (Space-Ops2004), 2004, in press.

[4] Consultative Committee on Space Data Systems, "Space Information Architecture", White Paper, Information Architecture Working Group. February 2004, in press.

[5] Foster I, Kesselman C, Nick J, Tuecke S., "The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration", Open Grid Service Infrastructure WG, Global Grid Forum, 2002.

[6] ISO/IEC 1999. Framework for the Specification and Standardization of Data Elements 11179-1, Specification and Standardization of Data Elements 11179. International Organization For Standardization.

[7] The Planetary Data System- System Requirements Document (SRD), Version 1.2, JPL Internal Document, November 2003.