The Planetary Data System for the Next Decade

Sean Hardman [(1)], Dan Crichton [(2)], John S. Hughes [(3)], Reta Beebe [(4)]

[(1,2,3)] ***Jet Propulsion Laboratory, California Institute of Technology***
4800 Oak Grove Drive, Pasadena, California, 91109, USA
EMail: Sean.Hardman@jpl.nasa.gov[(1)], Dan.Crichton@jpl.nasa.gov[(2)], *John.S.Hughes@jpl.nasa.gov*[(3)]

[(4)] ***New Mexico State University***
P.O. Box 30001/MSC 4500, Las Cruces, New Mexico, 88003, USA
EMail: RBeebe@nmsu.edu

ABSTRACT

Over the past two decades, the planetary science community has seen major changes in both computing capabilities on the ground and with the spacecraft flown for planetary science missions. These capabilities allow for more diversity and greater opportunities to perform scientific research and support analysis of the data. However, this paradigm must be embraced and planned for by scientific data systems to ensure that these systems can scale appropriately to support the demands placed on them by both the missions and the users. In the area of planetary science, there is also a growing trend towards internationalization with access to massive data sets. For the Planetary Data System (PDS), this requires an architecture to capture, preserve and deliver data through an online, national system as well as support sharing and interoperability at an international level with other planetary science data archives.

Keywords: NASA, JPL, PDS, planetary, data, archive

INTRODUCTION

As the next decade unfolds, it is clear that scientific computing infrastructures (or cyberinfrastructures) will continue to gain traction. These computing infrastructures will require additional services to better enable scientific analysis within distributed environments and to ensure those services integrate with the systems and data in place. Therefore, it is essential that scientific archive systems, the facilities established to support long-term research and analysis as well as preservation of the data, have a plan in place to evolve and leverage these newer paradigms.

The Planetary Data System (PDS) is faced with these very challenges requiring a modernization of its architecture and technical implementation to ensure it can meet the demands of the next decade. As a result, the PDS has undertaken a modernization effort to position itself to support both the missions and scientific users for the next decade with version 4 of the PDS. "PDS4" includes three fundamental project goals:

- Provide more efficiency for the delivery of data by data providers to the PDS
- Enable a more stable, long-term usable planetary science data archive
- Enable computing services for the data consumer to find, access and use the data they require in contemporary data formats

In order to achieve these goals, the PDS is upgrading both the technical infrastructure and the data standards in order to support increased efficiency in delivery of data as well as usability of the data system in light of a decade of new plans for robotic exploration of the solar system. Efforts are underway to continue to work with missions as early as possible to ensure they have adequate tools and that the system streamlines the preparation and delivery of data to the PDS. Likewise, the PDS is working to define and plan for data services that will aide in searching, accessing, and using data in formats and

structures that will enhance the ability of researchers to perform analysis in cost-constrained environments.

Over the past year, the PDS has made great strides in designing and developing PDS4 including an improved set of data standards and an initial set of distributed services to support online management and sharing of data. In the coming year, the PDS is planning a series of builds to mature the system and begin working with new mission starts to baseline them on PDS4.

PDS IN CONTEXT

The PDS has two main stakeholders that interact with the system; they are the data providers and the data consumers. The data providers are represented by the missions, instrument teams and NASA-funded researchers who are involved with delivering data to the PDS. The data consumers are represented by the planetary scientists, which include those experienced with solar system exploration missions and those without active mission experience. To a lessor extent, the general public is also considered a consumer of PDS data. Figure 1 depicts the flow of planetary science data from the data providers, through the PDS system and out to the data consumers.

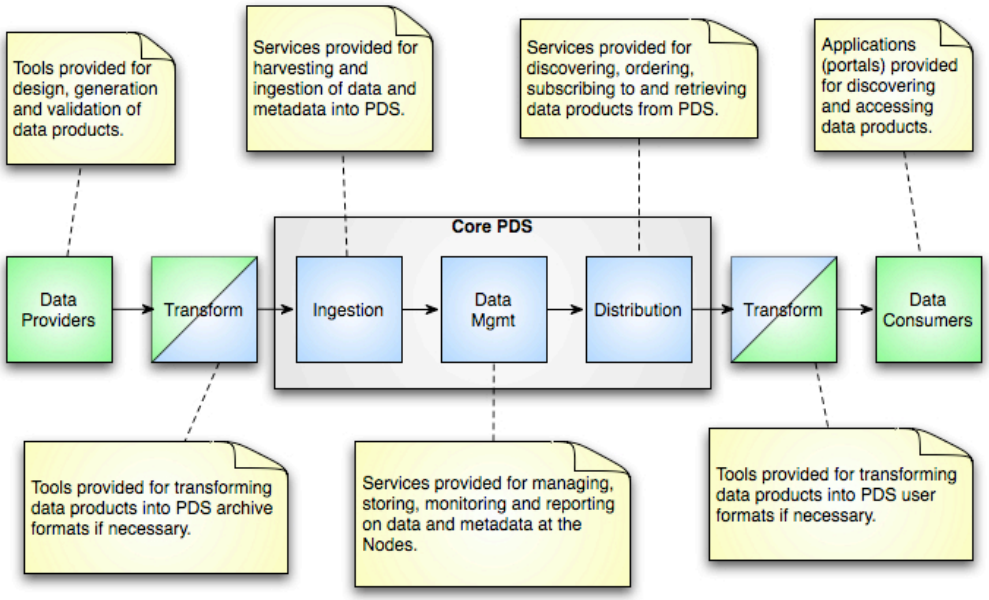

Figure 1: PDS in Context

One of the main features detailed in the diagram above involves the on-demand transformation of data products as they flow into and out of the PDS system. This allows the data to be collected in one format, archived in another and distributed in yet another format that is more user friendly for the consumers. The dual colored Transform function indicates that such transformations may occur on the client side or on the system side of the interface. For client-side transformations, PDS will provide the users with tools and software libraries for incorporating this functionality into their local pipelines. The sections that follow in this document will fill in the details regarding the data standards and technical implementation that realize the goals for PDS4.

DATA STANDARDS

The effort to upgrade the data standards for PDS4, involved transitioning a 20-year old collection of data standards [1] into a modern set of standards constructed using the best practices for standards development. The upgrade was accomplished through:

- Design of fewer, simpler, and more rigorously defined data structures for science data products.

- Use of the Extensible Markup Language (XML), a well-supported international standard [2], for data product labeling, validation, and searching.
- A data dictionary built to the ISO/IEC 11179 standard [3], designed to increase flexibility, enable complex searches, and make it easier to share data internationally.

Figure 2 depicts the four fundamental data structures that may be used for archiving data in the PDS.

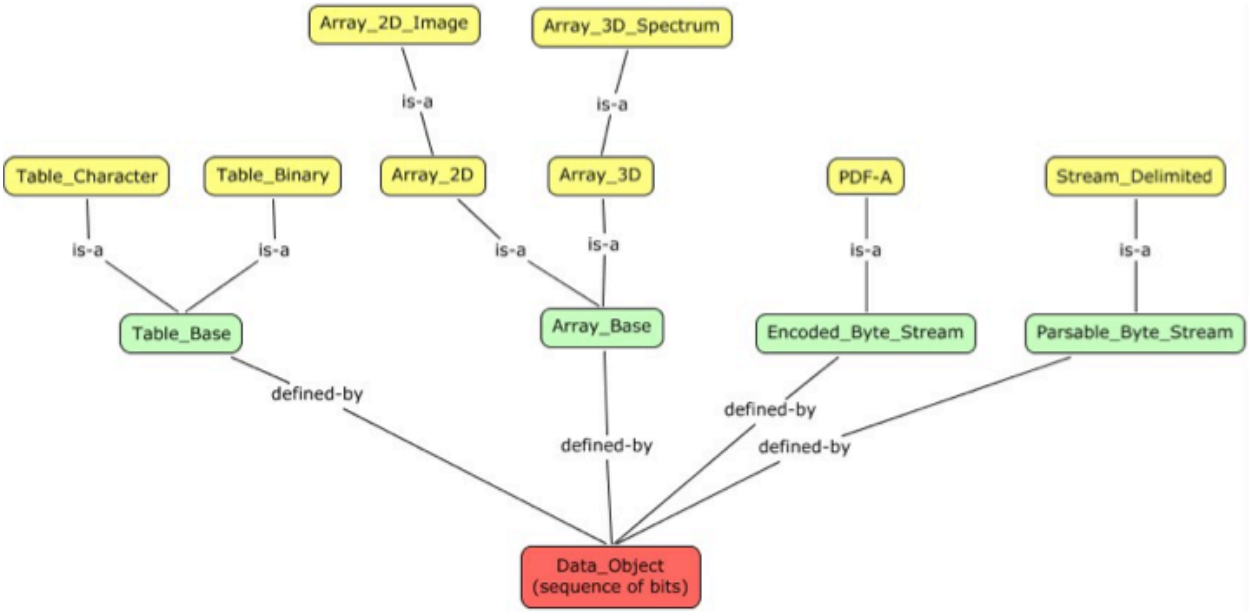

Figure 2: Base Formats and Extensions

The data structures include:

- **Table Base** - a set of repeating heterogeneous records of scalars (binary and/or character fixed-length tables).
- **Array Base** - a homogeneous N-dimensional array of scalars (images, spectral qubes).
- **Encoded Byte Stream** - data that must be interpreted according to a set of limited, PDS-authorized encoding standards; used primarily for documentation and browse files (PDF files, JPEG images, MPEG movies).
- **Parsable Byte Stream** - a stream of bytes that can be interpreted using simple parsing rules defined by a widely recognized standard (text files, XML files, CSV tables).

PDS4 data objects will actually be defined using extensions of these four fundamental structures. For example, there are Table Character and Table Binary extensions of Table Base, each with its own set of required and optional keywords, associations, and child classes. But every table in PDS4 can ultimately be traced back to the Table Base definition.

These data objects are described with labels using XML. XML provides a fixed and well-defined syntax for creating document structures and is widely utilized in software systems for information exchange. XML is only useful as an interchange standard if both the sender and receiver are cognizant of the definitions of the various keywords, called tags, used in the XML document. XML defines the syntax for the tags, but not their names or meanings. PDS uses XML Schema to define those tags and constrain their content and organization. The XML Schema standard [4] is widely supported, having a number of predefined data types useful to PDS, and offering support for the types of constraints and structure definitions required in PDS4 labels.

The keywords specified in the XML labels are defined in one or more data dictionaries. While there will still be a common PDS-wide data dictionary, PDS4 data dictionaries may also be domain or mission specific with each dictionary managed by its own steward. This delegation of authority alleviates the need for PDS-wide review of mission-specific keywords while facilitating intra-mission and intra-node cross-

correlation. Along with delegation of authority, the ISO/IEC 11179 standard has been adopted to provide a standard model for representing the PDS data dictionaries. The standard provides a common structure, a mechanism for defining keywords and promotes a common understanding of data definitions within and across organizations, including internationally.

All of this is facilitated through model-driven development. Figure 3 depicts the model-driven process.

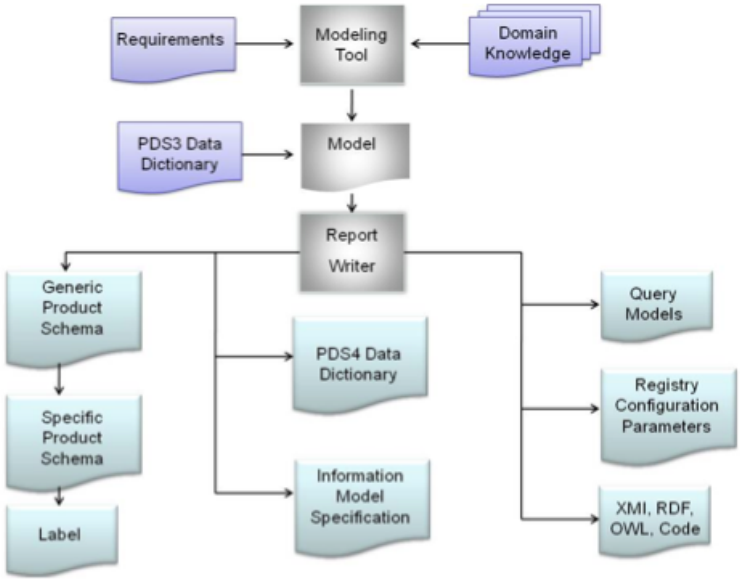

Figure 3: Model-Driven Process

The process begins with requirements and domain knowledge captured in a modeling tool. PDS is currently using Protégé for capturing the PDS4 information model. The model is updated frequently to reflect design decisions but at any time during the build cycle the artifacts detailed in Figure 3 can be generated including the product schemas, data dictionary and information model specification. Also included in the artifacts, is a configuration file for the registry service detailing the valid product types and associations for ingestion into the system. The registry service will be discussed further in the next section.

In the PDS4 information model, everything is considered a product. All products are treated in an equal manner requiring the same rigor with respect to labeling and documentation. Figure 4 depicts the product in context.

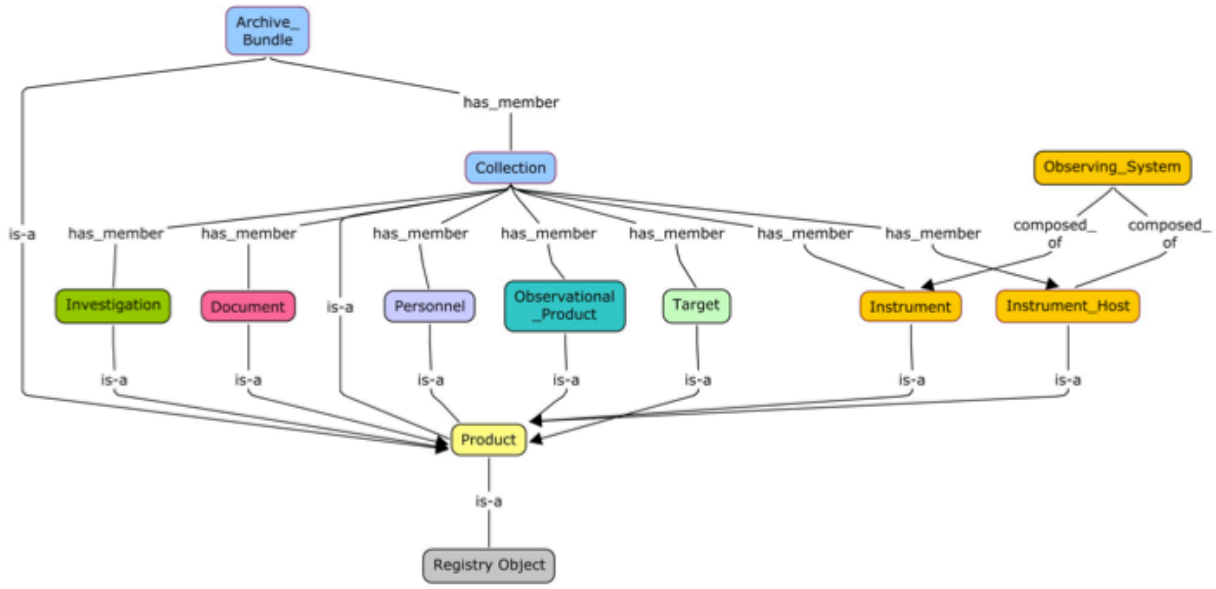

Figure 4: Product Context

Defining everything as a product ensures the ability to cross-reference products throughout the archive holdings. Additionally, registration of every product with an instance of the registry service throughout the system facilitates tracking and reporting of the archive holdings.

TECHNICAL INFRASTRUCTURE

The technical infrastructure for PDS4 offers a balance between a distributed system and a centralized system. The system includes distributed services for product registration and search to be deployed at the PDS Discipline Nodes along side their local data repositories. The system also includes centralized services for high-level search, tracking, reporting and monitoring to be deployed at the Engineering Node and made available to the Discipline Nodes. The following core concepts drive development of the system:

- **Service-Based Design** - supports remote access to data and services to bring the federation together both for ingestion and distribution.
- **System of Registries** - supports improved tracking and access.
- **Common Search** - a publicly available layer facilitating search across PDS.
- **Enhanced Tool Suite** - a tool-based approach is still appropriate for certain functions.

Service-based design in the system will focus on public interfaces for search, retrieval and value-added processing (science services) of data. Where web-based service interfaces are planned, a REST-based interface will be implemented. REST stands for Representational State Transfer, which is an architectural style for distributed systems [5]. Other services that integrate commercial off-the-shelf or open source solutions, such as a directory service, will utilize their provided interfaces.

In order to facilitate tracking, auditing, locating, and maintaining artifacts within the PDS, a system of registries has been designed. In the PDS, artifacts include data products consisting of data and label files, schemas, dictionary definitions for objects and elements and service definitions. The following types of registries were identified to capture these artifacts:

- **Inventory** - captures catalog and product metadata.
- **Dictionary** - captures the data dictionary, which consists of object/element definitions and their associations.
- **Document** - captures project documents, product label schemas, etc.
- **Service** - captures descriptions of PDS services and their associations with data collections.

Instead of designing a different registry for each of the above types, a single registry implementation was developed based on the Electronic Business using Extensible Markup Language (ebXML) standards [6,7].

With all of the artifacts in PDS captured in the registries, the system includes a common search service for discovering these artifacts. This service serves as the public interface for information contained in the registries. The metadata in the registries represents the contents of the archive holdings. This service allows product metadata to be annotated for the purposes of search, which may include updated geometry for products based on improved calibration or incorporation of taxonomies to enhance product discovery. The service also provides a common search protocol that facilitates parameter passing and integration of search amongst the Discipline Node search services.

Although the PDS4 system employs a service-oriented architecture, there are still certain functions within the system where a tool-based solution is appropriate. These include:

- **Design** of product label schemas
- **Generation** of product labels in a pipeline
- **Validation** of products and collections of products

- **Transformation** of product formats (labels and data)
- **Visualization** of data

These capabilities are provided in the form of Java and Python-based software libraries with command-line and graphical user interfaces where appropriate. Portions of the library, especially in the area of data transformation, will be open source enabling contributions from the PDS community.

Figure 5 depicts the layered architecture utilized for PDS4 providing details as to how the different components (services, tools and applications) in the system interact and build on each other:

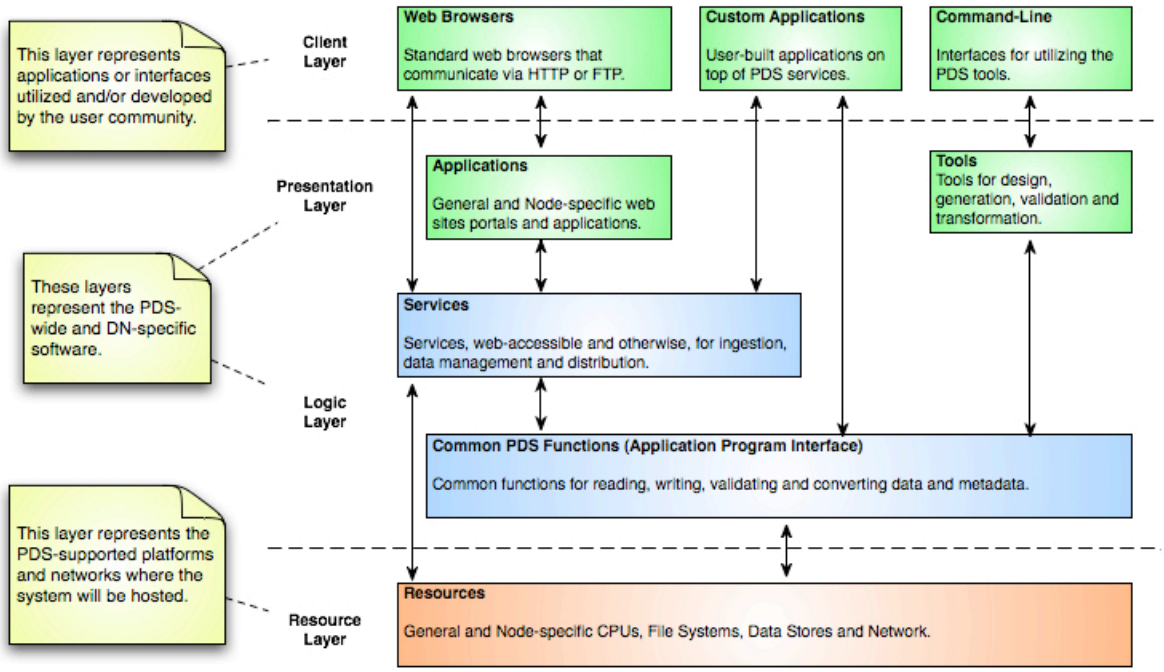

Figure 5: Component Layering

Adherence to this layered architecture, allows common functionality to be developed once and then utilized by the various components of the system as well as for user-developed applications. Not only does this approach reduce development costs, but it also improves reliability and consistency of the software within the system.

Beyond the tools detailed above, the following components are central to the function of the system in PDS4:

- **Harvest Tool** - provides functionality for capturing and registering product metadata. The tool will run locally at the Discipline Node to crawl the local data repository in order to discover products and register associated metadata with the registry service.
- **Registry Service** - This is the single registry implementation based on ebXML described previously in this paper.
- **Search Service** - provides functionality for accepting queries from data consumers for registered products and retrieving search results. This is the implementation of the common search capability described previously in this paper.
- **Transport Service** - provides functionality for moving data across the network. This service provides access to search results (products) for data consumers.

Figure 6 depicts the components identified above in a configuration detailing population of the registries and facilitation of search throughout PDS.

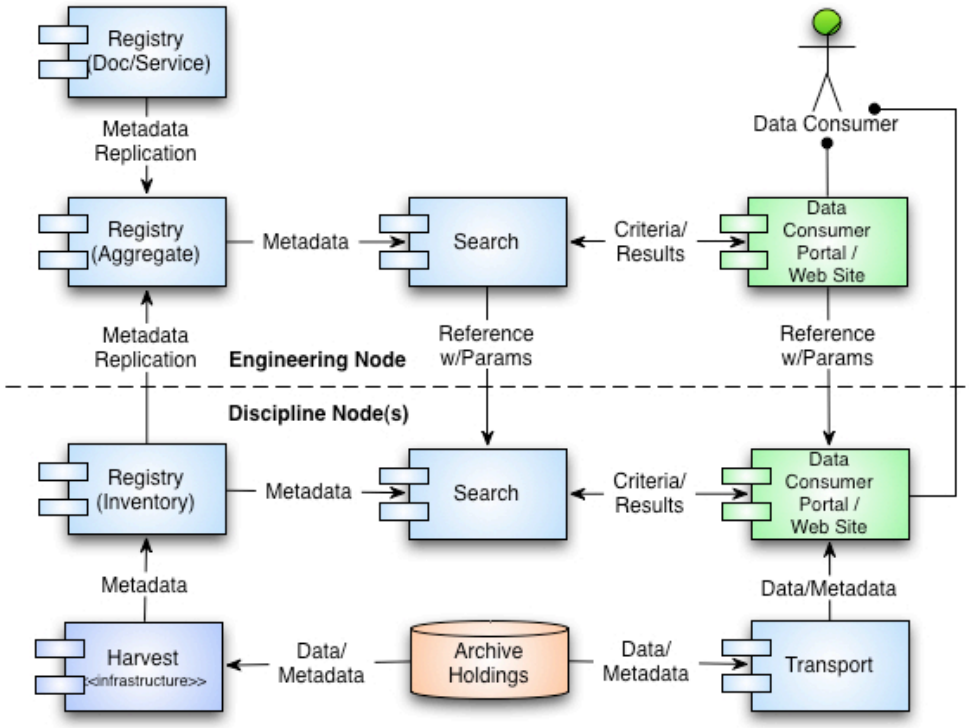

Figure 6: Registry Population and Search

Each Discipline Node in the PDS maintains a local data repository containing the archive holdings for which they are responsible for curating. As new submissions of data are received from the data providers, the Node will run the harvest tool against those submissions in order to register the new products with their local registry service. The local search service periodically pulls metadata from the registry service to populate a locally customized search index, which is designed to satisfy discipline-specific search requests through the Node's portal. Requests involving download of data are satisfied by the transport service including requests for transformation and packaging of data.

At the Engineering Node, the aggregate registry service periodically pulls new and updated entries from the registry service instances at the Discipline Nodes. This information is utilized to populate the search service that satisfies the catalog-level search capability hosted at http://pds.nasa.gov. Besides being able to access catalog-level information for data collections, instruments, targets, etc., users are also provided with a list of product-level search tools and interfaces corresponding with their search criteria. When redirecting or forwarding the user to one of these product-level search tools, the search protocol for PDS4 supports passing the user's initial search criteria (parameters) to the target search tool. This alleviates the need for the user to reenter these search criteria and allows the target search tool to immediately provide the user with an appropriate result set.

CONCLUSION

Design and development of PDS4 continues with the first operational build scheduled for delivery at the end of October 2011. This delivery supports ingestion and the initial capabilities for search and distribution. A subsequent build planned for delivery in the summer of 2012 will start to address software and tools for reading, transforming and visualizing PDS4 data products. The first missions to utilize the PDS4 standards will launch in 2013 and start to deliver data to the PDS in 2014.

In an environment where budgets are tight, the PDS4 project has continued to meet its original milestones while adhering to budget constraints and a schedule that phased its implementation and deployment over a three-year period. Although PDS4 is a significant overhaul of the data standards and technical infrastructure, efforts have been made to minimize the impact on data providers, data consumers and the Discipline Nodes. For data providers, the documentation of the PDS4 data standards is much improved over what was available for PDS3 allowing for improved communication with respect to PDS

expectations for data submissions. For data providers and consumers, several transformation options are available to offset the restricted set of supported data structures. Finally, for the Discipline Nodes, although the technical infrastructure is designed to replace existing capabilities at some Nodes, it will also integrate well with current software running at the other Nodes allowing for a smooth transition to PDS4.

REFERENCES

[1] - PDS, Planetary Data System Standards Reference, Version 3.8, February 27, 2009.

[2] - W3C, Extensible Markup Language (XML), Version 1.0, November 26, 2008.

[3] - ISO/IEC 11179, Information Technology -- Metadata registries (MDR).

[4] - W3C, XML Schema, Version 1.0, October 28, 2004.

[5] - R. Fielding, Architectural Styles and the Design of Network-based Software Architectures, 2000.

[6] - OASIS, Electronic Business using Extensible Markup Language (ebXML) Registry Services Specification (RS) v3.0, May 2005.

[7] - OASIS, Electronic Business using Extensible Markup Language (ebXML) Registry Information Model (RIM) v3.0, May 2005.

ACKNOWLEDGEMENTS

This effort was supported by the Jet Propulsion Laboratory, managed by the California Institute of Technology under a contract with the National Aeronautics and Space Administration.