The Next Generation PDS Archive Data Standards

J. Steven Hughes [(1)], Anne Raugh[(2)], Mitch Gordon[(3)], Edward Guinness[(4)], Ron Joyner[(1)], Lyle Huber[(5)], Elizabeth Rye[(1)], Dan Crichton[(1)], Steve Joy[(6)], Dick Simpson [(7)]

*[(1)] **Jet Propulsion Laboratory***

California Institute of Technology

Pasadena, CA 91109, USA

{jshughes,crichton, ron.joyner,elizabeth.d.rye}@jpl.nasa.gov

*[(2)] **University Of Maryland***

PDS Small Bodies Node

College Park, MD, USA

raugh@astro.umd.edu

*[(3)] **SETI Institute***

PDS Rings Node

Mountain View, CA, USA

mgordon@seti.org

*[(4)] **Washington University***

PDS Geoscience Node

St. Louis, MO, USA

guinness@wunder.wustl.edu

*[(5)] **New Mexico State University***

PDS Atmospheres Node

Las Cruces, NM, USA

lhuber@nmsu.edu

*[(6)] **University of California, Los Angeles***

PDS Planetary Plasma Interactions Node

Los Angeles, CA, USA

sjoy@igpp.ucla.edu

*[(7)] **Stanford University***

PDS Radio Science

Stanford, CA,USA

rsimpson@magellan.stanford.ed

ABSTRACT

The Planetary Data System (PDS) data standards were developed in the late 1980's to define the concepts and terms needed for archiving science data in the planetary science domain. Even though the data standards were innovative for their time, ambiguity has crept in after almost two decades of use and has caused significant problems for PDS operations, data providers, and end-users. Prompted by the results of an International Planetary Data Alliance (IPDA) project to identify the core requirements of the PDS

standards for adoption by the IPDA, the PDS started development of its next generation data standards, PDS4. A data design working group consisting of information technologists and science data experts from each of the PDS discipline nodes is developing the data standards with the following design goals: simplification of data formats, long-term stability in the archive, efficient archive preparation, and efficient data services including location, retrieval, reformatting and distribution. Using an ontology modelling tool, domain knowledge is captured and the contents are used to support a data-driven development methodology.

Data Standards, Information Model, Ontology, Archive, Data-Driven

INTRODUCTION

The Planetary Data System (PDS) in 2008 began a major data system upgrade project (PDS 2010). PDS 2010 will leverage modern data base and Web 2.0 technologies in order to produce a data system that ensures improved data standards and efficient, effective storage, search, retrieval and distribution of scientifically useful planetary data in the coming decades. Under the PDS 2010 project a Data Design Working Group was formed in December of 2008 to develop the next generation data standards, referred to as PDS4. Starting from first principles and leveraging over 18 years of data standards development experience, a working group of science data experts from each of the PDS discipline nodes and data engineering staff have captured the planetary science information model in an ontology modeling tool [11, 12, 13, 14]. The ontology is subsequently used to generate the documents and configuration files needed to create and validate archive quality science data products and drive the implementation of the information system. There is a broad body of research [e.g., 1-10] indicating that such an approach can be very successful in developing science information systems that meet modern expectations for information interconnectedness, correlative science, and system interoperability

As illustrated in Figure 1, the data architecture, highlighted in yellow, is one of four parts of the next generation PDS archive information system architecture. The PDS4 information model, a component of the PDS4 data architecture, is comprised of several related models – archive, data format, query, and archive organization. Although the initial focus was the preliminary development of the archive and data format models, all four models are interrelated and development of all four, while staged, proceeds in parallel in order for each to be supported by the others.

The archive data model provides the organization for PDS data and its descriptions. The model is cross-disciplinary and intended for the long term preservation as well as distribution and use of the data. It provides standards for packaging data products and for providing the mission project context within which the data resides (e.g., missions, instruments, and targets).

The data format model, new for PDS4, defines standard data structures for the data ingested into the archive. In order to reduce the complexity of PDS data formats, all data products within the archive will be stored using four fundamental component structures. Software to convert between these fundamental structures and popular modern data interpretation and analysis formats will be provided.

The query model supports data search and retrieval from both a PDS wide interface and discipline specific constraints within the distributed PDS data system.

The archive organization model addresses the organization of the data repositories.

Tightly coupled to these models is the data dictionary. The data dictionary contains the definitions of the data elements used in the models. Finally XML has been adopted as the grammar to capture the descriptive information in the archive.

The PDS4 data design group has made significant progress and has produced preliminary versions of the data format, product, context, and general query models. The group has also produced an information model specification document, generic XML schemas for each product type, data dictionary content files, and registry configuration files.

Existing standards have a significant role in the development of the system. The ISO/IEC 11179:3 Registry Metamodel and Basic Attributes specification [15] has been adopted for the PDS4 data dictionary structure. Elements from the Dublin Core Metadata Element Set are used as product attributes to facilitate the finding, sharing and management of PDS products as resources on the Web. The concept of an information object as defined by the Open Archival Information System (OAIS) Reference Model [16] is being used to unify digital, conceptual, and physical data objects in the repository. Finally the PDS4 products are being designed to support a standard federated registry reference model. [17]

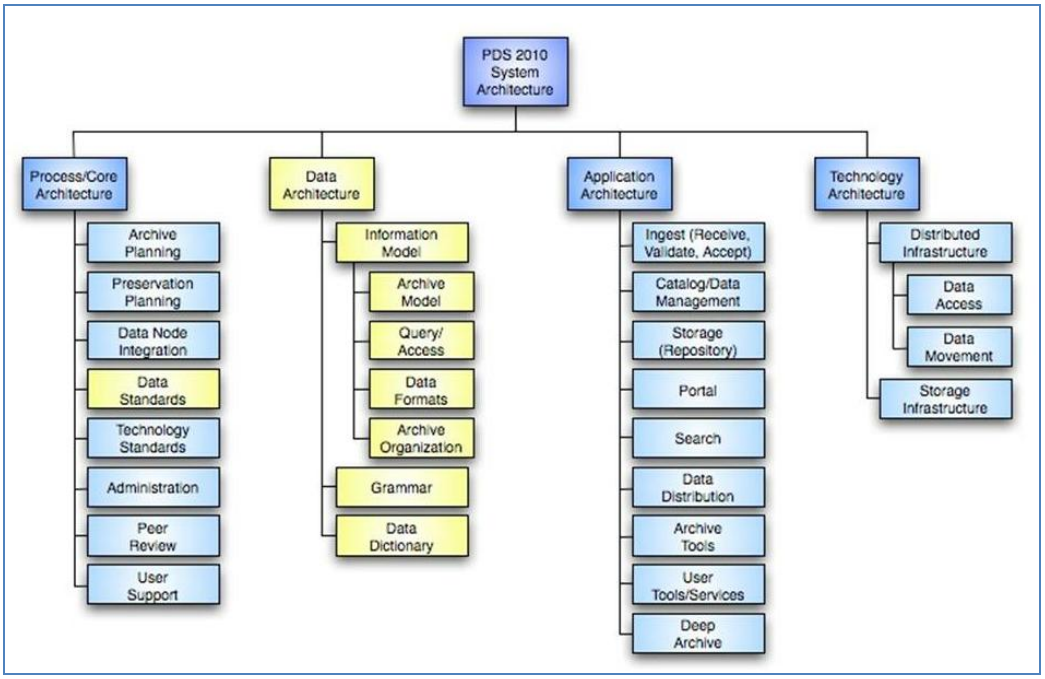

Figure 1- Next Generation PDS Architecture

DESIGN GOALS

Several design goals have been established for the PDS4 data design task. A key design goal is the simplification of the data formats allowed in the archive. Under the PDS3 data standards a standard set of descriptions was provided to describe data formats. However the data structures that were the basis of the data formats were never formally defined. To address this issue, the PDS4 data design working group defined four fundamental structures and the metadata to describe the structures. The four fundamental data structures are array, table, parseable byte stream, and encoded file. Captured in the ontology, these structural classes are formally defined, logically consistent, and can be extended and combined to handle a significant percentage of the data submitted to the PDS archive.

These few simple data structures promote long-term stability in the archive. Under PDS3 standards, almost any data structures that could be described were allowed into the archive. After nearly two decades of operations many of these structures became obsolete, superseded, or had supporting software that was too complex to maintain. Even though well documented, these data structures now need to be converted into structures compatible with the PDS architecture. For example the Digital Equipment Corporation (DEC) VAX architecture record and data formats, once widely used in the science community, are seldom used

today. However most of the early PDS Voyager data products were captured using these formats and now must be converted for use.

Restricting the archive to a few simple well-defined data structures promotes more efficient archive preparation for data providers. First, they should be easier to understand. The structural elements also tend to be defined early and remain static and are partitioned off from the interpretive elements that are generally defined later and are more volatile. Finally the PDS will be in a better position to provide low-level I/O and data structure conversion support that could save some coding effort on the part of pipeline programmers.

The proposed set of data structures also promotes code re-use, and consequently the efficient development of services. For example, a generic n-dimensional array reader can be written once and reused in the development of web services and desktop applications that display and analyze all arrays, including their extensions, 2-d and 3-d images and spectra.

Finally, the PDS4 information system architecture shown in Figure 1 suggests the separation of the data architecture from the technology and application architecture. The data architecture has been developed independent of almost all knowledge of system implementation details. This requires that generic software and services be developed with the notion that they will be configured for use using input from the information model. This approach makes parallel development of the data standards and the system software possible while also ensuring changes in the data architecture are more easily reflected in the operational system. This is especially important in science domains, where instrument design and the types of data generated constantly change.

A BRIEF OVERVIEW OF THE PDS4 INFORMATION MODEL

The OAIS Information Object is central to the PDS4 information model. It allows all data objects - digital, physical, and conceptual - to be defined and managed in a uniform way. For example, under the PDS4 data standards a digital image of Mars, the planet Mars itself and the concept of a Mars Pathfinder mission are all considered information objects.

A digital information object consists of a digital object (a sequence of bits) and its metadata. For example, an image of Mars is itself a digital object, while the elements that describe its dimensions, the time of observation, the geometry of the spacecraft, the filter used, etc., constitute the metadata. The PDS4 approach separates purely structural metadata (dimensions, data type, etc.) from the higher-level descriptive metadata (time of observation, filter, etc.), and then organizes these separate elements into data object descriptions that reflect basic form and function (an image, a table, a spectrum, and so on). This is illustrated for the case of a simple grayscale image in Figure 2.

The use of the PDS4 data structure Array_Base both defines and provides structural information about the digital object by describing how the sequence of bits is organized. The Image_Grayscale extension of the fundamental Array_Base structure is used to provide information about how the bits are to be interpreted for science data processing. This descriptive information together with the digital object and its structural information is equivalent to the *information object* defined in the OAIS reference model. *Information objects* such as Image_Grayscale and Histogram can be associated by making them members of a set. Finally one or more *information objects* are packaged as a Data_Product.

Even though the actual planet Mars and the Mars Pathfinder mission do not physically exist in the archive, from a modeling perspective they are data objects. So as in the digital object case, descriptive information must be provided. For example the planet Mars information object consists of descriptive metadata and a physical object, the actual planet Mars. Likewise the Mars Pathfinder mission information object consists of descriptive metadata and a conceptual object, the Mars Pathfinder mission. This approach unifies the digital, physical and conceptual information objects in the archive and focuses the information system's data and technology architecture on the management of information objects.

Identifiable is a generic *registry object* defined by a standard federated registry reference model. The standard federated registry model ensures that *registry objects* are uniquely identified at registration and that versioning, classification, and cataloging requirements are handled automatically. The registry model also supports such functionality as event notification and the federated query, linking, and replication of *registry objects*. As implied in Figure 2, products defined in the PDS4 information model are types of Identifiable.

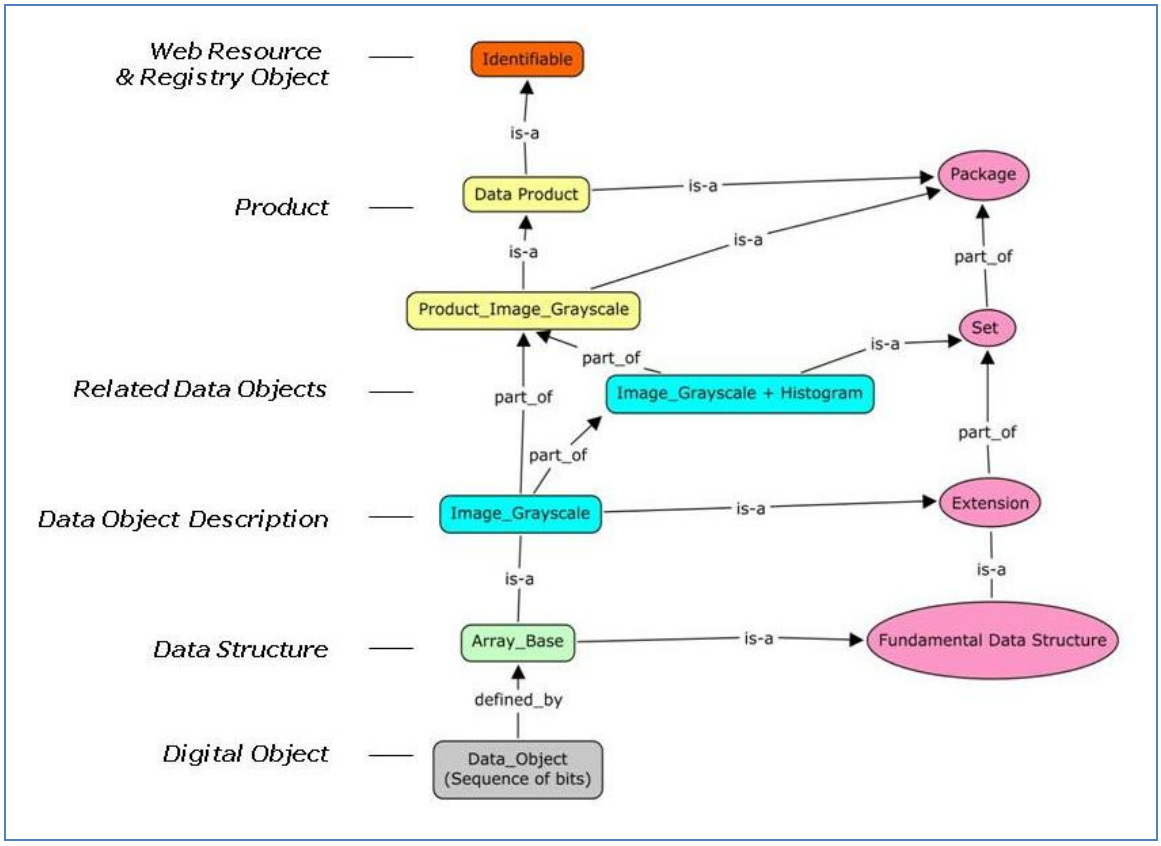

Figure 2 - Basic Components of a Data Product

The four simple data structures defined for PDS4 form a solid foundation for the information model. The current vision is that these basic structures will not change significantly over time. However, extensions to and combinations of the data structures are and will be allowed to meet the community's needs. For example, 2- and 3- dimensional spectra are defined as extensions to Array_Base; Table_Binary and Table_Character are defined extensions to Table_Base.

THE DATA-DRIVEN METHODOLOGY

The development and management of the PDS4 information model independent of the system software has been important to the success of the effort to-date. Experience suggests that in a domain as large and complex as planetary science the information modeling task alone would typically take no less than three years. The use of an ontology modeling tool and the ability to automatically generate documents for review and files for prototyping have allowed the PDS4 data design working group to quickly define the core components of the information model. Also, with only a year and a half allotted to deliver the entire system, a data-driven paradigm where parallel development can be performed is essential. This data-driven methodology is illustrated in Figure 3. The data standard artifacts are produced in the design layer and used to configure the operational federated registry system in the archive layer.

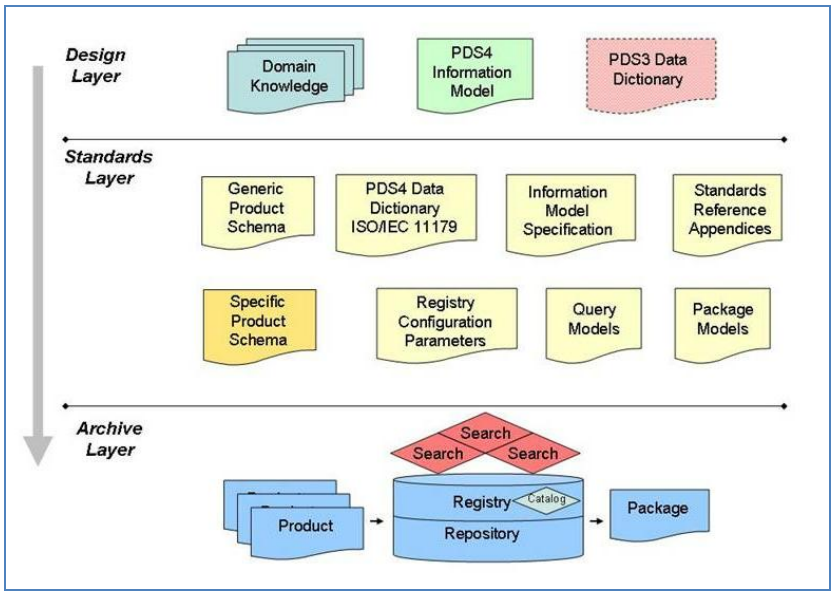

Figure 3 - Model Driven Methodology

At the design layer, domain knowledge is captured in an ontology modeling tool and used to produce the PDS4 Information Model. PDS3 data dictionary elements are reused as necessary.

Key data standard artifacts are the Generic Product Schemas. These are generated from the information model, one for each generic product type, and written as XML schemas. For example, a generic product schema is written for Product_Image_Grayscale. This schema is then used to design a Specific Product Schema for the specific product type to be ingested into the archive. For example a Mars Pathfinder Product_Image_Grayscale is defined to specify the descriptive information unique to Mars Pathfinder grayscale images and might include camera modeling information. The Specific Product Schema is subsequently used to produce and validate PDS4 product labels.

A file containing the content for an ISO/IEC 11179 compliant preliminary data dictionary is produced from the information model. The content of the information model is also used to automatically generate the PDS4 Information Model Specification document. This document represents the information model using object-oriented notation. Example labels and label schemas are also generated for inclusion in the PDS Standards Reference Document and a data engineering handbook. Finally configuration files for federated registries are generated to specify product types, classification schemes, and catalog search parameters.

The PDS4 information model also supports the development of customized searches. First, schemas for traditional forms-based search, contemporary text- and facet-based search, and advanced semantic search can be generated from the information model. Second, metadata compliant to the schemas can be harvested from the data product registries. Both the schemas and the metadata are subsequently used for search development as a service layer above the registries.

CONCLUSION

The PDS data standards were developed in the late 1980's. Even though innovative for their time, ambiguity has crept in after almost two decades of use and is causing significant problems for PDS operations, data providers, and end-users. Starting from fundamental principles and leveraging their combined experience the data design working group has made significant progress in the development of the next generation PDS data standards. Several design goals including the simplification of the data formats allowed in the archive will promote efficiencies in several functional areas. Using an ontology modeling tool and leveraging several information technology standards, the data architecture is being

developed independently and in parallel with the system software and services. The use of shared ontologies, existing standards, and the data-driven methodology are producing an archive information system that will provide information interconnectedness, correlative science, and system interoperability to support the planetary science community in the coming decades.

REFERENCES

[1] M. Uschold and G. M. Gruninger, "Ontologies and Semantics for Seamless Connectivity," SIGMOD Record, vol. 33, 2004.

[2] H. Wache, et al., "Ontology-Based Integration of Information — A Survey of Existing Approaches," In Proc. IJCAI-01 Workshop: Ontologies and Information Sharing, 2001.

[3] M. Uschold and M. Gruniger, "Ontologies: Principles, methods and applications," Knowledge Engineering Review, vol. 11, pp. 93-155, 1996.

[4] J. S. Hughes, et al., "An Ontology-Based Archive Information Model for the Planetary Science Community," In Proc. Spaceops, Heidelberg, Germany, 2008.

[5] S. Hughes, et al., "The Semantic Planetary Data System," In Proc. 3rd Symposium on Ensuring Long-term Preservation and Adding Value to Scientific and Technical Data, The Royal Society, Edinburgh, UK, 2005.

[6] H. Eriksson and M. Musen, "Metatools for Knowledge Acquisition," IEEE Softw., vol. 10, pp. 23-29, 1993.

[7] A. Cañas, et al., "Managing, Mapping, and Manipulating Conceptual Knowledge," In Proc. AAAI-99 Workshop on Exploring Synergies of Knowledge Management and Case-Based Reasoning, 1999.

[8] N. Noy, "Semantic Integration: A Survey of Ontology Based Approaches," SIGMOD Record, vol. 33, pp. 65-70, 2004.

[9] H. Knublauch, "Ontology-Driven Software Development in the Context of the Semantic Web: An Example Scenario with Protege/OWL," in International Workshop on the Model-Driven Semantic Web. Monterey, CA, 2004.

[10] G. Singh, et al., "A Metadata Catalog Service for Data Intensive Applications," in Proceedings of the 2003 ACM/IEEE conference on Supercomputing: IEEE Computer Society, 2003, pp. 33.

[11] J. S. Hughes and S. K. McMahon, "The Planetary Data System. A Case Study in the Development and Management of Meta-Data for a Scientific Digital Library.," In Proc. ECDL, 1998.

[12] J. S. Hughes, et al., "A Planetary Data System for the 2006 Mars Reconnaisance Orbiter Era and Beyond," In Proc. 2nd ESA Symposium on Ensuring the Long Term Preservation and Adding Value to Scientific and Technical Data (PV-2004), Frascati, Italy, 2004.

[13] J. S. Hughes, et al., "Preliminary Definition of the Core Archive Data Standards of the International Planetary Data Alliance (IPDA)," In Proc. PV 2007, 2007.

[14] D. Crichton: "Core Standards and Implementation of the International Planetary Data Alliance." 37th COSPAR Scientific Assembly, Montreal, Canada (2008).

[15] ISO/IEC, "ISO/IEC 11179:3 Information Technology -- Metadata registries (MDR), 2007-05-27. International Organization for Standardization, Geneva, Switzerland.

[16] "Reference Model for an Open Archival Information System (OAIS)," CCSDS 650.0-B-1, 2002.

[17] " Registry and Repository Reference Model," CCSDS W-1, 2009.

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