An Ontology-Mediated Space Science Digital Repository

J. Steven Hughes[a,1], Daniel J. Crichton[a], and Ronald S. Joyner[a]
[a] *Jet Propulsion Laboratory*
[a] *California Institute of Technology*

Abstract. The Planetary Data System, NASA's official archive for Solar System Exploration data, has transitioned to an archival information system based on ISO standards for the long-term preservation of digital data. The ontology-based PDS4 Information Model provides the informational requirements to address the system's mission to efficiently collect, archive, and make accessible the digital data and documentation produced by or relevant to NASA's planetary missions. Adopted internationally, this ontology-mediated information system brings together inputs from a variety of sources in an open and interoperable fashion.

Keywords. Digital, Repository, Ontology, Information, Model, Mediated, Science

1. Introduction

The Planetary Data System (PDS) [1] is NASA's official archive for Solar System Exploration science data. It is a federation of science discipline nodes formed in response to the findings of the Committee on Data Management and Computing (CODMAC) [2] that a "wealth of science data would ultimately cease to be useful and probably lost if a process was not developed to ensure that the science data were properly archived."

Starting operations in 1990, the stated mission of the PDS is to "facilitate achievement of NASA's planetary science goals by efficiently collecting, archiving, and making accessible digital data and documentation produced by or relevant to NASA's planetary missions, research programs, and data analysis programs."

After about twenty years of successful operations, the PDS transitioned to a more modern system [3,4,5] using lessons-learned and foundational principles from the Open Archival Information System Reference Model (OAIS-RM) [6]. The OAIS-RM states that the digital repository must define the designated community and its associated knowledge base. Complementing the key CODMAC finding that "the science community must be engaged in all aspects of a science data repository if the data are to remain scientifically useful to the community over the long term" this suggests that ontologies would be useful for capturing the planetary science knowledge base.

A task was subsequently initiated to create ontologies that would remain independent but actively drive the development and evolution of the archival system. The resulting ontologies form the core of an agile data management and curation system

[1] J. Steven Hughes, Architecture And Systems Engineering, Jet Propulsion Laboratory, 4800 Oak Grove Drive, Pasadena, California, USA; steve.hughes@jpl.nasa.gov.

having characteristics of adaptive planning, early delivery, evolutionary development, continuous improvement, and rapid and flexible response to change.

2. The PDS4 Information Model

The Protégé [7] information modeling tool was chosen for the development of two ontologies. The first ontology was an implementation of the ISO/IEC 11179 [8] Metadata Registry (MDR) standard. This standard meets many of the requirements for the detailed definitions required for science object classes and their attributes. These requirements include the ability to define domain data types, value ranges and character lengths, units of measure, and terminological names, aliases, and sources.

The MDR standard also provides strategies that help address significant issues associated with metadata governance, primarily the impact on metadata due to changes in the science community. The PDS Information Model adopts a key strategy of the MDR standard that provides and implements a multi-level governance hierarchy. The ontology is partitioned into namespaces and each namespace is governed independently by a steward.

The second or core ontology is "a representation of concepts, relationships, constraints, rules, and operations to specify data semantics for a chosen domain of discourse" [9]. It provides a sharable, stable, and organized structure of information requirements that supports an agile data curation environment. The combination of the two ontologies is called the PDS4 Information Model.

2.1. Foundational Concepts

The foundational concepts for the core ontology of the PDS4 Information Model are derived from the Information Model provided in the OAIS RM, starting with the "Information Object". Its extensions include object classes for the following information categories:

- Identification – allows information object to be discovered and accessed.
- Representation/Format - allows a data object to be interpreted.
- Fixity - ensures the information object has not been unintentionally altered.
- Provenance – provides essential authenticity of information objects.
- Context - describes the environment in which the data object was created.
- Reference - allows the information objects to be referenced.
- Access Rights - identifies the access restrictions pertaining to the data.

The PDS4 Information Model addresses these concepts that are required for long-term preservation and provides the means for the PDS mission's data and documentation to be efficiently collected, archived, and made accessible.

The knowledge acquisition phase for the core ontology of the PDS Information Model was a multi-year effort that required the collaboration of information architects and experts from the diverse scientific domains in the Planetary Science community. This effort resulted in what is called the "Common dictionary" consisting of the object classes and relationships used across the science domains. The Common dictionary, illustrated as a concept map in Figure 1, has rapidly stabilized after a few years of use.

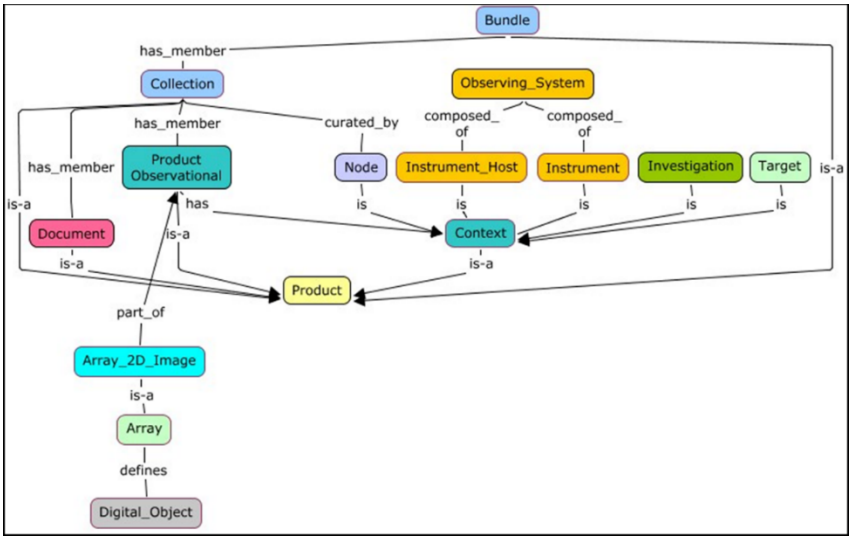

Figure 1. Common Model

Two useful characteristics result from the use of ontologies to drive agile development. First, as mentioned previously, the information architecture remains independent of the implemented system's architecture including its implementation choices. This allows the science domains and the information technology to evolve separately. This reduces the impact of change.

The second property, the Information Model's multi-level governance hierarchy, partitions the model into a common and many discipline and local models, each governed by a steward. Each steward is given significant autonomy but is constrained by the ontology's modeling principles. This creates a loosely coupled set of consistent models.

In order to improve data interoperability, an original large set of diverse and often complex data formats are reduced to a small standard set of simple data formats that are designed for long-term stability. The reduced standard set is sufficient for most planetary science data. More complex data formats are accommodated by using compositions of the standard formats. However, this conceptualization increases the complexity restrictions, such as no longer allowing the interleaving of two or more data objects. For example, engineering data may no longer be prefixed to each row of a simple raster image but must be grouped into a separate data object.

The Information Model defines one Product class. This class is extended into a set of products sufficient for the various categories of data, including observational data, ancillary data, contextual information, and documents. Each Product consists of a detached metadata label with a unique, immutable identifier and one or more data objects. The identifier can be versioned. The Product label is defined from and validated against the PDS4 Information Model. The Product label provides data format, identification, reference, integrity, provenance, and context information.

PDS4 Products may reference each other. This results in the formation of a semantic network of linked data. There are two aggregate products, Collection and Bundle. A Collection groups related "base" products. A Bundle groups related Collections.

The content of the PDS4 Information Model is filtered, translated, and written to various file formats as shown in Figure 2. Since the PDS chose XML to label Products,

the content of the Information Model is written to XML Schema and Schematron files. These files are subsequently used to create and validate the product's XML label. Product label validation is largely accomplished using the XML Schema and Schematron files, effectively tens-of-thousands of lines of auto-generated declarations.

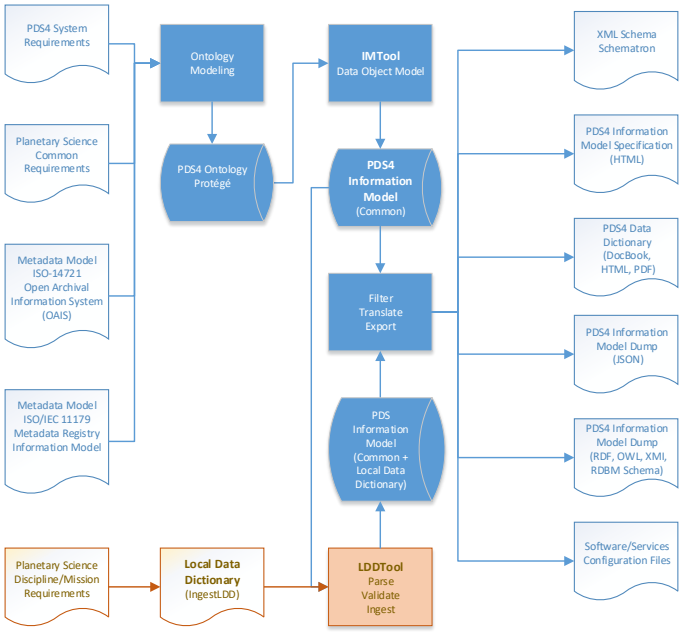

Figure 2. Information Flow

2.2. Maintaining Relevancy in a Diverse and Evolving Science Discipline

After the completion of the Common model, the development of discipline and local models started. These models are called Local Data Dictionaries (LDDs). Each discipline LDD involves specific areas of expertise, for example Cartography. To create a discipline LDD, one or more discipline specialists take "stewardship" responsibility to design and maintain an LDD for their specific area of expertise. To shield discipline experts from the complexities of the data modeling process, a modeling "template" and associated validation tool were developed that constrain the designer to a simple design methodology and selected references to classes and attributes defined in the Common model. The LDD template itself is defined in the Common model and so has an XML Schema and Schematron file. The LDD designer populates the LDD template to create the LDD.

The tool validates the populated LDD template by temporarily "ingesting" the LDD into the Common model to check for consistency. This framework allows stewards to design and maintain their models in an environment that is loosely coupled to the Common model and other LDDs.

This paradigm is repeated at the local level for missions and projects. The tool allows the "stacking" of two or more LDDs when cross-referencing is desired. References between LDDs for reuse of object class definitions are negotiated between stewards. The

resulting hierarchy is illustrated in Figure 3. Currently there are fifteen LDDs at the discipline level and nine LDDs at the mission level. However new missions and the migration of legacy mission data will substantially increase the number of LDDs at the mission level.

Common	Information Object Product, Array, Table, …
Discipline	Imaging Cartography Spectral …
Mission	InSight Cassini Voyager 1 Voyager 2 …

Figure 3. Model Hierarchy

3. Conclusion

The PDS4 Information Model consists of two stable ontologies and a rapidly expanding group of discipline and local ontologies. These mediating ontologies form the core of an independently managed model that provides the information requirements for software and services configuration. An agile development cycle of use, feedback, planning, change, test, and release keeps the PDS relevant in a diverse and evolving science discipline and ensures future planetary scientists will have useful data for new science inquiries.

4. Acknowledgements

The research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

The authors wish to acknowledge the PDS4 Data Design Working Group (DDWG) and the Systems Design Working Group (SDWG) for their significant efforts in the design, development, and implementation of the PDS4 information and system architectures. These discipline experts remained committed, sought excellence, and provided first-rate information without which PDS4 would not have been possible. The authors also wish to acknowledge Sean Hardman for his PDS Systems Development leadership and David Giaretta and the various teams responsible for the ISO standards. Finally, they would like to recognize the support of the PDS Management Council and NASA Headquarters.

References

[1] Special Issue: The Planetary Data System, Planetary and Space Science, European Geophysical Society, ISSN 0032-0633, Volume 44, Number 1, January, 1996.

[2] National Research Council. 1986. Issues and Recommendations Associated with Distributed Computation and Data Management Systems for Space Science, Committee on Data Management and Computing, Space Studies Board, National Academy Press, Washington, DC, pp. 95.

[3] Hughes, J. S., Crichton, D. J., Mattman, C. A., "Ontology-Based Information Model Development for Science Information Reuse and Integration", 10.1109/IRI.2009.5211603, IEEE International Conference on Information Reuse & Integration, 2009.

[4] Hughes, J. S., Crichton, D. J., Hardman, S., Law, E., Joyner, R., Ramirez, P., "PDS4: A model-driven planetary science data architecture for long-term preservation," Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on , vol., no., pp.134,141, March 31 2014-April 4 2014.

[5] Crichton, D., Hughes, J.S., Sean Hardman, S., Law, E., Beebe, R., Morgan, T., Grayzeck, E., "Scalable Planetary Science Information Architecture for Big Science Data", 2014 IEEE 10th International Conference on e-Science, Volume 2, 20-24 October 2014.

[6] ISO 14721:2012 - Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model, ISO, 2003.

[7] (2013) The Protégé Ontology Editor and Knowledge Acquisition System website. [Online]. Available: http://protege.stanford.edu/.

[8] ISO/IEC 11179: Information Technology -- Metadata registries (MDR), ISO/IEC, 2008.

[9] Lee, Y. T., "Information Modeling: From Design To Implementation", Proceedings of the Second World Manufacturing Congress, pp 315-321, 1999.