PDS4 DATA MODEL WHITE PAPER
January 14, 2008

Todd King (UCLA)
Steven Hughes (JPL)
Ed Guinness (WashU)
Chris Isbell (USGS)
Lyle Huber (NMSU)

1. Introduction

At the August 2007 meeting of the PDS Management Council three working groups were formed to define different aspects of the next generation of the PDS system (PDS4). One of those working groups was the Data Model working group. Members of the working group include Todd King, Steven Hughes, Ed Guinness, Chris Isbell, and Lyle Huber. The working group held several telecons, conducted e-mail discussion, presented draft proposals and analyzed existing data models. At the December 2007 Management council meeting the Data Model Working group conducted a day long face-to-face session where a draft information model was envisioned. A discussion of importance of data modeling and the results to date of the Data Model Working Group is presented in this white paper. More details of the work of the PDS4 Data Model group is captured at

http://pds-engineering.jpl.nasa.gov/index.cfm?pid=100&cid=120.

What is a data model?

One of the fundamental questions with data modeling efforts is "What is a data model?" The answer to this question can vary and the differences are related to the tendency to used the terms "information model" and "data model" interchangeable. In fact, there is significant difference between the two. An "information model" is an abstract but formal representation of entities including their properties, relationships and the operations that can be performed on them. Where as, a "data model" is the expression of the information model for use in a software implementation. That is, an "information model" is implementation neutral, and a "data model" is implementation specific. From a logical/physical viewpoint an "information model" is a logical model and a "data model" is a physical model derived form the "information model". While developing a data model for use by PDS for its next generation implementation is the goal of the PDS4 Data Model Working Group, the group is actually developing an information model form which a data model can be derived.

Why do we need a data model?

Another question that arises is "Why do we need a data model?" For information systems an information model is a conceptual representation of the entities along with their attributes and operations which comprise the body of knowledge (corpus) which is stored and operated on within the system. An information model serves as a guide for the design and development of the system. When an information model is formally specified it can be expressed as application specific data models. For example, an information model can be expressed as a SQL schema, as Java classes, or an XML schema. Formally specifying an information model can make

management of the model easy since revisions can be tracked and consistency within the model can checked (enforced) with existing modeling tools.

2. Principles

The development of the PDS4 system follows a set of principles so that individual efforts such as those of the PDS4 Data Model Working Group can be coherent with the overall vision. Some principles related to the PDS4 information model are:

Interoperability – The PDS works to ensure interoperability among planetary science archives by seeking community consensus on a core set of common objects and data elements.

Partitioning - The data model is logically separated into partitions in order to allow for management and evolution of components of the data model independently. For example, the image model is managed by the imaging community.

Formal Specification - The data model is explicitly and unambiguously defined using a formal data engineering notation and/or language.

Standards - PDS applies commonly accepted and documented standards that address fulfill its requirements.

Evolvability and Flexibility – The data model should be extensible and flexible enough to meet new requirements.

Model Expressions – The data model is implementation neutral and can have different expressions to support subsystem functions.

a) Database development. A database developer needs an Entity-Relationship, UML model or schema.

b) User Interface. A web developer needs a taxonomy for navigation and keyword/values for facet-based navigation system.

c) Persistent storage. Instances of all or part of the data model needs to be expressed in a form that can be stored indefinitely. (XML, ODL, etc)

d) Model-to-model mapping. Where possible and necessary, translations from the PDS data model to other data model expressions in order to support the interchange of information.

e) Notation conversion. The model can be expressed using a number of notations or languages. (e.g. Ontology, UML, ER, Taxonomy, etc)

3. Other Considerations

The Management Council has stated that backward compatibility of PDS4 to PDS3 is not a requirement. The Management Council (at its December 2007 meeting) also directed the

working group to define a PDS4 Information Model independently from the PDS3 specification. The assessment of costs or impact will be performed after the PDS4 specification is complete.

Another consideration was what formats should be supported by the Information Model. It was decided that the Information Model should be format neutral and that any requirements for archive or exchange formats would be decided outside the Data Model Working Group.

4. PDS4 Information Model

The process of developing an information model follows these steps:

1. Identify your domain
2. Determine the partitions of domain.
3. Identify the entities in each partition
    a. Identify relationships of entities
4. Identify the attributes for each entity.

Prior to the December 2007 Management Council meeting the efforts of the working group were focused on identifying the issues and problems related to the PDS3 data model. A concurrent, non-PDS effort (The Rosetta Model) to assess existing data models [1] was occurring and portions of the effort were reported to the working group. At the December 2007 Management Council meeting the rough outline of Rosetta Model was presented to the working group. The working group used this outline as a starting point to define the PDS4 Information Model. To date, the efforts of the working group has reach step 3 (Identify entities). Each of the completed steps is described as follows.

4.1. PDS Domain

For the PDS an enduring data model is a critical system component since the primary function of PDS is the preservation and distribution of information. Developing and maintaining an enduring data model as one expression of a PDS information model is core to the PDS enterprise because it is the descriptions in the selected data model which are included in the archived. This is expressed in the PDS Roadmap [2] which describes PDS archives as:

> The PDS archives and makes available space-borne, ground-based, and laboratory experiment data from over 50 years of NASA-based exploration of comets, asteroids, moons, and planets.
>
> The archives include data products derived from a very wide range of measurements, e.g., imaging experiments, gravity and magnetic field and plasma measurements, altimetry data, and various spectroscopic observations.

These artifacts of the Roadmap define the PDS domain. It is a very broad domain because of the diversity of sources and type of data products. This presents some challenges, but PDS has the advantage of many domain experts with decades of experience archiving data.

4.2. Model Partitions

The PDS4 Information Model has been divided into the following top-level partitions.

Participants
: Individuals, organizations or objects which contribute to or define the context for other defined entities.

Products
: Entities which result from an observation or analysis; provide additional supportive information; or describe the operation or analysis process.

Resources
: Entities which are components of the system.

Collections
: Groupings of entities.

Query
: Additional information to support locating entities in the system.

4.3. Model Entities

Under each of the partitions a set of entities have been defined. The full outline of the entities is:

Participants

- Mission
- Observatory
- Instrument
  - Detector
- Person
- Reference
- Target

Product

- Sample (Physical)
- Data Structure (Digital)
  - Catalog (record collection)
  - Table (row, column)
  - Image (x, y, z)
  - Movie (x, y, z, t)
  - n-Array
  - Compound Structure (?)
- Documents

Resource

- Repository
- Registry
- Web Link
- Service

Collection

- Dataset
- Event
- Campaign

Query

Details regarding each entity will be provided in a subsequent report.

5. The Core Questions

In November 2007 the Management Council defined a set of questions for PDS4 Working Groups to respond to.

1. How will PDS-4 enable "one-stop shopping", i.e., seamless access to data that reside at multiple nodes?

   Adopting a common information model makes distributed searches more achievable, integration of services more possible and interpretation of data easier. Seamless access requires a seamless model. The model should specify the metadata needed to support science, engineering, administration and operations. We need to capture the context in which data are acquired to aid in its interpretation. We need to preserve information relevant to the data.

2. How will PDS-4 help users by delivering derived data products in the format, coordinate system, and map projection the user requests?

   The information model must include models for supported formats, coordinate systems and map projections. The information model must support the delivery, conversion and transformation services required to respond to a user request. The information model should be based on stated requirements for supported formats, coordinates systems and map projections. An efficient and lean information model containing those terms that add value and benefit the system is the goal.

3. How will PDS-4 help data providers by automating the design, production, and delivery of PDS data sets?

   A formally defined information model for specifying entities and relationships will make creating compliant data sets easier and more efficient. Expressing the data model in application appropriate languages (XML Schema, ODL, OWL) will enable adoption in a variety of environments. Any tools developed which rely on the formal specification should be portable (useful) to other providers.

4. How will PDS-4 ensure that PDS standards are simple, straightforward, and consistent so that data providers and users can easily understand and apply them?

   The Information Model should be expressed as a formal specification in a data engineering language. The formal specification must be translated into a version that

users can easily understand and apply. The Information Model should be prescriptive with an accommodation to allow descriptive information (additional terms) to be included, but clearly separated from the PDS Information Model. The PDS Information Model should include only those terms with a well specified purpose and clear intent.

5. How will PDS-4 ensure that data sets can be safely and efficiently archived in NSSDC and retrieved on demand?

   While the data model aids in archiving by preserving information using standardized terms, it does not aid directly with NSSDC.

6. How will PDS-4 improve the data transfer, data integrity, and maintenance of PDS data sets?

   A well specified information model will ensure coherent transfer of data and allow for checking data integrity and long term maintenance.

6. Recommendations

That the PDS Management council approves continuing the PDS4 Data Model Working group with the objective of providing a formal specification of the information model outlined in this white paper. The goals of the working group are to:

1. Compile known issues and problems associated with the current model; determine probable cause.
   This will be assembled into a separate document and used to evaluate the effectiveness of the proposed PDS4 Information Model in addressing the problems and the limitations of the PDS3 data model.
2. Define the relationships
   The outline of the information model presented in this paper only identified the top level entities. The relationships between entities is the next step in defining a fully specified information model.
3. Define attributes for each Entity.
   More detail is needed for each entity in order to derive a usable data model from the information model. The additional detail is needed in order to define a migration path from the PDS3 data model to the proposed PDS4 data model. Additional details are also needed to assess the impact and cost of implementation including the migration of data.
4. Vet with tech group.
   The more individuals who evaluate the Information Model the more likely it will address the needs of the planetary science community.
5. Draft Information Model Specification document (must include PDS3->PDS4 mapping)
   A formal specification will allow a more rigorous analysis.
6. Final report (recommendation for control authorities, partition alignment architecture decomposition)
   Upon delivery of a formal PDS4 Information Model specification there should be a well defined path to move the effort from a working group to an operational setting. The final report should include recommendations on how to proceed.

7. Summary

This white paper is intended to serve as a report back to the Management Council regarding the efforts of the PDS4 Data Model Working Group.

8. References

[1] The Rosetta Model, Poster, Fall 2007 AGU.
[2] PDS Roadmap February 2006