June 10, 2009

PDS4 Information Model Requirements

PDS4 Data Design Working Group

History:

- First release December 11, 2008

Reviews:

- Review; June 10, 2009

Preface

Members of the PDS4 Design Working Group

Steven Hughes (Chair)
Mitch Gordon
Ed Guinness
Lyle Huber
Ron Joyner
Anne Raugh
Elizabeth Rye
Dick Simpson - Observer

Table of Contents

1. Scope

This document defines the requirements for the fourth generation of the Planetary Data System (PDS) Information Model -- "PDS4". The intent of the PDS4 Information Model is to provide an enduring data model that can describe the relevant attributes of data and ancillary information collected as part of planetary science activities. This information is to be included in the Planetary Data System archive that is described in the PDS Roadmap [1] 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.

The PDS has defined a set of level 1, 2, and 3 requirements [2] that set the scope for the PDS4 Information Model.

2. Audience

The expected audience includes the following member of PDS:

- Managers and administrators
- Engineering Staff
- Technical Staff

and other individuals who wish to define the scope and extent of the PDS4 Information Model.

3. Introduction

The PDS4 Data Design Working Group (DDWG) has compiled a list of requirements for the PDS4 Information Model. These have been derived from the PDS Level 3 requirements and compiled from the problems, issues, anomalies and items of note captured during the PDS3 Information Model review.

The DDWG established a baseline set of objectives for the PDS4 Information Model Architecture:

1. Consistency and robustness -- the model should define and enforce business rules

2. Ease of maintenance -- corrections in documentation should only be actioned once, with automatic propagation of the change throughout the model. It must be possible to upgrade easily to a new version of the base standard

3. Ease of explanation -- it must be possible to communicate the model to users having little experience in information modeling

4. Requirement for artifacts -- it must be defined which artifacts are generated and how and where should they be deployed

5. Auto-generation of artifacts -- schemas, scripts and documentation must be automatically generated from the model

6. Normative elements -- normative elements of the Information Model, associated artifacts, and inter-relationships must be defined

4. PDS4 Information Model Requirements

The PDS4 Information Model requirements have been derived from the PDS Level 3 Requirements and additionally compiled as an artifact of the PDS3 Information Model review.

1. The Information Model shall be developed and maintained independent from any specific technology choices, implementations, or expressions.

   Rationale: One of the major objectives of any Information Model is platform / implementation independence -- the model, the contents of the model, and the systems modeled can represent one or more different types of actual physical systems. All implementation (physical) models will be derived from the logical information model; including, data models for subsystem implementation, applications for generating and validating PDS data holdings, and expression of the information model in any language or notation.

2. The Information Model shall be defined using a formal data modeling notation.

   Rationale: One of the major objectives of any Information Model is ease of maintenance -- corrections should only be actioned once, with automatic propagation of the change throughout the model.

3. The Information Model shall encompass all PDS stakeholder viewpoints, including the contextual, conceptual, logical and physical.

   Rationale: Requirements definition, a critical activity within information systems development, involves many stakeholder groups: managers, various end-users, and different systems development professionals. Each group is likely to have its own ‘viewpoint’ representing a particular perspective or set of perceptions of the problem domain. The Information Model ensures, as far as possible, that the systems meet the needs and expectations of all involved stakeholders. The Information Model becomes

a conceptual framework for understanding and investigating viewpoint development approaches; as well as, managing potential inconsistencies and conflicts.

4. The Information Model shall be rigorous and prescriptive.

   Rationale: A prescriptive approach, in its most rigorous application, specifies that which is correct, sanctioned, and authorized. The Information Model will establish "correct and unambiguous usage" of data maintained in the PDS archive.

5. The Information Model shall allow the definition and inclusion of data models from many domains.

   Rationale: The Information Model will be able to reference, adopt, and/or incorporate data models from other domains which will promote interoperability and the use of common standards.

6. The Information Model shall formally define "context" classes.

   Rationale: The Information Model defines "classes" (e.g., Mission, Instrument, Data Set, Document, etc) and identifies the viewpoints of the context classes within explicit boundaries. The various expressions of these classes are left to implementation.

7. The Information Model classes shall formally define relationships between classes.

   Rationale: As many of the PDS3 classes were not formally related (e.g. document and data set), the Information Model will specify the normative elements and the inter-relationships between the classes.

8. The Information Model shall define a standard set of "attributes" and "attribute values" for describing data maintained in the PDS archive. [PDS 1.4.2]

9. The Information Model shall define a set of data formats that are based on standard data structures. [PDS 1.4.2]

   Rationale: The Information Model will define the number of PDS approved data formats.

10. The Information Model shall be tightly coupled with a data dictionary to capture information that is not captured within the Information model. [PDS 1.4.2]

    Rationale: Many Information Models do not capture the breadth of information required for defining an individual attribute.

11. The Information Model shall support long term preservation (e.g., archiving) of the data in the PDS archive. [PDS 4.1]

    Rationale: The Information Model has the overall responsibility for capturing information relevant to the long term preservation and archive of the data in the PDS archive.

12. The Information Model shall support long term use of the data in the PDS archive. [PDS 4.2]

    Rationale: The Information Model has the overall responsibility for capturing information relevant to the long term use of the data in the PDS archive.

13. The Information Model shall support distributed discovery and access to the data in the PDS archive. [PDS 2.8.1; PDS 2..8.2; PDS 2.8.3; PDS 3.1]

    Rationale: The Information Model has the overall responsibility for capturing information relevant to the discovery and access to the data in the PDS archive.

14. The Information Model shall support cataloging the holdings in the PDS archive. [PDS 2.6.2]

    Rationale: The Information Model has the overall responsibility for capturing information relevant to cataloging the holdings in the PDS archive.

15. The Information Model shall support tracking of the data in the PDS archive. [PDS 2.2.2; PDS 2.4.5]

    Rationale: The Information Model has the overall responsibility for capturing information relevant to tracking the data in the PDS archive.

16. The Information Model shall be maintained in accordance with the PDS Data Standards process. [PDS 1.4.6]

    Rationale: The Information Model has the overall responsibility for capturing information relevant to tracking the data in the PDS archive.

17. The Information Model shall formally define the following classes: Data Set, Product, Mission, Instrument, Host, Target, Node, Person, Reference, Document, and Software.

    Note: 1.4.4 PDS will establish minimum content requirements for a data set (primary and ancillary data)

18. The Information Model shall formally define the following classes: data object, data structure, data interpretation, data identification, and data metadata.

    Note: 1.4.1 PDS will define a standard for organizing, formatting, and documenting planetary science data

19. The Information Model shall formally define a data dictionary model.

    Note: 1.4.2 PDS will maintain a dictionary of terms, values, and relationships for standardized description of planetary science data

20. The Information Model shall have one or more grammars into which to express the information model classes.

    Note: 1.4.3 PDS will define a standard grammar for describing planetary science data

21. The Information Model shall formally define the following classes: resource, release, and housekeeping.

    Note: 2.6.2 PDS will design and implement a catalog system for managing information about the holdings of the PDS – Assume that classes derived from 1.4.4 also support 2.6.2.

22. The Information Model shall formally define the following classes: repository, registry, and identifiable.

    Note: 3.2.1 PDS will develop and maintain online mechanisms allowing users to download portions of the archive

23. The Information Model shall formally define the following classes: coordinate system.

    Note: 3.3.4 PDS will provide tools for translating archival products between selected coordinate systems

24. The Information Model shall formally define the following classes: manifest.

    Note: 4.1.2 PDS will develop and implement procedures for periodically ensuring the integrity of the data.

Appendix A – PDS Level 3 Requirements

The Level 1-3 requirements that relate to the PDS4 Data model are:

1.4 Archiving Standards: PDS will have archiving standards for planetary science data

- 1.4.1 PDS will define a standard for organizing, formatting, and documenting planetary science data
- 1.4.2 PDS will maintain a dictionary of terms, values, and relationships for standardized description of planetary science data
- 1.4.3 PDS will define a standard grammar for describing planetary science data
- 1.4.4. PDS will establish minimum content requirements for a data set (primary and ancillary data)
- 1.4.5 PDS will establish minimum sets of archival data from missions and other major data providers
- 1.4.6 PDS will develop, publish and implement a process for managing changes to the archive standards
- 2.1.2 PDS will track the status of data deliveries from data providers through the PDS to the deep archive

2.3 Peer Review: PDS will conduct peer reviews of all data submissions to ensure completeness, accuracy, and scientific usability of content.

- 2.3.5 PDS will track the status of each peer review

2.4 Acceptance: PDS will accept or reject submitted data.

- 2.4.1 PDS will develop and publish procedures for accepting archival data

3.1 Search: PDS will allow and support searches of its archival holdings

- 3.1.3 PDS will provide a means to locate caveats and errata associated with archival data sets, volumes, and products

4.1 Long-Term Preservation: PDS will determine requirements for and ensure long-term preservation of the data

- 4.1.1 PDS will develop and implement procedures for periodically ensuring the integrity of the data

Not Used

- The information model shall endorse the set of PDS supported technology platforms.

25. The Information Model shall be unambiguous.

    Note: Standards are unclear as to requirements versus recommendations; have not separated requirements from recommendations (e.g. and standards from policies). There is no clear set of requirements for archives. This is mostly an editorial issue with respect to the contents of the standards reference document. It would be preferable to have all policies clearly differentiated from the standards.

26. The Information Model shall formally define instrument so that it correctly models the radio science instrument.

    Note: The radio science instrument has components on both the ground and on a spacecraft is poorly modeled.

27. The Information Model shall specify globally unique identifiers for any class with instances that can be referenced from outside the PDS.

    Note: Individual products are not easily relocated. The interpretation of some attributes in a product description is dependent on an external file system organization or the dataset/volume context. This makes it difficult (or impossible) to form new collections (possibly based on a user's selection) consisting of portions of existing collections. The "volume" structure which requires different types of products to be stored in fixed locations (documents in "document" folder, data in the "data" folder) can result in file name conflicts when products from multiple "volumes" (or datasets) are combined into a new collection. This is an issue more for the delivery of archived products, rather than for the archive itself. An "archive system" manages and stores products and collections "as delivered" to the system.

28. The Information Model shall formally define Target.

    Note: All targets (planets, satellites, small bodies, etc.) are treated as a single category. Because of the large number of targets the use of the list is cumbersome for providers. Also, some targets which are different in class (planet, satellite, small body) share the same name. Other organizations also have standard names for bodies. Examples of targets that share names include Halley (comet) and Halley (asteroid) or Amalthea (asteroid) and Amalthea (moon of Jupiter). The Small Bodies Node has defined a set of formation rule for creating target names

eliminate ambiguities, but this method is not universally applied in PDS, hampering the use of the target name in locating related resources. There is no mechanism to describe observed regions such as fixed geographic areas or features in atmospheres. The GAZETEER object has the potential to address this issue, but the object definition refers to the " PDS Data Preparation Workbook" for the specification of the GAZETEER table record format. This document is not longer available from PDS.

- The Information Model shall formally define Products.

  Note: Software, document, SPICE, and data products all need to be formally defined and treated as first class products. - The PDS Information model did not formally define products. This resulted in many strange variations. For example, products consisting of multiple files (compound products) were either poorly supported or impossible to describe. A product consisting of multiple files organized in a directory tree cannot be described because paths are not allowed in pointers. The current standards need to better handle compound products.

- The Information Model shall formally define File.

  - Note: The implicit file object either needs to be eliminated or made explicit. The handling of RECORD_TYPE definitions for files containing multiple objects of different types should be better defined. Similarly, DATA_OBJECT_TYPE is a nonsensical required keyword in DATA_SET_INFORMATION except for completely homogeneous data sets. There are standards based on archaic hardware / software - (e.g. 80 character per line limit, use of RECORD_TYPE and RECORD_BYTES, case, underscores, etc.).

- The Information Model shall formally define instrument so that it correctly models the radio science instrument.

  Note: The radio science instrument has components on both the ground and on a spacecraft is poorly modeled.

- The Information Model shall formally define Target.

  Note: All targets (planets, satellites, small bodies, etc.) are treated as a single category. Because of the large number of targets the use of the list is cumbersome for providers. Also, some targets which are different in class (planet, satellite, small body) share the same name. Other organizations also have standard names for bodies. Examples of targets that share names include Halley (comet) and Halley (asteroid) or Amalthea (asteroid) and Amalthea (moon of Jupiter). The Small Bodies Node has defined a set of formation rule for creating target names

eliminate ambiguities, but this method is not universally applied in PDS, hampering the use of the target name in locating related resources. There is no mechanism to describe observed regions such as fixed geographic areas or features in atmospheres. The GAZETEER object has the potential to address this issue, but the object definition refers to the " PDS Data Preparation Workbook" for the specification of the GAZETEER table record format. This document is not longer available from PDS.

- The Information Model shall define a packaging (i.e. logical volume) class for partitioning and distributing data sets.

  Note: The role of archive volumes needs to be re-evaluated in response to data sets primarily being on-line and distributed to uses electronically.

- The Information Model shall have unambiguous definitions of pointers.

- The Information Model shall specify globally unique identifiers for any class with instances that can be referenced from outside the PDS.

  Note: Individual products are not easily relocated. The interpretation of some attributes in a product description is dependent on an external file system organization or the dataset/volume context. This makes it difficult (or impossible) to form new collections (possibly based on a user's selection) consisting of portions of existing collections. The "volume" structure which requires different types of products to be stored in fixed locations (documents in "document" folder, data in the "data" folder) can result in file name conflicts when products from multiple "volumes" (or datasets) are combined into a new collection. This is an issue more for the delivery of archived products, rather than for the archive itself. An "archive system" manages and stores products and collections "as delivered" to the system.

4.1. Data Dictionary

- The data dictionary shall maintain the registration authority, submitter, and steward for each data element and its valid values.

- The data dictionary shall maintain aliases or synonyms for attributes.

  Note: A preferred alias will always be designated.

- The data dictionary shall identify and manage aliases for valid values.

Note: The standard values for some keywords have been poorly controlled. For example, there are currently several standard values of instrument_type that identify a magnetometer, which hinders using this keyword in searches.

- The data dictionary shall specify the number of values allowed for an attribute.

- The data dictionary shall allow the definition and inclusion of data elements from many domains.

  Note: The use of name spaces will allow the data dictionary to reference, adopt, or incorporate data elements from other domains. This promotes interoperability and the use of common standards.

- The data dictionary shall maintain versions of each data elements and its valid values that indicate major and minor releases.

- The data dictionary shall set reasonable constraints on data elements and their valid values.

  Note: Arbitrary limitations on the number of characters allowed for some keyword values requires an update to the data dictionary each time a new value exceeds the current size limit. Increasing and standardizing the size of certain classes of keywords such as *_id and *_name would reduce the number of data dictionary updates. Size (or length) limits are necessary in many instances. Some limits may have been imposed because implementation constraints. In some cases these constraints no longer exist. Since the reasons for some limits may be archaic, a re-assessment of current limits is warranted.

- The data dictionary shall allow the specification of dependencies between attributes.

  Note: Dependencies between attributes is first handled in the model using classes. When the model fails to address a specific type of dependency the data dictionary should be responsible to address it. - The current Information Model does not account for dependencies between keywords. For example, the presence of certain optional keywords may require the presence of other keywords (bands, band_sequence, …).

- The data dictionary shall have a comprehensive list of data types for valid values.

  Note: An standard list of data types should be considered for adoption. - Some elements use a data type and then specify limitations or extensions to the data type definition through a narrative in the standards reference. This has lead to differing interpretations of the standard. Data types and any constraints should be an integral part of the data dictionary to permit consistent application. A possibly related issue is that the Data Dictionary characterizes keywords as being CHARACTER, REAL, INTEGER, etc. But the standard values for the keyword DATA_TYPE cover a much wider range. In fact the data types REAL and INTEGER in the Data Dictionary should be ASCII_REAL and ASCII_INTEGER based on the DATA_TYPE values.

- The data dictionary shall specify the allowed character sets for data elements and their valid values.

  Note: The standards do not support characters with diacritical marks (e.g., accented and non-English characters), which are important for some international missions and archive partners. There is useful information in A.16.5 in the Standards Reference; whether this constitutes adequate support could be debated since it applies strictly only to GAZETEER_TABLE. There is a much broader question of how to handle non-Roman characters.