PDS3 Assessment

March 22, 2010

PDS4 Data Design Working Group

Agenda

- Introduction
- Assessment Findings
- Recommendations

Introduction

- A thorough review of the PDS3 standards was conducted as part of developing a recommendation for the PDS4 data architecture
  - The effort focused on attempting to formally model the PDS3 standards and determine the gaps
- An assessment team from across PDS was formed and ran from January 2008 – July 2008
- An assessment team from IPDA was also formed to get international input

PDS3 Issues - Summary

- The information model was never formally defined.
- The standards are riddled with ambiguity and inconsistencies.
  - 74+ issues were identified during the analysis of the PDS3 standards.
- There are no defined data structures.
- The class hierarchies are not consistent.
- A weak grammar is used to express the model.
- Software solutions are required to fill in the gaps.
- The data dictionary meets less than half of the modern requirements for a data dictionary.
- We tried to fix it but there were no simple fixes.
- We decided to start back at the architecture.

Findings – PDS3 Specification Work

- **The core concepts of the PDS3 standards and years of lessons-learn provide a good basis for designing PDS4.**
  - Examples of core concepts include:
    - Product as a package of "objects" that describe science data.
    - Data set as a collection of products and supporting information.
    - Data sets and products are given context through their association with Instruments, Missions, Targets, Nodes, ...
  - The standards are flexible.
  - The archived metadata is human and machine readable.
  - Science discipline experts are intricately involved in the development and evolution of the standards.
  - In general, the PDS3 data standards were innovative for their time.

Findings – PDS3 Specification Work

- **The WG has compiled 74+ anomalies, problems, issues, and items of note from the PDS3 specification work, gaps in the PDS requirements, PDS4 SCRs, and ESA/PSA input.**

- **These items have been grouped into four categories.**
  - General (Methodology, Catalog, Operational, …)
  - Data Products
    - Data Structures
  - Data Dictionary
  - Object Description Language (ODL)

- **A key issue identified is the lack of explicitly defined data structures.**
  - PDS3 "objects" *describe* data structure but do not *define* data structure.
  - For example a PDS Image object provides the number of lines and line samples in a simple image but the 2 dimensional structure must be assumed. Typically it is through the omission of the BANDS keyword.

Findings – General (Methodology, Catalog, …)

- **Problem Summary**
  - The PDS data model in general is ambiguous and is riddled with assumptions.
  - Generally accepted data engineering principles were not being applied.
- **Proposed Fix**
  - Capture and manage the data model in a modern data engineering tool.
  - Fix or replace the PDS "objects" that have problems. E.g. Volume
  - Fix miscellaneous problems.
    - Use unique identifiers;
  - Maintain the PDS Data Model independent from any implementation.
- **Source**
  - 36+14 Issues – PDS3 Spec and L3 Gaps, PDS4 SCRs
  - 6+ issues are associated with Volume alone
- **Impact**
  - Fixing any specific "object" (e.g. Volume) would affect labels, tools, and documentation and have impact on users, data providers and the system as a whole.

Findings – Data Product

- **Problem Summary**
  - The PDS3 data product is not explicitly defined.
  - The PDS3 data structures are not explicitly defined.
  - The PDS3 "objects" in general are ambiguous, have too many assumptions, and are tightly bound to ODL.
- **Proposed Fix**
  - Capture and manage the data product model in a data engineering tool.
  - Fix the problematic "objects" and other outstanding issues
  - Capture the data structures that are implicit in the archive.
- **Source**
  - 28+14 Issues – PDS3 Spec and L3 Gaps, PDS4 SCRs
  - E.g. "either eliminate or replace the implicit File object"
- **Impact**
  - Changes to almost any "object" would affect product labels, tools, and documentation and have some impact on PDS users, data providers, and the system as a whole.
  - The definition of a set of data structures for all products currently in the archive would be almost impossible.
  - A set of "factored out" data structures would make some percentage of the PDS3 archive non-compliant.

Findings – Data Structure

- **Problem Summary**
  - The PDS standards do not explicitly define a set of data structures.
- **Proposed Fix**
  - Define a set of Data Structures
  - Derive a set of Data Formats (i.e. PDS "objects")
- **Source**
  - 1.4.1 PDS will define a standard for organizing, formatting, and documenting planetary science data
- **Impact**
  - A new set of "objects" would affect product labels, tools, and documentation and have impact on PDS users, data providers, and the system as a whole.
  - A significant percentage of the PDS3 archive might be made non-compliant.

Findings – Data Dictionary

- **Problem Summary**
  - The data dictionary has limited capability and does not meet modern requirements.
- **Proposed Fix**
  - Adopt a new data dictionary model
  - Migrate the data dictionary content
  - Clean up the data dictionary content
- **Source**
  - 9+8+10 Issues – PDS3 Spec and L3 Gaps, PDS4 SCRs, PSA input
  - E.g. “The PSDD does not capture relationships between data elements such as "has similar meaning" and "has similar valid values."
- **Impact**
  - A new data dictionary model would affect the data dictionary, related tools, services, and documentation.
  - Change to the data dictionary model should have minimal impact on users and data providers.

Findings – ODL

- **Problem Summary**
  - The Object Description Language (ODL) has inconsistencies and limited capability.
- **Proposed Fix**
  - Fix or replace ODL.
  - An additional grammar makes sense for operational purposes. (e.g. XML)
- **Source**
  - 2+3 Issues – PDS3 Spec and L3 Gaps, PDS4 SCRs
  - E.g. "When are ODL "Groups" appropriate?

    "To quote or not to quote"
- **Impact**
  - Fixing ODL would affect product labels, tools, and documentation and have impact on users, data providers and the system as a whole.
  - Depending on where it is used, a new language such as XML could have significant impact.

Recommendations - Part 1

- Fix the identified problems, issues, and anomalies in the General, Data Dictionary, and ODL categories.

- They include:
  - Use a formal data modeling methodology to define the Data Model
  - Make use of class hierarchies for extensibility
  - Maintain the Data Model independent of any implementation
  - Address issues with the catalog objects and other area of the PDS standards. E.g. Data Set, Volume, Target, Repository, …

Recommendations - Part 2

- Fix the data dictionary model
- Fix or replace ODL
- Rigorously define the data product model

Backup

Ambiguity Example

- A PDS3 Image object only requires the number of lines and line samples in a simple image.
  - However the 2 dimensional structure becomes a 3 dimensional structure with the addition of the BANDS keyword.
  - A two dimensional structure is assumed by the omission of the BANDS keyword.
- Where is the first logical pixel?
- Are the pixels in row or column major order?

Image Description

```
OBJECT = IMAGE
  LINES = 800
  LINE_SAMPLES = 800
  SAMPLE_TYPE = UNSIGNED_INTEGER
  SAMPLE_BITS = 8
END_OBJECT = IMAGE
```

Simple Image