DATA MIGRATION STRATEGIES: PREPARING FOR THE MOVE TO PDS4

Lynn D. V. Neakrase[1], Lyle Huber[1], Shannon Rees[1], Matias Roybal[1], Reta Beebe[1], Daniel J. Crichton[2], J. Steven Hughes[2], Mitch K. Gordon[3], Joseph Mafi[4]
[1]Planetary Data System Atmospheres Node, Department of Astronomy, New Mexico State University, Las Cruces, NM 88003
[2]Planetary Data System Engineering Node, Jet Propulsion Laboratory, Pasadena, CA, 91109
[3]Planetary Data System Rings Node, SETI Institute, Mountain View, CA, 94043
[4]Planetary Data System Planetary Plasma Interactions Node, University of California, Los Angeles, CA 90095-1567

Introduction: The NASA Planetary Data System (PDS) is the distributed system of discipline nodes responsible for the archive of all planetary data acquired by robotic missions, manned missions, and observational campaigns through ground/space-based observation systems. Beginning late in 2012, the PDS will be publicly moving from version 3 to version 4 of its archival system. Of greatest concern moving forward is the preservation of the integrity of older data sets, while achieving improved accessibility and streamlined processes for new data entering the archive. Migration of the older data is of the utmost importance while maintaining seamless usability during the transition.

The Past (PDS3): The PDS3 system implemented a label/product duet for each item within the archive [1]. The labels were implemented in the Object Description Language (ODL) developed and maintained by JPL/Caltech (used only by the PDS), which could be used as attached or detached labels. ODL allowed for a human-readable, "KEYWORD = VALUE" structure, that was not always the easiest for software to parse or use efficiently. Directory structures were organized specifically with physical media in mind (e.g., Magnetic Tape, CDs, DVDs, etc.) [1]. Data sets were organized into *volumes* that could easily be written onto physical media, which were designed specifically for transfer from archive to user (or user to user). Retrieval of data was based on search routines maintained by the individual discipline nodes and was not well-suited for overlapping datasets present within the archive in possibly piecemeal fashion between different nodes.

The Future (PDS4): PDS4 is the latest incarnation of the PDS archiving system. With PDS4, many of the perceived problems and shortcomings of the PDS3 system have been addressed. Organization and implementation are designed around the modern idea of all data being delivered to users across internet-based systems. PDS4 is an object-oriented system based on a central core Information Model, from which everything within the system is defined explicitly [2]. This differs greatly from past incarnations and provides continuity across discipline nodes, which has not been present in the past. The catalog system has been replaced by the new central registry, which allows more information to be ingested and tracked across the system. PDS4 is product-centric. A "product" is defined as a label file and the object (data, document, etc.) it describes. The registry allows metadata to be registered across the PDS, allowing better cross-referencing between various data products and between other discipline nodes. This approach also facilitates search and retrieval at the individual product level. The new system replaces the use of ODL (managed by JPL/Caltech, used only by PDS) with the commercially available eXtensible Markup Language (XML)[2,3]. XML is widely accepted as a modern standard for encoding data for use on the Internet by providing enhanced machine readability, focused on simplicity and generality [3]. XML allows commercially available software to use the PDS archive without extensive modification, which should allow for better, more wide-spread usage of PDS data. The first public release of the PDS4 system should be expected in Q4 of 2012.

XML and the PDS: As in many web-based markup languages, XML employs a tag-based system of designating meta-attributes used to distinguish differences in the data [3]. Tags simply bracket the assigned value:

<title>This is XML for PDS4</title>
tag *value* *end-tag*

The tags can be arranged in a specific fashion into a blueprint or *schema* for the entire label. The structure of this schema is determined and generated by the Information Model ensuring that 'valid' labels are 'consistent' with respect to the framework of the Information Model. This approach is a more rigorous approach than in PDS3, in which validation was done by a tool and not against a model.

Contents Organization: PDS4 products are organized into focused "Collections" with separate collections for observational data products, calibration products, browse products, document products, etc. PDS currently identifies nine broad collection types. Related collections are then organized into logical Bundles. A Bundle, for example, might consist of the collections, which together contain all products from an instrument on a specific mission. Within the Bundle, there could be one collection for raw, calibrated, and reduced data products, or there might be three separate collections (one each for raw, calibrated, and reduced data products), a collection for calibration products, a collection for document products, and a collection for browse products.

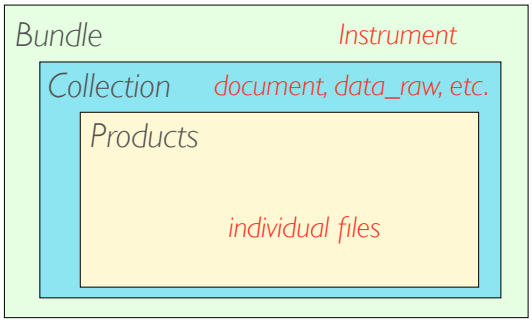

Figure 1. Bundle architecture diagram showing the nested structure for PDS4.

Migration Strategy: In order to be released for public use, the PDS4 system is being tested extensively with previously archived data that are *migrated* into the new system. Exercising many of the new features is a primary goal of migration efforts in late 2011 and early 2012. Many "test-case" holdings, of rather less complicated, complete data sets will be migrated into the new format from end-to-end. Currently migration efforts are occurring across the discipline nodes with the focus on testing all the fundamental data structures. At the Atmospheres Node for example, Python has been employed to bridge the two systems. We've constructed an XML translator that takes the validated blueprints (the schemas) and produces XML label templates and then populates the label by translating PDS3 keyword/value pairs into valid XML tags.

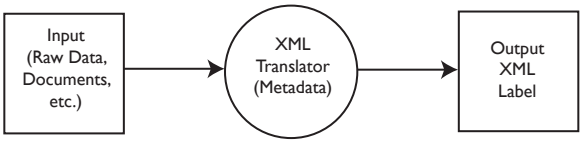

Figure 2. Simplified diagram of the migration process strategy.

This simple modular architecture for migration is also useful for designing data pipelines for new data coming into the PDS as well, in which the input data would be ingested directly from a data provider instead of a PDS3 archived product.

Our strategy has been two-fold using this modular approach as Step 1 and then using progressively more complicated data sets as Step 2. We began testing migration by working with the five atmospheric instruments of the Mars Phoenix Lander mission and have been refining the process over the last year. We chose this mission set because the data types are relatively simple, well-behaved files (ASCII Tables or Table_Character), and the mission is finished. We have moved on to prototype binary table files from recently archived (restored) Galileo UVS/EUV data, and simple FITS images from a ground-based observing campaign for Jupiter and Saturn.

Conclusions: PDS4 allows the PDS Archive to begin to transition into a modern, efficient system by using XML as the implementation of a product-centric, object-oriented, Information Model approach. We see this as being an improvement over previous versions of the archive as well as a step in the right direction for moving PDS into the future.

References: [1] PDS Standards Reference, version 3.8 (2009); [2] Data Preparers Handbook, version 4.0, *in prep.*; [3] Extensible Markup Language, http://www.w3.org/XML, (2011)