The PDS Approach to Science Data Quality Assurance

Anne Raugh,[1]

[1] *University of Maryland, College Park, Maryland, USA;* araugh@umd.edu

Abstract.

The Planetary Data System (PDS) was established by NASA in c. 1984 with the mandate not merely to preserve the bytes returned by its planetary spacecraft, but to ensure those data would be available to and usable by future generations. When PDS accepts data for archiving, it must be complete, thoroughly documented, and as far as possible autonomous within the archive (that is, everything needed to understand and use the data must be in the archive as well). In order to maintain usability, PDS must first establish usability of each incoming data submission. The two primary quality assurance tools applied to archive submissions are the PDS4 Information Model and the mandatory External Peer Review. The Information Model guides data preparers to producing well-formatted, well-documented data products that are programmatically accessible, while the External Peer Review ensures the archive submission is complete, usable, and of sufficient quality to merit permit preservation - and support - as part of the Planetary Data System archives.

1. Introduction

The Planetary Data System (PDS) was established in response to a "perception that data problems are pervasive throughout space sciences" (CITE-CODMAC), and a subsequent Planetary Data Workshop convened at Goddard Space Flight Center because it was "noted with increasing alarm by many in the science community that valuable data sets are disappearing" (CITE-PDW). NASA charged PDS with preserving and maintaining the usability of planetary mission data in perpetuity. PDS is NASA's guarantee of return on invest in planetary science. As such, when PDS accepts data for archiving the question of quality assurance is primary. In the PDS context, "data quality" is interpreted as meaning data that are complete, thoroughly documented, compliant with PDS4 data formats and metadata standards. The primary tools for ensuring data quality are the PDS4 Information Model, and the External Peer Review. The Information Model guides data preparers to producing well-formatted, well-documented data products using a standardized metadata system that can be programmatically validated, while the External Peer Review ensures the archive submission is scientifically complete, usable, and of sufficient quality to merit permit preservation and support as part of the Planetary Data System archives.

2. The PDS4 Information Model

The PDS4 Information Model codifies metadata not just for structure, but for provenance, interpretation, and analysis. The XML document structures defined for the current implementation of the model and its various constituent namespaces establish minimum requirements and present best practices for describing all these aspects of the archival data. The schematic enforcement of these requirements provides a simple, automated approach to ensuring the metadata are present and well-formed.

2.1. Content Management

The foundation layer of the PDS4 IM defines the common metadata types and structures that are used for identification, provenance, and data structure definition in the core, and for all metadata defined in discipline and mission namespaces that extend the IM (CITATION?). Metadata is categorized and organized into dependent groups that are included or not depending on context and observational data content.

The hierarchical organization into classes and subclasses allows metadata to be viewed and used as functional groups. For example, the metadata needed to cite a data product resides in a class contained in an "identification area," which may also contain classes documenting modification history. The hierarchy provides two main benefits: First, the structural organization itself imposes existential constraints for required attributes, units of measure, and so on - to ensure at least a minimal level of metadata is supplied; and second, the full structure as documented in the defining XML schemas provides a complete template to those preparing data for archiving to follow, which helps ensure consistency in what metadata are included and where. The overarching hierarchy that divides the PDS4 IM itself into namespaces extends these templates to cover entire disciplines - providing a standard model for, for example, defining common geometric values related to the observation.

2.2. Schematic Validation

Because the IM is expressed as a combination of XSD Schema, which enforce the structural and hierarchical constraints, and Schematron files, which enforce more complex dependencies related to metadata values and co-dependencies, a broad range of validation can be performed entirely mechanically. More importantly, the canonical validator (the PDS-produced Validate tool) can be deployed to any data preparer's environment and used locally to ensure compliance prior to submission for review.

3. External Peer Review

The PDS External Peer Review is required for all candidate data submissions prior to acceptance for archiving. Equivalent to the refereeing process for journal articles, The PDS External Peer Review presents the candidate data to discipline experts unaffiliated with the creators of the data. These reviewers exercise the data in its archival form by reproducing published results, doing comparative analysis between the candidate data and similar or correlated results, and so on, using only the archival resources. These reviewers then determine if the data are of archival quality and, where needed, formulate a list of corrections and additions required prior to archiving.

3.1. The Panel

External peer reviewers are chosen to be discipline experts in the data to be reviewed, but must not have been involved in the data preparation process for the data they are reviewing - neither as a member of the team that took the data or produced the archive, nor as one of the PDS consultants who advised the data preparation team. In general, two independent domain experts are asked to review the candidate data. PDS personnel do, of course, also validate standards compliance.

3.2. The Process

The typical review takes about two calendar months of time. The data are submitted to PDS, who will do a standards compliance check to ensure that tools that recognize PDS4 format will not encounter problems reading the data. The reviewers (previously chosen) then have one month to exercise the data. They are encouraged to perform some reasonable analytical process - something an end-user might do. Typical analyses include reproducing a published result, correlating the properties of the reviewed data set against another data set, or verifying calibration by calibrating a raw data set to compare to its reduced counterpart.

Reviewers also read accompanying documents, and will generally check metadata for consistency with expected standards - those familiar with the SPICE toolkit, for example, will frequently check the observational geometry included in the label against the results produced by the toolkit.

At the end of the month, a review meeting is held. The reviewers present their findings, along with revisions that should be required prior to archiving. Data preparers are invited to participate, in order to ask and answer questions. The goal is to produce a specific list of revisions, referred to as a list of "liens", to be applied prior to archiving, and to ensure that the data provider can address those revisions quickly and completely.

3.3. Revision and Archiving

Following the review, the data preparer typically takes a few weeks to perhaps a few months (where pipeline software must be revised and worked through configuration control) to make the revisions and submit the final dataset. PDS will do a final acceptance review to ensure both standards compliance and resolution of the liens list. Once accepted, the data are posted as archived.

4. Conclusion

The PDS4 Information Model, by requiring a minimum set of metadata, providing templates for data providers to follow in designing metadata for their products, and facilitating rigorous schematic validation, provides quality assurance for metadata syntax and content, as well as data structure compliance (via the metadata constraints on structural descriptions). The PDS External Peer Review ensures the data are usable by recruiting field experts to literally use the data, and by declining to accept data for archiving until its usability has been thus demonstrated. The IM and the External Peer Review work together to ensure that data are well-described and usable when they enter the archive.

5. References

This template has no bibtex file. Look for the larger template and Makefile how to do this. By default the Makefile will create an empty O7-2.bib. When you add references to this, uncomment the line `\bibliography` below, make use "make" to create your beautifully looking PDF.