DRAFT - PDS4 USER SUPPORT WHITE PAPER - DRAFT

Mark Sykes (PSI)
Mike A'Hearn (UMd)
Lisa Gaddis (USGS)
Ray Walker (UCLA)
Mark Rose (NASA Ames)

I. ISSUES AND MOTIVATIONS FOR PDS4

PDS has remained fundamentally unchanged as a system for more than a decade during which it has faced challenges with the data volume and diversity from missions and increasing demands for access to its holdings from a growing user community. Now, PDS is faced with greatly increasing data volumes from missions with continuing diversity in delivered formats, and a new level of demands from new programs such as the Planetary Data Analysis Program, which supports analysis of PDS data and the creation of new data products to be archived in PDS. Access demands are expected to increase significantly as is the need for a large and diverse group of individual researchers who will be wanting to archive years of accumulated data into the PDS to make its analysis eligible for funding. In addition, PDAP represents an opportunity for the generation of higher-level mission data products that were not produced by past missions. In sum, PDS is even more central to maximizing the science return on the tax-payer investment in current and past missions and science research programs that generated new data. Thus, it is time to revisit PDS to see how and to what extent it needs to be reinvented to ensure that it can meet these challenges. This white paper considers the issues and motivations associated with user support and begins with an articulation of those issues and motivations:

Finding data - scientists have difficulty in finding data that may be scattered at the data product level or lower across the PDS. Once they obtain the data, they may have difficulty in understanding how it is properly used and what other information/data are needed for that purpose. There is usually no 'primer'. This has implication for organization of data, capability for drilling even down to in some cases the record level, and the transparency of PDS across DNs.

Format issues/description and use issues - data is archived in a wide variety of formats, reflecting popularity at the time of submission, but making data more difficult to use by scientists not familiar with those formats. Translation services should be provided, but this is increasingly difficult as the universe of format-to-format combinations increase. This has implications for the PDS data model.

Understanding archiving requirements (missions and individual researchers) - every mission is a new challenge because PDS guidelines place essentially no constraints on archivable data formats, NASA does little to enforce its policies on adherence to delivery schedules and PDS requirements, and requirements to ensure adequate resources are

available for data product generation and archiving are few and not adequate. Researchers desiring to submit data generated in planetary research programs face a barrier of having to understand PDS standards.

Search services and support services for users - these are scattered across the nodes, making it difficult for users to find and use. Information generated by nodes for supporting user searches of its holdings is not accessible to other nodes. This has implications for supporting a single point of entry for PDS services, as well as transparency of internal PDS organization. Some examples of queries requiring cross-node interoperability:

- I want to know which direction the magnetic field was pointing (PPI) when this CDA data was gathered (SBN).
- I want to find groundbased observations of the Jupiter satellite Elara (SBN) and any serendipitous observations by spacecraft at Jupiter identified by other nodes (e.g., Rings) by asking for all data containing observations of Elara in the PDS.

Information describing/characterizing data - this is too narrow, limiting search capabilities. Some examples of queries not currently supported:

- I've found something interesting about an aurora; give me all auroral images from this time period.
- I would like a list of lobate crater observations on planetary surfaces.
- Without having to query about specific instruments, give me a listing of all magnetic field data in the PDS.

This has implications for the PDS data model and architecture.

PDS is not well-designed to support the curatorial function of the Discipline Nodes - as information about data is gained and links among data understood, there needs to be a reasonable means by which this information can be added, removed or modified (with external reviews as appropriate), and connections among data made.

PDS usage is not adequately measured and monitored in a way that reflects meaningful usage by users - Statistics based on web hits and even downloads do not measure success at meeting user needs.

II. THE PDS USER MODEL

Planetary scientists are the users that the PDS should be designed to accommodate. Planetary scientists include those experienced with solar system exploration missions and those who are mission-naïve. They include graduate students. While the PDS is a NASA-funded program, many of its planetary missions are international partnerships and therefore the User Model must include non-US planetary scientists at some level.

The PDS User Model does not include educators, K-12 students, and members of the general public or others that are not planetary scientists. Those audiences are the target of public outreach and educational activities, which the planetary scientists and other non-PDS entities are funded separately to support.

Planetary scientists cover extremely diverse areas of study that roughly correlate with the range of targets (or portion of targets) being studied. These include planetary atmospheres, planetary surfaces, comets, asteroids, interplanetary dust, planetary plasma, ring systems, planetary formation and evolution, and dynamics. Consequently, there is substantial diversity in the data acquired and studied and the scientific questions being pursued using data from PDS.

III. PDS USER NEEDS

PDS users need to be able to put data into the PDS and retrieve data from the PDS. These include data from missions and research programs. These capabilities must be reliable and straightforward. Once ingested, data must be available indefinitely.

A. Input Support

The need for users engaged in missions and individual research programs to put data into the PDS mandates the provision of input support services by PDS. These services would include documentation on required data formats and standards, online interfaces, and direct support for those services by PDS personnel. PDS personnel must also be available to support activities such as peer reviews, expert data user assistance, and PDS standards. Source acknowledgement by PDS is also required.

B. Retrieving PDS Data

Users retrieving PDS data must be able to search for and identify the data they want and have it delivered to them or made available for retrieval. Identification requires a broad spectrum of access modes and query tools that depend in part on the use of standard data formats. Delivery also requires a broad spectrum of capabilities suited to the nature and quantity of data. Direct support by PDS personnel must also be available.

Deliverables to PDS Users include data and documentation required to understand and interpret archived data. They also may include software or other data required to support that interpretation (e.g., calibration data) as well as metadata associated with query results. Access to data should be seamless regardless of location in the PDS. They should be available online through data access points that are easy to find and simple to use. Expert assistance in the use of these deliverables should also be available within the PDS.

C. Long-term Stability and Usability of PDS Data

In principal, planetary data do not have a finite shelf life as they capture characteristics that are often dynamic in time. When users submit data in the PDS, it is presumed that information will be available in perpetuity - analogous to a journal article. Similarly, because planetary data have ongoing value, users will expect that they should be able to find any data that have been ingested into the PDS, regardless of the passage of time.

The data formats required to support long-term archiving requirements of PDS is a separate and distinct issue from the data formats either submitted or used by PDS Users (with the exception of missions for submission). Popular formats are often commercial and short-lived (e.g., a particular version of an Excel spreadsheet) and not necessarily expected to be even readable easily after decades. The PDS Data Model must necessarily focus on the very long-term.

Consequently, PDS should define a minimum set of archival data formats to which data archived in the PDS will be restricted. Missions would be required to submit their data in these formats and individual scientists would submit data generated by a planetary research program either in those formats or have that data converted to those formats using tools provided by PDS. It is contemplated that a mission proposal to produce products in a format other than these archival data formats would be considered to be high risk. If a mission proposal team felt that they were producing data of a new and different nature that required expansion of the accepted archival data formats in order to insure its long-term viability, it would be incumbent upon them to contact the PDS prior to proposal submission, make their case, and if successful work with the PDS to develop and define a new archival data format.

An internal archival data format requires that PDS provides some level of reasonable translation services. This does not mean that the PDS must translate to any format a user provides or desires, rather it is part of the responsibility of the curatorial function of the Discipline Nodes to understand the limited set of formats for which translation support should be requirement and the circumstances under which it is appropriate to require the user to make their own translation. For instance, if the PDS data model required that all images binary arrays with detached labels, the nodes might recognize the need for a translation capability to FITS, a popular format among planetary astronomers. For purposes of input of data generated by a planetary research program, a similar reasonable service should also be available (and might even be transparent) to the user.

IV. SUPPORTING USER DATA DELIVERY TO PDS

A. Mission-generated data

As part of their funded operations, NASA space missions are required to generate specified data products, compliant with PDS standards, and archive them in the PDS on timescales dictated by NASA. PDS science discipline nodes support mission and/or instrument teams in these activities and support them through the procedures required for PDS ingestion.

Missions need to allocate and preserve adequate resources for the generation and archiving of their data products. Towards this end it is essential that it be required that a budget for data product generation and archiving be required as a specific line item in the proposal and subsequent mission budgets.

Mission-delivered data must conform to the PDS data model (i.e., can't use a non-standard data formats just because it is designed to accommodate internal analysis software of the instrument team).

Currently for mission proposers and missions, the PDS provides online access to the Proposer's Archiving Guide and the Archive Preparation Guide, laying out requirements and procedures for the generation of data products and their ingestion. These documents support the more complex and formal PDS Standards Reference and the Planetary Science Data Dictionary. However, experience demonstrates that making documents available alone is not adequate. It does not guarantee that they are read. Some missions have attempted to design their product labels on the basis of a cold reading of the PDS Standards Document, with the result of products being generated that were neither compliant with PDS standards nor in a usable form. This can be obviated by NASA (not PDS) imposing clear data definition and archive planning requirements for missions once selected (for Phase A? Phase B?). Compliance needs to be monitored and enforced by the mission Program Executive at NASA HQ. These requirements should include:

o Regular contact with the lead PDS Discipline Node assigned to the mission.
o Generation and signoff of PDMPs, SISs, etc. by a time specific, and adherence to delivery and review schedules therein?
o Compliance with the PDS Data Model.
o Adherence to peer review lien resolution requirements. For instance, a review panel may find that data products are in compliance with PDS standards, but that additional information may be required to make them usable. If liens are large, the changes made to resolve those liens are very likely to result in new liens and require another iterative review.
o The review of the pipeline can ensure compliance with PDS standards, but cannot ensure that the data will satisfy peer review for scientific usefulness. The missions must understand this.

There needs to be a more consistent cross-mission procedure for the generation and delivery of data products to the PDS. Part of a solution may be to require that appropriate missions develop a PDS peer-reviewed, configuration controlled pipeline to generate standards-compliant EDRs and RDRs. Such a pipeline would be required to be subject to regression testing to help determine when changes to the pipeline require another PDS

review. This would mean that missions would have to deliver sample data products very early for review by PDS to allow for timely pipeline modifications. Any changes in the pipeline would probably require at least internal PDS review to ensure standards are maintained. The need for scientific review would have to be separately assessed (and may result in required modifications of the pipeline). It would be the responsibility of the missions to provide information to PDS on a schedule that would allow it to remain in compliance with NASA policy on data delivery deadlines.

The above NASA requirements should result in the straightforward ingestion of compliant mission data by PDS. Additional reviews would be required for non-pipeline higher-level or derived products (e.g., maps) generated by the mission.
It is not the responsibility of PDS to generate mission data products of any level.

It is not the responsibility of the PDS to compel mission compliance with NASA policies regarding mission archive requirements and schedules. PDS may report on compliance, but it is the responsibility of NASA HQ, informed by PDS reports, to compel such compliance. The failure of NASA HQ to do this has resulted in PDS attempting fill the gap (with absolutely no leverage) with the result that the relationship between PDS and missions is often viewed as adversarial, undermining the ability of PDS to do its job and undermining the quality of products generated. For NASA HQ to do its job, it is essential that it track data products initially proposed, how those are modified with mission development (and whether or not such modification translates to loss of mission science), delivery schedules for documents and products and then compliance with its policies. If a mission refuses to generate a promised high-level product, for instance, it is up to NASA HQ to take action. It is not the responsibility of the PDS to use its limited resources to generate that product itself.

In addition, it is sometimes the case that the delivery of scientifically useful data lags considerably the development of the data pipeline. It is the responsibility of the PDS to review data for scientific validity and usability, but it is the responsibility of NASA HQ to compel resolution of liens against the data and pipeline found during the review process.

B. Research program-generated data

In ROSES 2008, the Planetary Data Analysis Program will be announced. PDAP will provide funding for the analysis of data archived in the PDS, restoration of NASA data not yet in the PDS, as well as the archiving into the PDS of data obtained under other NASA funded programs (e.g., including higher level mission products, ground based telescopic observations, laboratory data). PDAP will increase the demand by individual users for support to have their research program data ingested.

There is potential for PDS personnel to be overwhelmed by large numbers of users wanting to archive small amounts of data (relative to mission-sized) deliveries. This necessitates the development and maintenance of tools, such as OLAF, to minimize that impact by maximizing automation.

Submitted data (including documentation) must be compliant with the PDS data model. In the event that there is a divergence between standard user formats and the PDS data model (say the model for images was a binary array with detached label and the user standard was a FITS image), then it is in the interest of NASA to support the generation of tools (which can be hosted by PDS) to convert one format into another ready for archiving. This is not a service that will be provided to missions, since their pipeline needs to be designed to generate PDS compliant data.

It is not the responsibility of the PDS to follow up on promised data deliveries by researchers. That is the responsibility of NASA HQ. It *is* the responsibility of the PDS to provide technical assistance to those researchers.

V. SUPPORTING USER DATA RETRIEVAL FROM PDS

A. Data Access Model

Data access has two components: identification and retrieval. The ultimate results of a data search, prior to retrieval, should be a listing of the data products or datasets desired by the user with access to other data and information needed to interpret, understand and make use of those data products or datasets.

1. Identifying/Finding Data in the PDS

Because of the diversity of the users supported by PDS, the methods of finding data within the PDS must necessarily be diverse. These would include (but not limited to) search by reference, generalized parameter searches and specialized search interfaces addressing the needs of specific user disciplines. The main page of the PDS is a logical single point of entry to access all search interfaces supported by the different nodes in a transparent fashion.

a. Citation reference

A user may find PDS data cited in a scientific paper. In such citations there will be a unique identifier (currently dataset_id or dataset_id:product_id) given in the reference. A user should be able to go to the PDS and be given a direct means of inputting that unique identifier and accessing the data referred to in the scientific paper.

b. General parameter search

There are a variety of high-level parameters (keywords) common to all datasets. Some means of drilling down through these parameters should be made available. This is analogous to the current PDS-D interface currently available on the PDS home page and

represents a basic capability.

c. Specialized search

Within a given planetary discipline, there may be need to search on parameters not common to all datasets as well as to search at a product and even record level. Tools designed to accommodate these needs require the knowledge specific to Discipline Node scientists and must therefore be developed and implemented at those nodes. These capabilities would tend to be target or target-class specific, such as “Mars”, “asteroids”, and “rings.” Examples would include a geographic or map-based search for image data of the Moon, or a feature-based query covering imagery of all planetary surfaces.

Some thought also needs to be given to queries that cross discipline boundaries and whether specialized interfaces provided by one node may allow for access or query through the specialized interface of another node.

Discipline Nodes may also wish to provide access to simple specialized interfaces developed elsewhere such as Google Moon and NASA WorldWind Moon (Clementine) upon their positive evaluation.

d. Relative importance of search techniques

Citation reference search represents a very specific capability where the exact product or dataset is identified in advance. The general parameter search is a very superficial capability. Citation reference search and general parameter search together are not sufficient to satisfy the needs of the PDS user. For PDS to be relevant to its user community, specialized search capabilities need to be developed and supported by the Discipline Nodes. Such capabilities are dynamic with time, requiring ongoing development and maintenance. The Discipline Nodes have, to various extents, developed specialized search interfaces already, but their capabilities have been limited in scope due to PDS resource limitations. The recognition that such specialized search capabilities are an essential part of PDS support for the scientific community would represent a sea change in philosophy and necessitate a substantial investment of resources.

2. Retrieving Data from PDS

Having identified the data desired, the user will want to acquire that data. These data may be a few bits to terabytes. They may have associated files that are critical to their interpretation that the user needs to know about and be able to also acquire. Consequently, a variety of modes is needed for data transfer, depending upon size of package:

o direct download (e.g., via html)
o transfer and electronic download (e.g., sftp, wget)
o transport by physical storage device (e.g., DVD, storage drive)

3. User Support

For both data identification and download, direct and expert user support should be available - either by email, electronic chat, or voice-to-voice.

4. Data Formats

Assuming that the PDS data model adheres to the principal of simplicity in identifying the minimum number of necessary fully-documented standard formats that allows for long-term viability of the range of PDS holdings, conversion to more sophisticated formats by either the user or the PDS should be straightforward. Popular formats will always be changing, driven by changing analysis packages.

The first priority is the delivery of data in the formats delivered in adherence to the PDS data model. Conversion to other formats (e.g., a raster scan with a detached label to a FITS or JPEG2000 image) is a useful service.

5. PDS Services

PDS is resource limited. User access can be enhanced through PDS services such as the creation of higher-level products (excluding those failed to be generated by missions) including graphical displays, coordinate transformation, remapping, and others. What is necessary, however, should be determined by Discipline Nodes on behalf of the communities they serve.

6. Application Program Interface (API) Support

It is commonly true that accommodating programmatic access to search interfaces spurs the development of new applications by 3rd parties. It is in the interest of the PDS to support the scientific community in the development of such external applications, to the extent that it is reasonable, given limited PDS resources.

VI. IMPLICATIONS FOR THE PDS WEB INTERFACE AND TRANSPARENCY OF NODES

Users should not need to know to go to a specific Discipline Node's website to access data. Knowledge of the internal structure of the PDS should not be required in order to make use of the PDS. All nodes should be transparent to the user.

The PDS main page should be a single point of entry to the services provided by the nodes in support of data searching and archiving. However, this does not mean that it should be the only point.

Some recommendations:

Services maintained by the Discipline Nodes should have addresses such as pds.nasa.gov/sbn/olaf. (Or perhaps sbn.pds.nasa.gov) Current PDS sites such as pds.jpl.nasa.gov and pds-smallbodies.astro.umd.edu should redirect permanently to the new, unified URLs.

Facilities provided by data nodes (sometimes set up as mission functions) should also appear in the same unified URL scheme. (E.g., themis.asu.edu would be better as geo.pds.nasa.gov/themis, or something.)