Planetary Data Ecosystem Independent Review Board Interim Briefing

PDS Focus

Talk Outline

- IRB Structure & Membership
- Background information
- IRB Core Values
- Categories of PDS Findings
- PDS Findings and Recommendations
- Questions and discussion

The IRB Team

NASA Review Manager
Becky McCauley Rench

Contract Manager: Autumn Manecke
NRESS Support: Susan Borden
Latessa Tuck
Jason Silveira

Chair
Melissa McGrath

Subcommittees

Archiving
Co-Chair Moses Milazzo

Executive Secretary:
Sarah Black

NASA Ex Officio Members
Meagan Thompson
David Smith

Searching
Co-Chair Emily Lakdawalla

Executive Secretary:
Katya Gilbo

NASA Ex Officio Member
Lindsay Hays

Utilization
Co-Chair Sebastien Besse

Executive Secretary:
Don Hood

NASA Ex Officio Member
Sarah Noble

Mining & Automation
Co-Chair Bruce Wilson

Executive Secretary:
Hannah Dattilo

NASA Ex Officio Member
Megan Ansdell
~~KC Hansen~~

Inter-Relational
Co-Chair Caroline Coward

Executive Secretary:
Blake Mendoza

NASA Ex Officio Members
Steven Crawford
Lucas Paganini

Members

Kate Crombie
Martha Maiden
Joe Masiero
Flora Paganelli
Corrine Rojas
Priyanka Sharma
Marshall Tabetah

Julie Castillo-Rogez
Vandana Desai
Mary Ann Esfandiari
Laszlo Kestay
Matthew Miller
Lynn Neakrase

Alyssa Bailey
Ross Beyer
Henry Hsieh
David Mayer
Chase Million
Danielle Wyrick

Abigail Azari
Benjamin Burnett
Amitabha Ghosh
Jason Laura
Jennifer Shin
O. James Tucker
Alexandria Ware

Reta Beebe
Shawn Brooks
Stephane Erard
Robin Fergason
Patricia Lawton
Xiaogang Ma

Background

- Chartered October 2020
- 4 Full IRB meetings (public) in Nov, Dec, Jan, Feb; 53 subcommittee meetings
- Full IRB meeting recordings and minutes available at science.nasa.gov/researchers/science-data
- Our final report will be delivered to NASA on 26 March 2021
- Our final report is *non-consensus* and will contain prioritized recommendations. Neither of these has been implemented for this interim briefing.

"The data gathered from missions is our National Treasure. It should be treated as such."

- IRB member Kate Crombie

The data gathered by the planetary science community is humanity's treasure. With you, we strive to preserve and ensure its usability for posterity.

IRB Core Values

- **First, do no harm**: Avoid the law of unintended consequences.
- **FAIR:** Facilitate participation in the PDE by adhering to the FAIR data principles of Findability, Accessibility, Interoperability, and Reusability.
- **Open**: Advocate for open science practices, including open access, open data, open code, open software/tools, and others.
- **Collaborative:** Encourage international collaboration. Welcome new participants from both inside and outside the professional space exploration community.
- **Effective:** Provide timely, useful support to user communities, especially data producers.

Categories of PDS F&Rs

- General
- PDS Structure and Governance
- Scope of PDS
- Standards
- Usability

Neither the categories nor the material within the categories is prioritized.

F1: The IRB applauds NASA's concept of a Planetary Data Ecosystem. The foundation of the Ecosystem is the planetary sciences community: the people who produce, provide, and use data (missions, R&A scientists, lab technicians, sample collectors, engineers, citizen scientists, etc.).

F2: The PDS is a cornerstone of the Planetary Data Ecosystem and is critical to its success.

F3: The PDS successfully preserves data, documentation, and expertise from NASA's planetary missions.

F4: The PDS is a trusted, reliable source of planetary data, enabling data discovery, search, retrieval and analysis.

F5: The federated structure of the PDS allows it to customize along science themes and capture science expertise.

F6: The PDS has been evolving to meet the archiving and preservation needs of planetary scientists and researchers, leading not just domestically, but internationally.

F7: The PDS is one element of the archival subset of the PDE.

F8: Within limits of available funding, the PDS does a good job of meeting its primary mission of data preservation.

R1: Future evolution of the PDS should preserve these positive attributes.

F9: The PDS is structured as a set of federated discipline-specific nodes. This sometimes leads to duplication of effort, duplication of data, inconsistency in tools, fragmented access to data, and potential confusion for community members. Some examples include Data Management Plans unique to nodes; node-specific archiving requirements; search mechanisms unique to each node; lack of systemwide cybersecurity standards and infrastructure; and ontology used to align various node metadata schema.

R2: The PDS would benefit from a greater emphasis on systemwide, perhaps centralized governance in regard to structure, standards, and related processes, in order to make decision making and communication more effective.

R3: The makeup and distribution of nodes should be examined more closely to ensure that the PDS contains the appropriate and relevant node elements and subject matter expertise, that unreasonable duplication of effort and data do not occur, and that appropriate flexibility regarding scope and content is built into policy.

R4: The PDS should continue its emphasis on data by looking to hire additional expertise in information sciences.

F10: Attributes of a more systems approach include robust, standard metadata across PDS with the capability to filter by node, and standard search tools which provide cross-node functionality for all users.

F11: PDS4 is intended to encompass many of these attributes. However, this work is likely years from reaching the essential functionality envisioned by the PDS.

R5: NASA should prioritize this work and find ways to accelerate this timeline.

F12: The PDS receives direction from NASA that is sometimes outside the scope of the original agreements under which the PDS Nodes were funded ("unfunded or underfunded mandates," e.g., PDS3 to PDS4, PDART and other R&A archival requirements).

F13: Funding levels for archiving data (mission, R&A, samples, etc.) are not always appropriate.

> **R6:** NASA should fund PDS nodes at a level appropriate to the full scope of work defined by the selected proposals as well as accumulated duties. For the money being spent on acquiring planetary data, are we allocating enough to make that data usable?
>
> **R7:** Along with the appropriate expectations for their data and metadata, NASA should fund other, non-PDS repository elements at appropriate levels.

F14: There is a mismatch between the explicit mission of the PDS (preserve data) and the perceived mission for the PDS by the broader user base (distribute usable data). Data archiving is a non-trivial activity that requires careful consideration. Data discovery and delivery to expert and non-expert users and programs/automated tools are equally challenging and potentially at odds with the goal of data preservation.

R8: The prioritized goals and scope of PDS need to be carefully and explicitly defined by NASA, and clearly articulated to the community. Mandates above and beyond the agreed upon scope must be accompanied by commensurate funding.

R9: Consideration should be given to how to disentangle the data preservation mission from the distribution of usable data mission.

R10: Each PDS node and the PDS as a whole should allocate additional time to identifying and using appropriate data use and citation metrics to drive improvements in the overall data FAIRness of the PDS.

R11: The PDS should regularly assess the FAIRness of each PDS node and the PDS as a whole, both for interactive access to data and for automated/programmatic access to data, using established criteria such as those at go-fair.org. This assessment should be used to identify the areas of highest value for improvement and to highlight the improvements which have been made over time.

F15: There is much important planetary science data which exists outside the PDS. Some examples include:

- Mission spacecraft engineering data.
- Many data from planetary-relevant samples (terrestrial and non-terrestrial) are not yet archived according to FAIR Principles (some are, some are not).
- Data from laboratory experiments are not typically archived according to FAIR Guiding Principles.
- There is currently no place to archive simulation or modeling input or output data, nor results.
- There is currently no place to archive software (not just the source code, but also including the development environment required to compile software, and etc.).
- There is currently no place to archive mission processing pipelines (the infrastructure is larger than just software and often includes databases and configuration files that make these unique).
- There is currently no place to archive historical data, press releases, internal communications, etc., that might be useful for archeological studies of NASA missions past (and present-that-will-become-past).

R12: The PDE IRB strongly recommends that these critical planetary science data be preserved. [The above list is not comprehensive.]

R13: NASA should consider ways of archiving that are amenable to creating FAIR and standards-based archives or repositories of these growing "data" sets.

- A primary archive for mission spacecraft engineering data is needed which meets FAIR Guiding Principles. This could be the PDS if that is appropriate.
- A primary archive for planetary-relevant sample data is needed which meets FAIR Guiding Principles. See the AstroMat archive as a non-PDS example of a PDE archival entity. This archive should be future-proof such that samples collected from Europa in 30 years can be archived.
- There needs to be a primary archive for planetary-relevant laboratory data which meets FAIR Guiding Principles. This could be the PDS if that is appropriate.

Two examples have been called out for more urgent attention:

F16: Arecibo Observatory (AO) data are not, generally, currently archived in an Ecosystem element, although some data are archived in the Small Bodies Node (SBN). This is also true of other ground-based radar telescopes. Some valuable data have already been lost.

F17: AO data processing tools and software are not archived, and higher-level processing is science-need-dependent; it is not, at this time, realistically possible for a new scientist interested in accessing AO data to process them to appropriate levels.

> **R14:** NASA should meet with experts to consider whether AO (and other radar telescope) data and processing procedures (or software) can appropriately be archived in the PDS. The total data volume is not particularly high, but funding is likely needed to accomplish compliance with data, metadata, and documentation standards will need to be available at the appropriate level.
>
> **R15:** Because of the uniqueness of the AO data and situation, this should be accomplished sooner rather than later.

F18: There is a clear need for analysis-ready data that are in well-known formats, available through automatable methods, are labeled, and are well-documented.

F19: High-level data products created by scientists and other researchers and data providers, including citizen scientists, are not easily findable in the PDS.

F20: There is not a NASA-approved home, within the PDS or elsewhere, for the derived data products produced by Machine learning, artificial intelligence, and advanced analytics (ML/AI/AA) projects working on planetary data. This can lead repeatedly to duplication of effort.

[NON-CONSENSUS]

R16: NASA should reconsider whether individual R&A-funded researchers and others should be required to archive with the PDS or an equivalent archive. ESA's Planetary Science Archive Guest Storage Facility is an example alternative model. Such alternatives should require data providers to meet standards, but those standards may not need to be as strict as the PDS standards.

F21: As a community we are often faced with the need to do more with, at best, a flat level of resources.

R17: The PDS and the entire Ecosystem should take advantage of the SBIR/STTR program.

R18: PDS should encourage and enable crowdsourcing to help with some of the "boring tasks" of data conversion, etc. DARPA has looked into gamifying crowdsourcing tasks like this to incentivize citizen involvement and "good" work.

R19: A carefully crafted strategy is needed to identify and prioritize the preservation needs of the planetary science community that are not currently being addressed.

F22: PDS is a trusted and reliable archive for NASA planetary data, and it is actively engaged with the international planetary data community. However, neither the PDS as a whole nor any of its nodes have completed any form of certification. Certification is one method of addressing journal and community expectations regarding data archives, and the certification process is useful in both identifying opportunities for improvement and improving transparency with the broader community.

> **R20**: The PDS nodes, with the potential exception of the Engineering node, should pursue CoreTrustSeal certification. CoreTrustSeal is the primary internationally recognized standard for operation of long-term data archives. This may require funding to meet the specifications and for joining the World Data System.
>
> **R21**: The PDS and PDE should engage in community with other kinds of domain repositories, such as the Council of Data Facilities, Research Data Alliance, CODATA, and the World Data System.

F23: The comparison of PDS with EOSDIS in the 2017 PDS roadmap study was incomplete, missing elements of EOSDIS such as work with field campaign data, airborne data, modeling output, and human dimensions data, which are far more analogous to problems tackled by PDS than the large satellite missions which comprise the majority of EOSDIS volume.

> **R22:** NASA should increase efforts that share principles, best practices, standards, and tools across the many NASA data systems.

F24: The PDS is working diligently to develop, manage, and maintain a complex set of data and metadata standards to meet a wide variety of planetary science data needs and expectations.

F25: The PDS Peer review of data and metadata for archiving is inconsistent. This includes peer review of the metadata standards.

> **R23:** NASA should help ensure that PDS has adequate expertise and funding to maintain current standards and to support ongoing improvements, including funding peer-review of data submissions.

F26: Metadata are one of the most important aspects of an archive, repository, or library.

> **R24:** FAIR-compliant metadata standards should exist and be usable across all Ecosystem/PDS archival elements so that all archived data are equally accessible. The PDS can help lead the way, but there are some data that may not be appropriate for the PDS.

F27: Long-term archiving of data for future infrastructure use (targeting of imagers and controlled products for landers, for example) requires standards for controlled, "foundational" data products and associated metadata.

> **R25:** NASA should ensure a long-term commitment to open planetary spatial data infrastructures (PSDI) that would enable access to such controlled, foundational data products. An essential part of any PSDI is standards-compliant, analysis-ready data, and this could be met through expansion (with appropriate budget allocation) of the PDS, or could be met by another PDE archival element.

F28: PDS now provides a Digital Object Identifier (DOI) for all new archived data at the bundle-level, which is a step in the right direction for improving search of and access to planetary data. Work is in progress, but incomplete, for assignment of DOIs for all existing bundles. This work is hampered by essential DOI metadata, as well as desired DOI metadata, that is not machine-accessible for PDS3 data.

> **R26:** NASA should accelerate work, to the degree possible, to ensure all data sets within the PDS have an associated DOI and that other elements have appropriate, persistent, and resolvable identifiers. NASA should consider the degree to which desired DOI metadata is manually assembled for historic data before at least getting DOIs registered with essential metadata.

F29: The PDS develops discipline-specific metadata dictionaries, at the discretion of the PDS Nodes, with input from the data providers. It is not obvious that there is an overall architecture for these dictionaries, nor is there an established peer-review process to ensure interoperability or non-duplication of these dictionaries.

F30: Accurate data dictionaries will be foundational to future AI/ML cross-science integration of multiple datasets

R27: PDS should adopt a systems approach in this area with a unified, controlled vocabulary.

R28: All data dictionaries and information models for the PDS and for other archival elements need peer-review before implementation.

F31: With the expansion of exoplanet data acquisition and sample collection on other planetary bodies, astrophysics, helio-based, earth-based, and laboratory-based data and planetary data are becoming more and more relevant to each other.

R29: NASA should continue to support non-planetary data archives and to encourage the cross-communication between planetary and non-planetary metadata developers.

R30: Working across these communities to employ existing metadata standards, or to collaborate on developing interoperable standards, is preferable to the PDS inventing new ones.

F32: Some users who have little to no experience archiving with the PDS struggle with its mission-focused archival demands. They also struggle with understanding the true costs of archiving. The PDS has substantial documentation, such as the PDS4 data providers handbook, yet there remain significant opportunities to assist data providers, particularly those from Research and Analysis projects to correctly understand data management needs and incorporate appropriate data management practices early in the project.

R31: The PDS should create a DMP template (e.g. on dmptool.org or something similar) for all of the various data types they accept. These templates should be linked whenever NASA mentions DMPs.

R32: Regular, broadly accessible, and effective training programs should be provided for scientists, mission specialists, and others who need to archive with the PDS. Entities with experience delivering to the PDS should be involved.

R33: Training should include peer-reviewer training.

F33: The PDS user interface could be more user-friendly.

- Metadata could be improved to help users find the data they need.
- Documentation exists but can be opaque to non-experts.
- New users frequently have difficulty learning how to search and use the PDS, creating a barrier for entry.
- Visualization could be enhanced so users can efficiently locate the subset(s) of data they seek, and don't waste time downloading data that doesn't meet their needs.

F34: Calibration can be a barrier to access.

- The usability of documentation for radiometric and geometric calibration methods is highly variable.
- It is sometimes difficult to determine what the best-quality version of a data set is or which version is most suitable for a particular purpose.
- Older data sets are often effectively inaccessible to most would-be users.
- The NASA PDART Program is insufficiently addressing these issues, meaning they are likely to persist for many years.

R34: To the extent practical, PDS archives should include not only raw data but also calibrated or otherwise analysis-ready processed versions of the same data sets.

F35: Data linking within the PDS is not yet mature:

- Few links exist between data sets and products derived from them in either direction.
- PDS4 standard includes powerful mechanisms for providing pointers between data sets and products, but application of these mechanisms has only started recently and is very incomplete.
- Derived data published outside the PDS do not take advantage of these mechanisms to link to data sets within the PDS.
- Additional link types are also needed

F36: Despite the significant efforts of PDS nodes, substantial barriers to search and discovery of PDS data remain. While migration to PDS4 will address some of these needs, that work will take years to complete. The differences among PDS nodes and the lack of a comprehensive search capability inhibit discovery. Further, users (particularly those who are not planetary data specialists) may come from a variety of non-PDS sites in searching for data.

- Users who are well-connected to NASA science missions and have professional or social access to NASA mission personnel have advantages in being able to locate the desired data due to community sharing of information and informal search techniques.
- Opaque language and lack of guided pathways to expertise impede entry to outsider or inexperienced users.
- There are multiple search systems at present, in different stages of evolution, and with a lack of clarity in which system is most appropriate for which purpose. While multiple search systems is not inherently problematic, particularly when systems are geared towards different types of users, it is nonetheless essential that those systems are comprehensive and that users can have confidence in high levels of both relevance and recall regardless of what search system is used.

F37: It is not practical to expect the PDS to push metadata to non-PDS systems to enhance discovery. However, there are broadly used standards in place that enable harvesting, and which can specifically enhance discoverability in more generalized search systems.

R35: NASA should consider the use of Linked Open Data (JSON-LD) and Schema.org metadata in dataset landing pages as a method to both improve discoverability (including through general Internet search tools) and improve machine actionability of PDS data. Completing at least the substantial majority of the DOI registration and the PDS4 migration is a higher priority, but that work should be done with the goal of broader discoverability in mind.

R36: NASA should have people whose expertise is in communicating with the public, or people who are less comfortable with the processes review tutorial materials, to help make them more easily understandable to the non-expert.

F38: The PDS cloud strategy is reasonable, within the limits of funding. It recognizes the value and roles of different computing modes (e.g. public cloud computing, hybrid cloud, private cloud/high end computing, and user-provided computing).

F39: Migration to public cloud computing, in and of itself, will not address any of the particular needs identified by the IRB. The PDS cloud strategy recognizes that cloud computing, including public cloud computing, is a tool to be used in addressing user and system needs, rather than an end in and of itself.

> **R37:** The PDS should continue its work to refactor systems into more cloud-native architectures, including containerization and well-defined application program interfaces, as means to reduce technical debt and create new capabilities, recognizing that this work will also enable migration to an appropriate combination of public, private, and hybrid cloud computing. Given other IRB recommendations, the PDS should not accelerate cloud migration plans, except where the PDS identifies such a migration as the preferred approach to addressing IRB recommendations or needs identified through other channels.

F40: Discovery of publicly available software relevant to planetary science can be difficult due to the diverse locations where such software may be found.

> **R38:** NASA should consider how to make publicly available software relevant to planetary science more amenable to discovery. While it is not appropriate for the IRB to suggest a particular implementation strategy, possible options discussed include to establish a centrally managed catalog of publicly available software (i.e., a registry), that can simply include links to available software or NASA could consider establishing a dedicated software node within PDS that would be responsible for curating, including interfacing and ensuring the successful implementation of NASA open source software policies.

F41: Machine learning, artificial intelligence, and advanced analytics (ML/AI/AA) have critical roles in mission planning, mission execution, and maximizing scientific value from mission data. However, data FAIRness, particularly for automated processes essential to ML/AI/AA, falls short of what PDS staff, PDS users, and would-be PDS users see as necessary to realize the full value of this expensive data. Here we emphasize factors more related to Interoperability and Reusability. While the PDS4 data providers handbook is an excellent and comprehensive technical reference for data providers assembling a PDS4 archive of their data, there is a separate, but related, need for guidance and documentation on data management practices geared towards the ultimate reusability of the data.

R39: We reiterate recommendation R11 from above: PDS should regularly assess the FAIRness of PDS data, including considering the perspectives of both interactive and automated access to data and metadata.

R40: Recognizing existing budget constraints and the proposition that many aspects of data reusability cut across the broad spectrum of scientific data management, the PDS should work with other planetary data ecosystem elements, other NASA data management projects, and other scientific data management groups to develop guidance, documentation, and training for Planetary data providers and data managers on approaches, specific practices, and tools that enhance the reusability of data, particularly reusability in the context of producing analysis-ready data products and data access by automated processes essential to ML/AI/AA.

F42: The PDS is considering implementing a user registration system, recognizing that it may be useful in better understanding the user community, increasing participation in the ACSI customer survey, and understanding use patterns across PDS nodes. Tracking users by Internet Protocol (IP) address provides no contact mechanism and is becoming increasingly inaccurate given the increasing use of institutional egress proxies that put large numbers of users behind a small number of Network Address Translation (NAT) IP addresses, highly mobile users, and users working across multiple computers. The PDS also recognizes that some form of user registration may become essential, particularly in the context of public cloud computing, in order to manage the egress costs associated with any given user.

> **R41:** The PDS should actively solicit input from, and factor in the experience of, other scientific data management systems that have implemented user registration systems to better understand the benefits and user impact of such a system, as well as best practices observed in looking across those other systems. The 20-year experience with user registration within EOSDIS, specifically, should be considered in evaluating the cost/benefit for a user registration system, as well as the appropriate level of interoperability of a PDS (or planetary data ecosystem) user registration system with other identity systems used in science applications.

[NON-CONSENSUS]

F43: The PDS software working group efforts, and efforts from the nodes in associating "tools" to the PDS4 machinery is greatly appreciated. Those tools help a lot in the preparation/validation of data, and with the usability of data in the end (i.e. Python PDS4 Tools, validate, transform).

> **R42:** There needs to be some sort of effort to maintain this to ensure the users are not left with PDS4 data with no support in reading them (this is a big issue in PDS3). This would be consistent with an open software policy.

Questions and Discussion