Increasing Discoverability Across Planetary Science Data Archives

WHITE PAPER

September 2019, Version 0.4

Vision: *Provide an integrated world-wide data services platform to enable discovery, use, and analysis of data within and across planetary science archives.*

Executive Summary

The Planetary Data System (PDS) has amassed 1.7 petabytes of data in highly curated data archives providing a standard model for stewardship of planetary data that has been adopted world-wide. As PDS looks forward, there is an immense opportunity to launch a project focused on developing the data services and standards to increase access, discoverability, and usability of planetary data archives across the PDS. In doing this, the PDS will lay the foundation for adopting even more capability to support new computing paradigms around data-driven discovery and the construction of an international planetary data ecosystem, enabling access, use, and analysis of data across not only NASA, but with international agency archives.

This white paper lays out the background, the current state of the PDS including an analysis of current data services as appendices, recommendations for increasing discoverability from planetary archives, support for increased access and computation for data analysis, and the shift towards a planetary data ecosystem. The white paper identifies key actions including (1) the development and adoption of a standard search API that should be supported by all PDS archives to enable product level search; (2) the implementation of search engine technology and capabilities pervasively across the PDS for improving discoverability at all levels of the archive from collections to base products; (3) the linking of search engines across archives to allow users to reference and find data; and (4) comprehensive and consistent mission archive pages to ensure users can bootstrap themselves into mission archives with sufficient instrument data user guides. The white paper also discusses leveraging modern technologies including cloud services, machine learning for dynamic labeling for search, integration of analysis toolkits, and support for interactive visualization. In addition to building the data discovery infrastructure, it is important to emphasize that PDS will also need to upgrade its web design to improve the overall user experience and ensure it can fully leverage the data services across the PDS as they are matured.

The white paper also describes related work in other science disciplines that are embarking on similar projects to address scale, use of technologies to support big data and data science, and to support a growing diversity of user demands and service requirements. The PDS is further challenged by its burgeoning data volume and increasing variety of data it must support in a constrained budget environment.

Finally, the white paper identifies a phased approach to launching and managing such a project. It is important that it be managed as a PDS-wide project with appropriate resources and schedules that include all nodes and stakeholders across the PDS.

1. Background

The Planetary Data System (PDS) has evolved from the early days of capturing archives on physical media [1] to the present as an international federated infrastructure with data distributed from different disciplines [2] and agency partners through online mechanisms coordinated through the International Planetary Data Alliance (IPDA) [3]. Figure 1 shows the federated structure of the PDS that includes discipline science nodes and support nodes from cross-node services and software support. As the PDS has evolved, it has made explicit architectural decisions to ensure that it could address key drivers including scalability and extensibility, most recently through its major upgrade to PDS4 [4]. This has resulted in PDS growing from approximately 10 TBs of data in 2001 to 1700 TBs of data in 2019. It has also allowed PDS to support the capture of a wide variety of data from different data providers representing approximately 600 different types of instruments with widely varying data structures, complexity and associated metadata. The shift has enabled the PDS to enhance its support of one of its primary requirements: *the preservation and stewardship of data*.

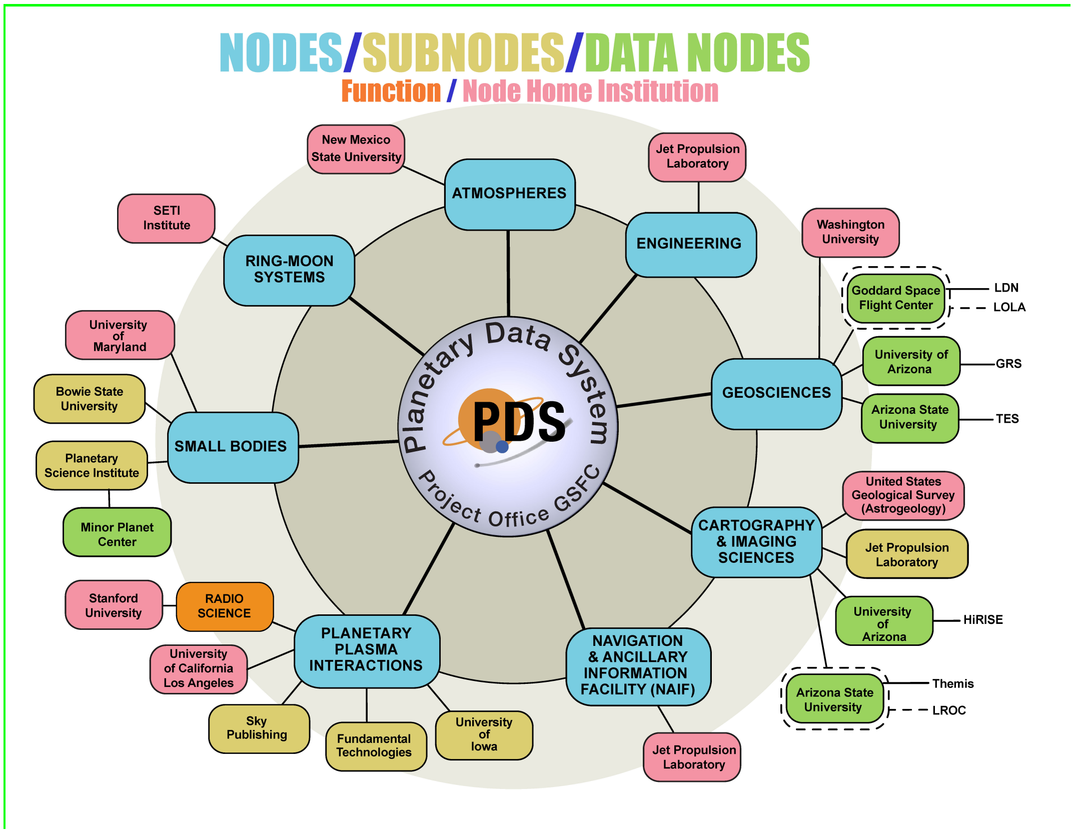

Figure 1: Federated Structure of the PDS

As the world of data intensive computing has evolved, so have the architectures, tools, and users. The architectures are shifting to not only scale towards the stewardship of "big data" but also to enable new capabilities in online services that allow users to discover, use, and analyze data. Search engines have now replaced traditional databases and common APIs are in place that allow users efficient access to find, download, and compute on data. In many cases, this technology leap is shifting the traditional science data processing paradigm towards a *compute on the edge* modality, which ensures that processing is occurring as close to the data as possible to reduce the bandwidth of data that is transferred over the network and to continue to support scalability as data collections increase. Users now expect the ability to quickly spin up data analysis environments with virtual machines using languages such as Python where they can access, compute, and wrangle data.

Emerging capabilities in cloud computing, search engine technology, content management, machine learning, visualization, data analysis environments, and distributed data systems are bringing stewardship, computation, and analysis together into platforms which can be used to transform how users work with data [5]. As PDS looks forward to increase how users interact with planetary data, there are immense opportunities for PDS to embrace these capabilities and to integrate them into a *planetary data ecosystem*.

At the core of the PDS is its *information model* [6] which captures a set of integrated, cross-disciplinary models for describing and structuring planetary data including common data elements, permissible values, relationships, and other information to structure archives. A major transition from PDS3 (version 3) to PDS4 (version 4) was in moving the model into the center of the system to ensure that changes to the model would directly update the *PDS* system software, thus enabling a *model-driven* architecture. The planetary data model labels data using the XML specification [7]. While in principle, many software architects would argue for this approach, in practice updates to information models and associated data standards often happen independent of software leading to ambiguous search results, which occurred for PDS3. Given the increasing internationalization, heterogeneity of disciplines, instruments, and data, linking the software system and the information model in PDS4 has been a major step forward to drive a self-consistent archive. In addition, a model-driven architecture paves the foundation to not only enhance stewardship capabilities, but also to improve search and to facilitate the labeling necessary to improve indexing, support discipline-specific searches, and taking advantage of emerging capabilities in areas such as machine learning.

As PDS looks at its forward strategy, evolving to strengthen three key capabilities becomes critical: *stewardship, search, and enable more efficient data analysis*. Positioning PDS to take advantage of PDS4 to extend its capabilities to provide data services that span all three areas will enable users to better exploit planetary data at the node and the central PDS search interface, as well as at the international level.

2. Current State of the PDS: A Stewardship Model

The early focus of PDS was on capturing, restoring, and preserving science data from NASA's planetary missions. PDS settled on version 3 of its standard, better known as "PDS3". While PDS significantly advanced NASA's ability to capture science data across disciplines, it was developed to largely document and steward data and provide the data via individual requests on CD and DVD-ROM. The movement of PDS to a fully distributed online system exposed the flaws in PDS3. These included: 1) inconsistent use of metadata across missions, nodes, and international partners 2) the use the Object Description Language (ODL) for structuring and capturing metadata 3) no required format structures for planetary data; and 4) an implicit information model, which was never explicitly validated. In order to drive systematic approaches through software for discovery and use, having consistent metadata is critical to improving search results and navigation. In addition, the data must be structured and consistently labeled to provide long-term opportunities to apply data-driven methods to analyze data.

Over the past decade, PDS made a major push to upgrade to PDS4 in order to drive self-consistency across its archives and to increase the overall *consistent* stewardship of planetary archives. When PDS originally architected PDS4, there were several decisions made which allowed PDS to begin accepting data in PDS4 format, while preserving access to existing PDS3 archives. This intentional strategy meant that PDS grandfathered in existing data pipelines and archives while allowing new data pipelines and archives to deliver data in PDS4. The move also implied that PDS could develop a PDS3-to-PDS4 migration strategy where migration to PDS4 could occur over time.

The PDS4 software architecture strategy follows a system-of-systems design. The top level system is developed by the Engineering Node. Discipline Nodes develop secondary level systems that capture the archival data and address the archiving needs for their community. Since all data is labeled using either PDS3 or PDS4 metadata, registries of those metadata are established that catalog the contents of data both across the system and within a node. *An important architectural tenet is that a PDS metadata label and its associated data can be physically separated.* This enables the Engineering Node to maintain a central registry that points to the appropriate node and their holdings, allowing for traversing from the top level of PDS to the appropriate node. The Engineering Node also has mechanisms to ingest a PDS3 label into the modern, PDS4 registry, supporting access to both PDS3 and PDS4 data collections archived at Discipline Nodes. This architecture drives a multi-level discovery process, but allows for the community to develop discoverable data resources. This includes both the Discipline Nodes and international partners that serve planetary data.

The PDS4 Information Model is comprised of a set of extensible models. New models can be introduced as new types of data are captured in the PDS as shown in figure 2. Models of existing standards can also be integrated to leverage standards, where available within disciplines. Each model is expressed as a data dictionary with stewards that are assigned for their management and curation. The overall information model is architected using modern computer science constructs as an object-relational model (e.g., classes, attributes, relationships, and inheritance). The model itself is captured as an ontology using Protégé, a Stanford-based open source tool. Capturing the model in a modeling tool such as Protégé is important. It provides a database for managing the contents of the model that can be validated for consistency. This is a major change from PDS3 where no such validation occurred. In PDS3, the standards were not explicitly captured as a model, which meant that any data captured in PDS3 would be captured with the notion of an implicit model, rather than an explicit model. This led to inconsistencies that affected everything from stewardship to discovery.

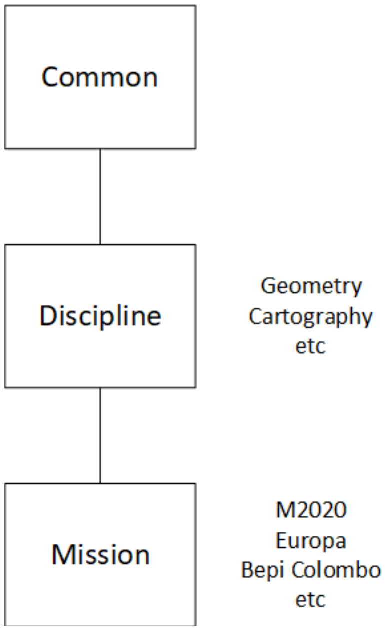

Figure 2: Integrating different models in PDS4

The foundational principles for the PDS4 Information Model are based on two standards, the Open Archival Information System Reference Model [8] and the Metadata Registry Specification [9]. These ISO standards provide the requirements for a “trusted” digital repository and help ensure that the data remain reusable, useful, and interoperable within the planetary science community for the long-term.

For PDS4, the PDS has largely focused on stewardship of data. PDS standards and tools ease the delivery of data supporting the mission-to-PDS interface and ensuring that high quality archives are developed. PDS has excellent processes to preserve and maintain the integrity of its data holdings and to ensure that the data can be interpreted for decades to come. The PDS4 information model is internationally recognized as the foundation on which to build archives. World-wide, agencies that are capturing planetary science results have adopted PDS4 as a standard for structuring their archives and data holdings. The development of such a standard represents a monumental achievement. At the same time, it represents an opportunity for architecting an evolution towards user services. Much of the core infrastructure is in place and while PDS has made some progress, which is detailed below, there remains an excellent opportunity to shift PDS into a new era of enabling more user and data-driven capabilities.

The central registry at the Engineering Node maintains top-level information about archival data as well as common contextual information such as instruments, instrument hosts, targets, and other information. All of this information, including common schemas, data elements, and relationships, are driven by the PDS4 Information Model and today are based on archival metadata labels. All data must conform to that model before it can be loaded.

Top-level searches at the PDS homepage use the contents of the central registry to build their indices. This is done using Apache Solr as a search engine infrastructure which can easily scale to billions of entries. Apache Solr allows for the tuning of rankings to order search results based on relevancy. Apache Solr is tuned based on the information model. It also runs as a backend data service and exposes an Application Program Interface (API) [10] for use by external users to search the central registry. That API is used by the PDS home page to retrieve results that can be presented to users. Search engines such as Apache Solr must be tuned to improve search results based on search use cases and the supporting metadata. PDS should use PDS4 metadata to ensure proper ranking of results for different types of searches (e.g., general vs discipline). For example, PDS may rank title higher than instrument host when performing general searches. This may potentially elevate some search results but lower the ranking of others. It also may include results which may not be relevant. It is important to perform usability testing to improve search engine optimization (SEO). These types of issues are often seen in mainstream search engines that continue to improve search algorithms, metadata, and perform user testing and validation of use cases to improve usability. This type of search validation is important for PDS to also conduct as it more generally moves to implement search engine capabilities across the PDS.

PDS Discipline Nodes steward and distribute the data holdings for their scientific discipline. Node-level searching, where it exists, searches across their respective data holdings to navigate a user to the appropriate data for distribution. In some cases, nodes organize their sites to support navigation to data based on a taxonomy of links and have not implemented full search engines. In other cases, they are using full keyword search engines. In either case, discovery is intended to be optimized for the community and support specialized access and discovery. Search capabilities are developed using a spectrum of technologies from databases to search engines and are all governed independently as described in *Appendix B. The heterogeneous nature of different search engine approaches and technologies can give the appearance of a disconnected search experience for users. However, this can be addressed. The distributed nature of metadata, data, and services themselves can be brought together to provide a more integrated experience by providing integrated APIs and data services that allow for discovery and linking.* Appendices C and D provide information on API usage across the nodes and linking to drive integrated search. Driving consistency in access and linking is a key foundation to improving search integration and usability across the PDS to increase discoverability.

One of the improvements that PDS has made over the past few years is to create mission archive support pages. These are intended to help users better navigate mission archival data with user guides, links to data, and other information to help users boot strap themselves. Currently, the PDS high-level search will point users to these support pages.

Improvements to discoverability and search engine integration will be described in the next section.

3. Increasing Discoverability from Planetary Holdings: Increasing Search Integration

The registry and search architecture is at the heart of what is needed to evolve PDS towards a more integrated data ecosystem. While the distributed, multi-level search approach (from high level to product level) enables broad integration at an international scale, it can also lead to siloed implementations and integration challenges. *As PDS moves forward, it is critical that the PDS focus moves towards unifying the DN search engines approaches in order to drive a more seamless user experience.* To achieve this, PDS must build on its PDS4 investment to:

1. Expand the information model to derive consistent common and discipline queries from the model
2. Support dynamic metadata to allow different types of searching including searching at the file level
3. Support cross-node, cross-product searching using modern search engines (*see Appendix B*)
4. Develop and use common APIs for searching, accessing and retrieving information between systems (*see Appendix C)*
5. Support integration across search engines including passing of parameters so search tools will be automatically configured (*see Appendix D*)
6. Continue to build and register archive support pages to help users boot strap into mission archives
7. Develop a next generation web design leveraging PDS4 metadata, common data services, and integration of search across PDS in order to improve the user experience

The following describes each of the above goals as 3.x. *Appendix E* lays out the current state of adoption of these goals across the PDS.

3.1 Expanding the information model to drive consistent common and discipline searches using data elements as search parameters from the model will enable a long-term extensible approach to build search interfaces that leverage the model-driven architecture. It will also provide a well-architected approach for creating search interfaces and systems from the model, rather than ad hoc implementations that are disconnected using different search parameters. In practice, this means that PDS will be able to launch different types of search interfaces under a search architecture that is compliant with the model. This will drive increased software reuse of search technologies and more consistent searches across the PDS and the different search interfaces.

3.2 Supporting dynamic metadata to allow different types of searching including searching at the file level will transform the PDS search in many ways. This includes the automated extraction of metadata from the data to improve search results from different features and attributes of the data. Some PDS search capabilities have evolved to improve searching on features and other additional metadata such as extraction and classification of different landmarks in images which has enhanced discoverability. As PDS moves forward, expanding search models to support both archival and extended metadata will help position PDS in the future to apply its model not only to PDS and IPDA archival data, but to build towards a global planetary data ecosystem. The model view, expanding from stewardship to improving the integration of archiving, searching, and support for analysis, is shown in the figure below.

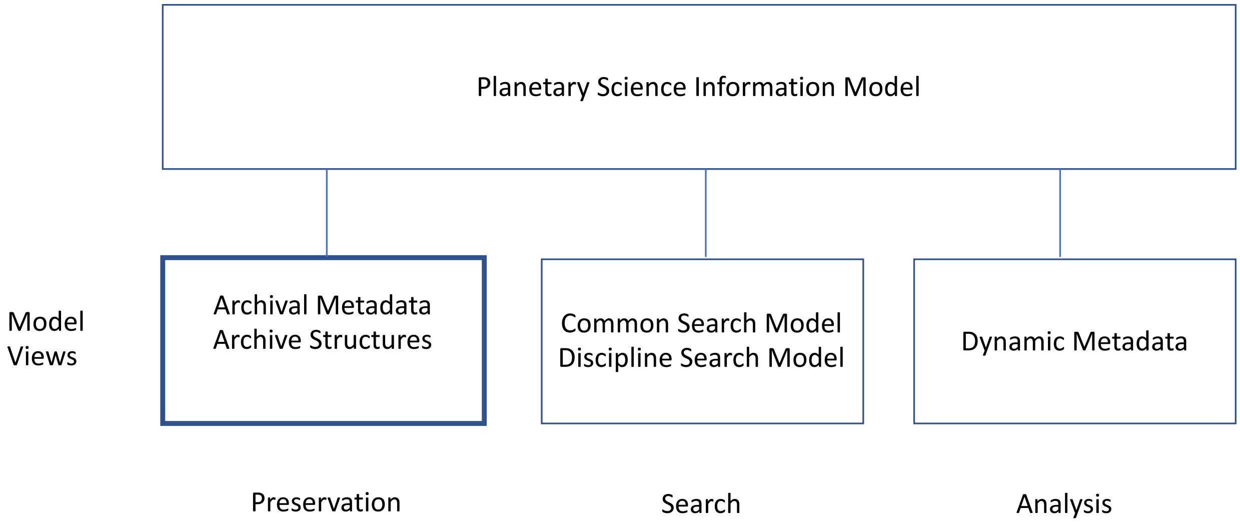

Figure 3: Extension of PDS4 Information Model

3.3 *Supporting integration across search engines including passing of parameters so search tools will be automatically configured* is important for creating a more seamless experience for users. Today, PDS does not consistently pass parameters from one search to another. If users come into a search, enter a set of parameters, and then are pushed to another search, the result can be that they enter at the top of that new search vs passing the original parameters to have the user as close to the data, as possible. *PDS should standardize the parameters that are passed across applications based on the information model*, as described above. Search interfaces should be upgraded to handle those standard terms. Large enterprises which have different tools that work with different data see similar challenges. Even Google has different applications, however, parameters from original queries are passed along so the users search experience appears integrated.

3.4 *Supporting cross-node, cross-product searching* will also bring PDS together as an integrated federation. This would improve how users navigate across the PDS which should appear more seamless. Similar to passing parameters for search integration is the ability to query other nodes to share and combine results from search services. This will require that each node deploy appropriate data services with PDS standard APIs which will be described in the next section.

3.5 *Developing and using common APIs for searching, accessing and retrieving information between systems are* important to expand PDS to fully embrace a data discovery model. As mentioned earlier, using standard APIs for searching data contents is critical. However, moving beyond that will be APIs for accessing, retrieving and even computing on data computing on data seems an odd way to state this will be valuable to allow different types of users the ability to discover, retrieve, and compute on data within the PDS (Sentence is awkward, not sure of meaning here). The implication is that not only will PDS be able to use such APIs, but PDS will publish APIs for users who are developing their own data analysis environments, portals, or local data analysis toolkits in order to significantly open up the access and use of PDs data and services. . The practice will also assist the PDS as it explores different storage and/or computing services (e.g., cloud computing) if systems are developed to decouple interfaces from backend services via APIs allowing the PDS architecture to evolve. This computational scenario will reduce data copying between the nodes by offering data services that better support metadata search and data access across the distributed archives of the PDS so data can be referenced and downloaded on-demand regardless of where it is stored.

*3.6 Build*ing *archive support web pages* helps users quickly boot-strap into mission archives. Such pages have been developed for several missions, including Cassini, and was the basis of the PDS4 deployment for LADEE and MAVEN at several nodes. These pages can be registered and made available to compliment search and discovery. Ensuring that PDS has consistent registered pages can be a basis for the next generation of web design.

3.7 Developing a next generation web design that leverages PDS4 metadata, common data services, and integration of search across the PDS and builds on the infrastructure investments in order to improve the user experience within the constraints of PDS4 - will be a major upgrade. Consistent metadata definitions will contribute to better search within and across PDS. As mentioned, data services will increase data sharing and consistency. With these elements in place, PDS can work to improve the user interface and user experience (UI/UX) by ensuring that the web design can both leverage these capabilities and improve the navigation, consistency, and accessibility of data from PDS.

4. Increasing Access and Computational Support for Planetary Data Analysis

Increasing technical capabilities from cloud computing, machine learning libraries, open source, and other data intensive software services is improving access and computation on massive datasets. These capabilities provide new opportunities to expand PDS data services without having to develop local solutions. *Cloud Computing Services* such as Amazon Web Services or Microsoft Azure have seen major investments in the last decade and provide a number of integrated capabilities that can improve storage, access and computation and can be leveraged in fully hosted vs hybrid architectural models (e.g., capabilities running both locally and in hosted environments). These services largely include storage and application hosting services providing numerous options for scaling, integrating, and accessing data and computation directly from applications. Amazon's Simple Storage Service (S3), for example, provides numerous services for building secure, large-scale data repositories providing standard APIs for uploading and downloading data, services for integrity management and checking, support for disaster recovery, and mechanisms for managing security. Similarly, Amazon provides EC2 elastic computing capabilities to dynamically bring up compute servers running virtual machines that can provide both small and large-scale application hosting and computing. With data locally available, this can provide ample opportunities for not only building web-based data distribution systems, but also facilitates computation directly on the server through hosted analytical data pipelines. Lastly, as toolkits for machine learning and other computational toolkits become more mainstream, there are opportunities to bring data, scalable computation, and analytical libraries together.

Exposing increasing data services including direct access to data, use of common analytic pipelines, on-demand computation, and other capabilities can continue to enhance the usability of PDS, particularly by those users who are developing data analysis tools in such languages as Python. This allows users to exploit the data services directly with their software providing more direct access to PDS data. *In building expanded data services, PDS should look to exploit a balance of cloud computing capabilities and supporting software services, rather than attempting to replicate such capabilities.* Commercial investments in cloud computing have resulted in highly relevant tools and computing capabilities that are co-located with cloud services which can be applied directly to the data and transform PDS to extend its usability. While there are cost constraints, working these at a project level (and above) should yield a strategy for cloud computing that can be coordinated across the PDS and bring forward common tools and services for the community and enable more computing at the edge.

Beyond cloud computing, there are numerous open source tool kits for working with massive data. These include search engines (e.g., Apache Solr, Elastic, etc), different database technologies (MongoDB, Accumulo, etc), toolkits such as notebooks for working with software and data (e.g., Jupyter, iPython, etc), machine learning libraries (TensorFlow, Kafka, etc), and others.

Modern, open source, search engine technologies provide a number of capabilities for searching, ranking, and classifying massively structured data into the billions of objects. Tools such as Apache Solr and Elastic Search allow for building custom indexes and tuning of ranking. Such indexes can use the PDS4 information model as a basis for classifying results. These capabilities also allow for development of faceted-based navigation capabilities that are widely in use by different web-based applications. Much of this approach supersedes early web-based navigation that used taxonomies to organize web links to data and services, or were driven by database parameters which provide drop-drop (drop-down?) lists and other widgets to launch database searches.

Long term, the movement of PDS to leverage search engine technology will enable it to modernize not only its access and search, but also its user experience in providing users more familiar approaches to working with data that is similar to that of Google or Amazon. In addition, as PDS uses newer approaches for deriving additional metadata from archival data to enhance search and discovery, it is important that such approaches can work to capture additional labeled metadata as part of a search index.

Finally, unifying the access to PDS archives and data services with toolkits both for supporting data analysis and bringing in computational libraries will support an emerging paradigm where students familiar with these toolkits can quickly access and wrangle data. Capabilities such as Jupyter Notebooks provide an environment for data analysis which can be integrated with data archives to provide support for pulling in, transforming, and working with data. Furthermore, libraries for working with planetary data and leveraging other analytical algorithms will be useful for analysis of image, time series, and other types of data which can take advantage of common statistical and machine learning packages.

5. Building a Planetary Data Ecosystem

As PDS looks forward, PDS needs an architectural strategy that brings storage, computational, and data services described above together into an integrated platform that can both scale and be extended over the next decade. This will provide a foundation for a planetary data ecosystem where data across the world-wide planetary community can be integrated over time with tools and services regardless of its location. The community should be provided standards which allow them to build supporting software and data services that can be made available in such an environment. Such an ecosystem should be coordinated through the *International Planetary Data Alliance (IPDA)* which should endorse the architecture, associated API standards for access, search, and computation, and agree on shared services. Figure 5 below shows this view with shared services hosted in the cloud environment. Today, much of this is ad hoc as shown in *Appendix A*. The figure separates stewardship including delivery and management of archives from capabilities to exploit search, access, transformation, on-demand computation, and scalable compute services to enable access to the planetary data and computational services by data portals, visualization tools, and user-driven analysis environments. The objective is a platform that works across agencies, institutions, and disciplines to increase access and analytic services.

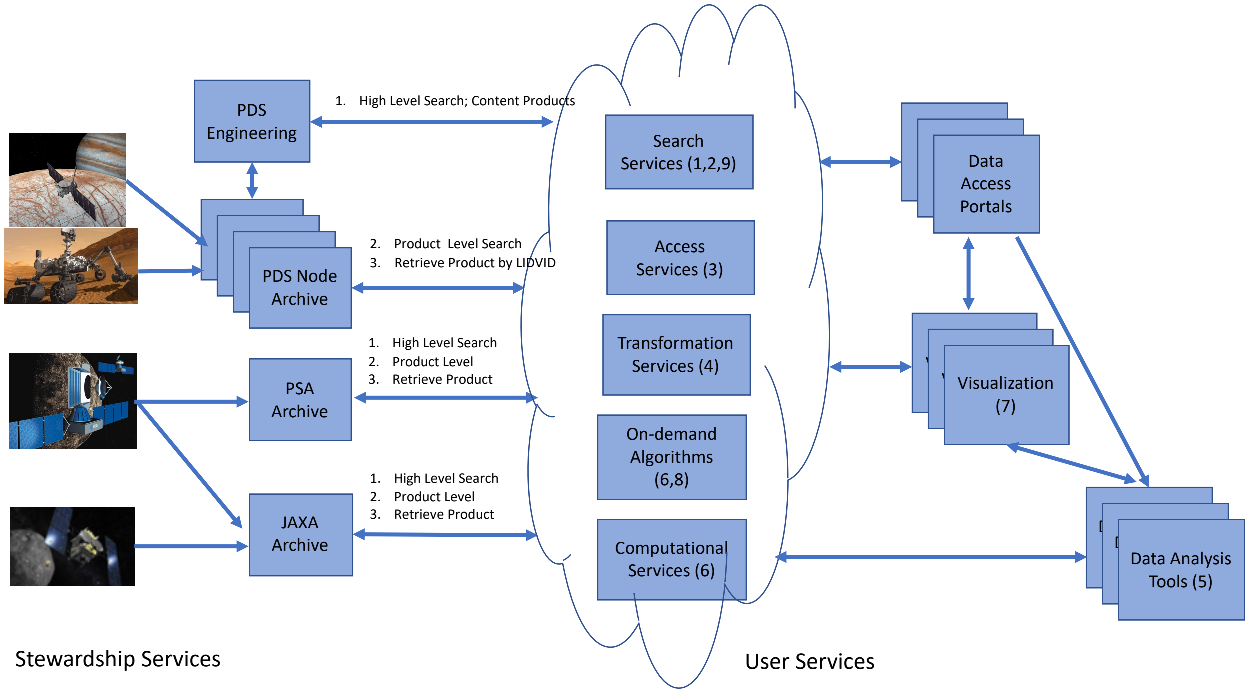

Figure 4: Enabling Discovery, Access, and Analysis

To enable a modern planetary data ecosystem, standard APIs for search, access, and transform as identified above are critical to building an integrated federated architecture. This will lay the foundation to build data services for a planetary data ecosystem. Beyond this, it is important that PDS extend its information model to support increased search of different types of data, both archived and not-archived, as more data resources are brought to integrate. It is also important that PDS provide tutorials and support demonstrating how analysis can be integrated as shown in Figure 4. Finally, it will be important to provide a central portal that can serve as a gateway across the entire international community to access data, services, tools, and to support analysis. PDS should collaborate with UI/UX experts to design the next generation web experience where such a portal capability can be brought up, realized, and become a critical starting point for planetary data discovery and outreach for the planetary science community.

As PDS continues to expand from stewardship to embracing a planetary data ecosystem vision, there is immense opportunity to engage the planetary science software community. In fact, it is critical that this community be involved in developing tools, services and capabilities that leverage PDS standard APIs and services to support the diversity of needs in working with data at an international scale. Coordinating with planetary science software community will create economies of scale that cannot be done by PDS alone. Delivering PDS software services as open source will help in building and coordinating with this community. Venues such as the Planetary Data Workshop and the Planetary Science Informatics and Data Analytics Conference provide an excellent venue for rolling out services and coordinating open source projects.

6. Related Activities

Many of the principles described are being discussed and developed as part of the broader data ecosystem for stewarding, discovering, and using data from scientific research. The FAIR principles [11], "Findability", "Accessibility", "Interoperability", and "Reusability", are being used by several groups to guide development of processes, models, and systems, to ensure that data systems address these principles. Efforts in the physical and biological communities are responding by ensuring the stewardship, discovery, and usability of data are built into the architectures and data services.

NASA established an ad hoc committee on Big Data, which recognized the need to expand from stewarding data to providing increased services to support discovery use across all disciplines within the Science Mission Directorate (SMD) [12]. The objectives initially included developing the appropriate architectures and computing capabilities and services to support the evolution of scientific data analysis as the data, community needs, and technology has both substantially grown and evolved. The study resulted in the establishment of the Strategic Data Management Working Group at NASA Headquarters to guide the roadmap and implementations within and across the divisions.

Significant investments are now being made into the NASA earth science archives to support increased use of data and integration of the Distributed Active Archive Centers (DAACs). Similar to PDS, the DAACs capture archival data at centers across the U.S. https://earthdata.nasa.gov has been developed as a unified portal for searching across the NASA DAACs. Tools such as Global Imagery Browse Services (GIBS) have been developed to provide visualization services for earth science data. Extended portals such as the Sea Level Rise (SLR) have been developed to provide analytical services. Some APIs are documented at https://earthdata.nasa.gov/api, although there is limited uniformity in their adoption and implementation across the DAACs. earthdata.nasa.gov has developed a Common Metadata Repository (CMR) as shown in figure 5 to catalog all data and service metadata records for EOSDIS. It is considered the authoritative source. The metadata is available vis standard APIs that provide a central search of metadata records that span the earth science archives. In addition, the metadata model handle aspects including stewardship and archiving as well as analytical capabilities for visualization, and dynamic metadata to handle future needs as shown in figure 6.

Several DAACs have adopted OpenDAP as a standard API and protocol for retrieving data from the DAACs. There are limited efforts to pass parameters with most DAACs operating different search engines and support differing parameters. Several are beginning to support the use of Jupyter Notebooks and other analysis mechanisms beyond common visualization tools.

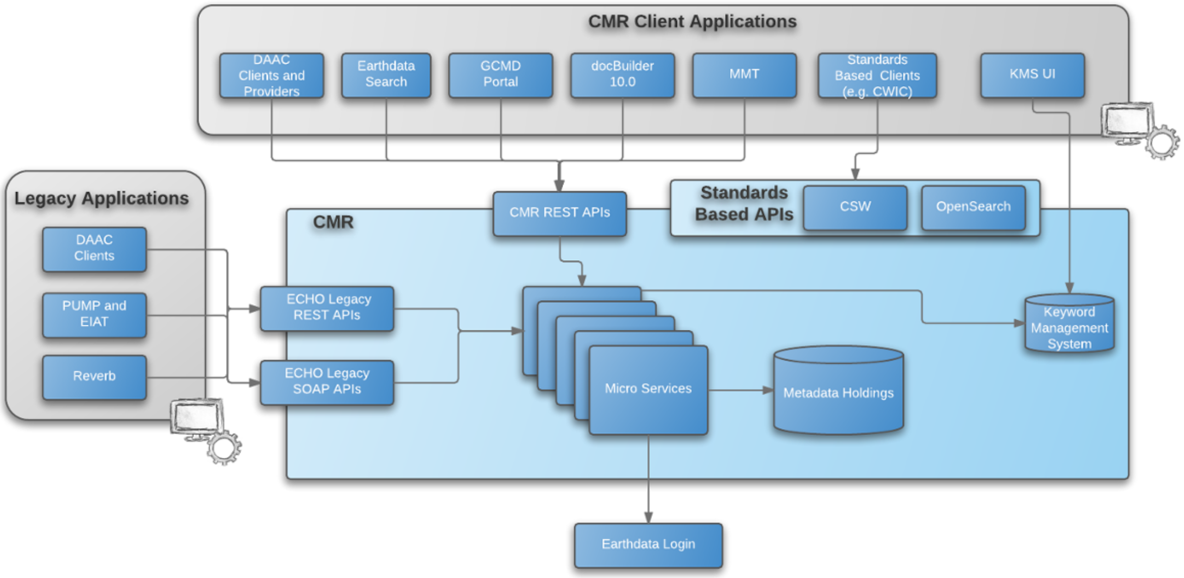

Figure 5: NASA Earth Science Common Metadata Repository

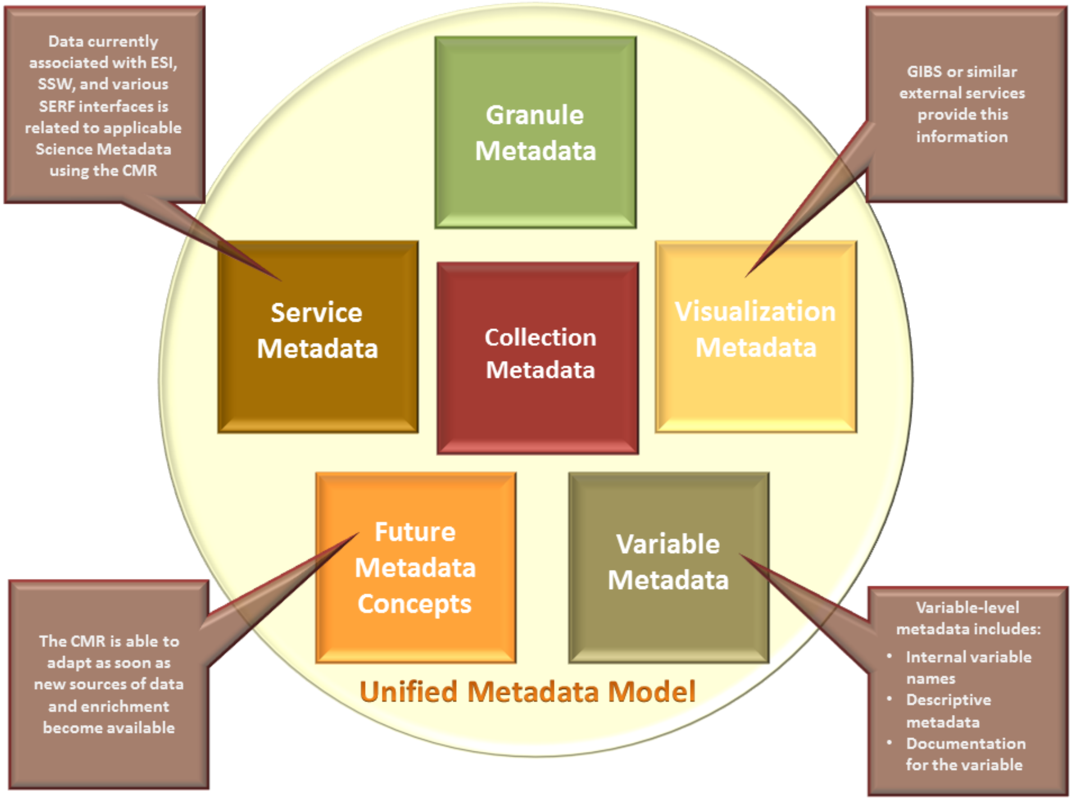
Data currently associated with ESI, SSW, and various SERF interfaces is related to applicable Science Metadata using the CMR
Granule Metadata
GIBS or similar external services provide this information
Service Metadata
Collection Metadata
Visualization Metadata
Future Metadata Concepts
Variable Metadata
The CMR is able to adapt as soon as new sources of data and enrichment become available
Variable-level metadata includes:
• Internal variable names
• Descriptive metadata
• Documentation for the variable
Unified Metadata Model

Figure 6: NASA Earth Science CMR Metadata Model

While astronomy has a smaller formal federated archival governance structure within NASA than earth and planetary science, the astronomy community has developed a number of standards as part of the International Virtual Observatory Alliance (IVOA) to promote discovery and analysis. These include common APIs for search, access, and transformation. Furthermore, several data analysis libraries have been developed that now integrate with open source analysis software stacks focusing on international collaboration for use of the data.

Heliophysics has developed information models for search with the SPASE [13] as well as APIs for accessing data in archives as part of the Heliophysics API (HAPI).

While all disciplines are working to transform themselves to increase data services and support more exploration of data through modern tools, PDS has a real opportunity to bring the community and data together by leveraging its PDS4 architecture to ensure that the model covers stewardship, discovery, and use, with data services using this model. This allows search and support for analysis to grow in a methodical manner to ensure that as data is added, it becomes part of the data ecosystem. This is particularly true for those missions and efforts which are starting with PDS4 from the onset.

7. Conclusions and Recommendations

Embracing stewardship, search, and support for analysis positions PDS to evolve towards embracing a planetary data ecosystem strategy for improving discovery and access for the world-wide planetary community. It also lays the groundwork for evolving to integrate more emerging data science capabilities into the PDS long-term to improve support for data analysis. This builds on the investments that NASA has made in PDS4 that provides a core foundation for discovery and use of planetary data. Further investments to expand capabilities to support the full suite of a modern data science platform will provide many options for increased access, search integration, and the use of new analytical tool environments. These can be implemented as a phased set of capabilities, both short and long term, focused on mission priorities, in order to bring capability to the community as early as possible, as shown in the figure below. These phases can also be scoped to fit the resources available to PDS since the conceptual vision identified in this white paper can scale based on resources.

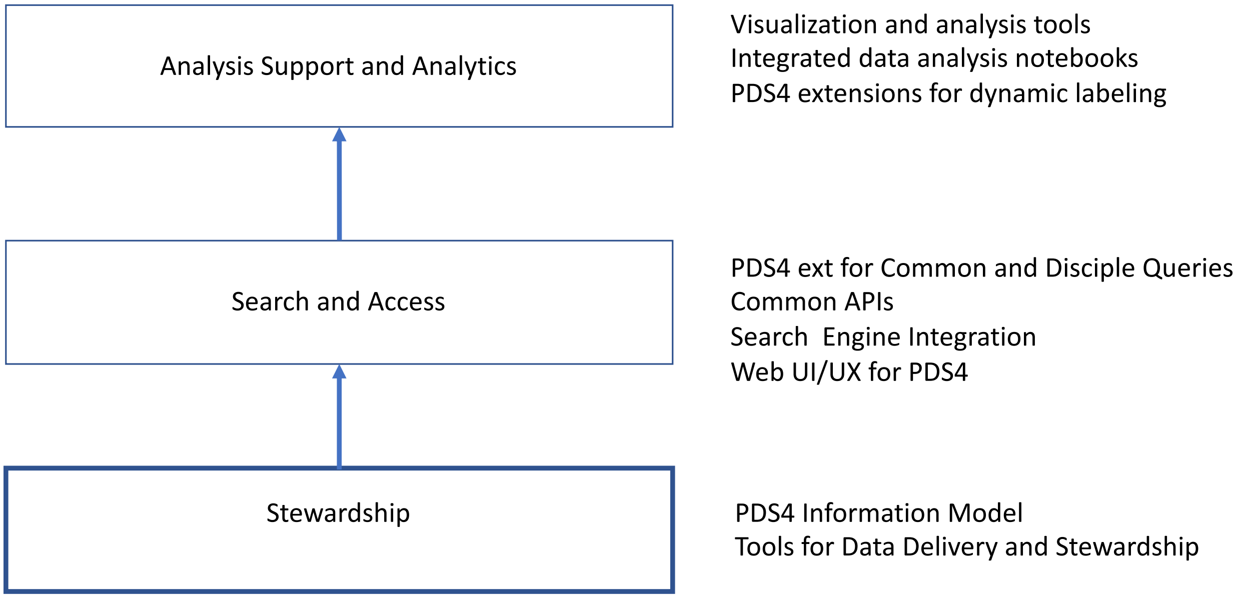
Analysis Support and Analytics
Visualization and analysis tools
Integrated data analysis notebooks
PDS4 extensions for dynamic labeling
Search and Access
PDS4 ext for Common and Disciple Queries
Common APIs
Search Engine Integration
Web UI/UX for PDS4
Stewardship
PDS4 Information Model
Tools for Data Delivery and Stewardship

Figure 7: Notional Concept to Move PDS Towards a Data Services Architecture

The notional plan identified below is broken into two phases. The first phase lays a foundation for increasing discoverability from foundational services, particularly integration with modern search services. The second integrates those services towards enabling improved use of the data with the integration of modern tools and an upgraded web presence. *Appendix F* lays out a mapping using Juno as a potential use case for driving a project plan forward.

Short-term (2 years): Increasing Discoverability from Planetary Holdings

Increase access and sharing of services through a common API, search integration, and deployment of transformation-as-a-service.

1. Develop an architectural specification for the planetary data services.

2. Expand the PDS information model to identify core search models for integration across the PDS.
3. Standardize PDS search, access, and transform APIs, in conjunction with IPDA, and post to the PDS website.
4. Implement transform as a common web service for use across PDS (Not clear here?).
5. Implement search engines at nodes, where needed, to support product level search
6. Adopt standard search APIs at PDS nodes.
7. Adopt parameter passing between PDS nodes.
8. Prototype common, PDS-wide shared services in the cloud.

Long-term (5 years): Building a Planetary Data Ecosystem

Drives new approaches for enabling data analysis for a next generation of researchers.

1. Work with NASA Headquarters and the IPDA to develop a planetary data ecosystem vision and associated requirements.
2. Upgrade PDS websites to new UI/UX design to support improved services.
3. Develop additional models in the PDS to support discipline search and analysis extensions.
4. Migrate all PDS4 data and applications to support common APIs, parameter passing, and data sharing across nodes.
5. Support integration with other international archives and databases.
6. Develop common operational concepts for data analysis services.
7. Support common data and computational service analysis needs on the cloud.
8. Support increased use of dynamic metadata generation for search indexing.
9. Support tool integration with Jupyter Notebooks and other analysis tools kits.

Appendix A – State of Standard Data Services

The following describes the current state of standards for implementing figure 4 in the conceptual architecture diagram. This includes international support.

Capability	**IPDA and/or PDS Support**
1a. High Level Search within an Archive (PSA, PDS, etc)	PDAP; EPN-TAP; PDS Search Service API; PDS and PSA supporting services interfaces.
1b. High Level Search across archives (PSA, PDS, etc.)	PDS Search Service API; Search integration (e.g., passing search parameters) inconsistently implemented.
2a. Product Level Search within a Site	Differing support for REST-based API access
2b. Product Level Search across Sites	No PDS or IPDA-wide product-level search; VESPA has some support

3. Retrieve Product	Most archives/node provide HTTP access for download. Inconsistent services for download (label, data).
4. Transform	PDS library exists but no service
5. Tool Integration	Integration with tools such as ISIS; no support for Juptyr notebooks; no support for languages such as R
6. Computational Support	Limited support for on-the-fly processing and running analytical results.
7. Visualization Support	Solar System Treks; VESPA
8. Metadata Extraction	Local node activities
9. Indexing on dynamic metadata	Local node activities

Appendix B – Current State of Search Engine Support Across the PDS

Different canonical search approaches are used across the PDS. These include:

- Full keyword search with faceted navigation (e.g., driven by modern search engine technology)
- Keyword searching with parameters (e.g, driven by either search engine or database technology)
- Parameter searching (e.g., generally driven by relational database technologies)
- Navigation via a taxonomy of links (e.g., Old Yahoo! web navigation model)

As PDS moves forward, given significant advancement of search engine technology and UI/UX approaches, PDS should adopt a unified search engine strategy with facet and search models that are derived from the PDS4 Information Model.

Node	**Search Engine Approach**
Atmospheres	Web-based navigation via mission/instrument *taxonomy* at dataset/bundle and product level. Atmosphere provides SPHERE as a tool for parameter-based searching.
Engineering	Provides *faceted-based search and navigation* across the PDS archive at the dataset/bundle level. *No product level search across PDS. Facets are derived from the PDS4 Information Model.*
Geosciences	Provides *parameter-based search* capabilities at the dataset/bundle and product level within Geosciences.
Imaging	Provides *faceted-based search and navigation* capabilities at the dataset/bundle and product level within Imaging. *Facets are unique to Imaging.*

NAIF	Web-based navigation via mission *taxonomy* at dataset/bundle level.
PPI	Provides *keyword search* capabilities at the dataset/bundle and product level within PPI.
Ring Moon Systems	Provides *faceted-based search and navigation* capabilities at the dataset/bundle and product level within Ring Moon Systems. *Facets are unique to Imaging.*
Small Bodies/UMD	Different taxonomy navigation approaches are applied to support UMD and PSI needs. Provides parameter-based search capabilities by target within PSI.

Appendix C – Current State of PDS Node Search APIs

Search of data holdings via an API allows for data to be discovered for different purposes from being able to refer users to data across PDS to enabling external access and use by community tools. The following identifies the current status of node API implementations.

Node	**API**	**Documented**
Atmospheres	PDS Search API for PDS4	Documented
Engineering	PDS Search API; PDAP support	Documented; In the tool registry https://pds.nasa.gov/tools/tool-registry/
Geosciences	ODE provides a REST-based API for data in ODE	http://oderest.rsl.wustl.edu In the tool registry
Imaging	REST Search API for IMG	https://pds-imaging.jpl.nasa.gov/tools/atlas/api/
NAIF		
PPI	HAPI	Documented
Rings	REST-based API	Documented; publicly accessible
Small Bodies		

Appendix D – Current State of PDS Search Parameter Linking

Passing parameters between search engines to support *product level search* increases the integration across the PDS federation by ensuring users are forwarded as close to the data, as possible, between node search tools. In many cases, product level search services either lack any parameter passing or parameters differ across node services. Long term, PDS should both provide consistent approaches to product level search and parameter passing that are derived from the PDS4 model and passed to services across nodes to support an integrated end-to-end search chain.

The following provides a status of the current status of parameter passing and integration from the high-level search at pds.nasa.gov to a discipline node.

Node	State of Parameter Passing
Atmospheres	For PDS3, accepts links to online repository for specific data. For PDS4, accepts links to online web resources (e.g., collections).
Engineering	Both accepts search parameters and provides parameter passing to each registered search service from the high-level search to integrate product level search.
Geosciences	Analyst Notebook does not accept passed parameters for product level search. Geosciences Web Services does support links to specific data. Online repository links to the data resources.
Imaging	Integration varies by dataset. Accepts limited set of parameters that are passed onto Atlas (e.g., mission name but a specific dataset identified as not passed on
NAIF	Accepts links to online data repository and various SPICE tools for specific data.
PPI	Accepts parameter passing to the PPI website to link to specific data.
Rings	Accepts links to specific pages for web pages, online repositories, and OPUS.
Small Bodies	Accepts links to specific pages for web pages.

Appendix E – Current State of Data Services Support Across the PDS

There are four elements in enabling improved access. These have been identified in appendices B-D and include archive support pages, product level search, remote API support, and linking of search parameters to support integration of the search chain. The integration of all four provide an integrated strategy for PDS to bring the federation closer together to support discovery and access.

Node	Archive Support Pages*	Product Level Search	API Support	Search Engine Linking
Atmospheres	Y	Minimal	N	N
Engineering	N/A	N/A	Y	Y

Geosciences	Y	Minimal	Minimal	Minimal
Imaging	Y	Y	Y (IMG specific)	Minimal
NAIF	N/A	N	N	N
PPI	Y	Y	Y (HAPI)	Y
Rings	Y	Y	N	Minimal
Small Bodies/UMD	Y	Minimal	N	N

* Many archive support pages are appropriately hosted by the lead nodes.

Appendix F – Use Cases Example: Increasing Discoverability of Juno Data in the PDS

This mission is a polar orbiter designed to investigate the near magnetic field, the mass distribution within the planet and the vertical structure of Jupiter's atmosphere. It is a spin-stabilized, solar powered craft in a 52-day elongated orbit where the craft is near perijove for about 10 hours.

Data from the mission is archived in 4 nodes. ATM –MWR. UVS, JIRAM and Gravity. Cartography and Imaging –JunoCam and PPI – FGM, JEDI, JADE and WAVES. SPICE data are archived at the NAIF node.

Data Types in the Juno Mission

Microwave Radiometer (MWR) has 6 antennae. It scans across the planet creating one row in an ASCII file for each observation in the scan. The team submits a data file and a geometry file, where each row provides the total information needed to understand that data point. They have arbitrarily decided to limit the time spanned by files to one-hour earth time. File sizes range from 1.5 to 85 Mb. The antennae are left on so there are 24 files/day. Near planet data deals with atmospheric structure and distant data can be used to understand Jupiter's interaction with the Solar Wind.

Ultraviolet Spectrograph (UVS) is a pulse counter where each data point involves the location of the photon on the detector and determination of the frequency of the photon. You need a lot of data points to map out a usable data product. File sizes are arbitrary and range from 14 to 1240 MB, mostly large. This data is delivered in a FITS file consisting of several planes containing data, housekeeping, etc. Data is obtained at different times in the orbit to acquire differing resolutions.

Juno Infrared Auroral Mapper (JIRAM) is an IR spectral mapper measures the atmosphere down to pressures of 5-7 bars. It is a combination instrument. In the forward focal plane the detector is split in half with 2 filters. One filter is sensitive in a frequency interval where methane is highly absorbing (of both incident sunlight and IR radiation coming up from below) causing the globe of Jupiter to be dark with the aurora, suspended above the methane, to be bright. The other filter, chosen at a frequency where methane is not a good absorber allows lower radiation to escape, thus mapping the heat loss from the planet. A slit in the second part of the focal plane passes the light to a second camera that documents where the instrument was looking. These data are delivered as standard 2-D files with detailed labels that specify many geometric parameters. This data is concentrated near perijove.

GRAVITY – radio occultation data PDS4 formats for this are still being formulated, however, the team is delivering data with initial PDS4 labels. The natural data block for this type of observation is essentially an experiment. The user would want to download the data for a selected orbital pass; thus, the natural delivery would be a compressed file for each orbit (~20 so far).

JunoCam is a wide-angle, 4-color, typical Malin camera; however, it was funded for educational outreach and has little user support. The raw and calibrated data are delivered in typical image files and files are less than 10 Mb. Amateurs decide what will be observed and JPL maintains a site for display of produced products – some useful for introduction to cloud structure and some pure art.

Magnetometer (FGM) produces the first global magnetic mapping of Jupiter. Each data product is similar to the MWR. It is an ASCII file containing a time series of magnetic field vectors in geophysical units (nanotesla, nT) that have been corrected for instrumental and spacecraft effects (calibrated). In addition, these data have been transformed into a physically meaningful coordinate systems. This geometry is incorporated as additional columns in the file. The file sizes are tens of Mb. Here, users are probably downloading the data from each perijove passage and building a dataset that allows them to refine their initial magnetic field model as the mission goes on. We now have a loose net around the planet and additional data will fill in the gaps.

The following 3 data sets would be used to refine the magnetic field model and JADE can be combined with UVS and JIRAM to study the magnetic field–upper atmosphere aurorae interaction.

Jupiter Energetic Particle Detector (JEDI) files are delivered in scientific units (particles/cm$_2$-sr-s-keV) as a function of incident energy and direction in time ordered ASCII CSV files. The position and attitude of the spacecraft and the measured pitch angle between the look direction and the local magnetic field for each telescope are included in this product. The file sizes are less than ten Mb.

Jupiter Auroral Distributions Experiment (JADE) Files are more than 100 Mb. There are 5 products :

Time ordered counts per second in energy vs. look direction,	Binary
Time ordered {electron or ion} flux vs. direction vs. energy.	Binary
Time ordered ion flux vs. energy vs. TOF.	Binary
Time ordered electron pitch angle distribution vs. energy.	Binary
Time ordered plasma moments vs. composition.	ASCII

Radio/Plasma wave experiment (WAVES) There are 2 major products delivered in ASCII with files in the range of 2 to 10 Mb.

SPICE A complete set of SPICE kernels used to compute many kinds of observation geometry parameters is archived at the NAIF node. SPICE data have already been used by the JUNO experiment teams to compute observation geometry parameters that are included in the instrument data archives. But future researches may wish to use the underlying SPICE data to compute additional parameters, to compute parameters at different times than those used in the instrument archive, or to compute improved parameter values if some portion of the underlying SPICE data are improved subsequent to delivery of the instrument archive.

Mapping Juno Use Cases to Data Services Capabilities

The following table provides a mapping between the identified strategic data services to be provided by PDS and its application to support increased access and discoverability to the data.

Data Services Capability	Application to Juno
Archive Mission Pages	Atmospheres registers the Juno Mission page as the official PDS landing page for all information about Juno. Other information may exist within PDS, but search engines refer novice users to the archive mission page which allows them to bootstrap themselves into the data with background information on instruments, archive organization, and structure of the data.
Cross-Node Searching	Users will use PDS4 keywords to search for data. Searching for Juno data will generate results of archival data at 3 nodes: Atmospheres, Cartography and Imaging, and PPI. Search results include links to datasets and services at nodes for further product level refinement. Users can launch searches and link to data services from any node in the PDS.
Search Engine Integration across the Nodes	Users searching PDS for Juno data will use PDS4 keywords which are passed between node search engines to get users as close to the data with as few steps as possible. In the case of Juno, users studying the magnetic field-upper atmosphere aurorae interaction would search for data sets including JADE, UVS, and JIRAM to support magnetic field model improvement and refinement.
Product Level Searching	Users perform searching of products within a dataset to refine search results. Discipline specific PDS4 keywords are used to support both file and record level searching to support extraction and subsetting of data. This can be organized so sets of data (e.g., all files for a specific pass) can be downloaded.

Remote API Access	Users can script downloads including specific products with constraining PDS4 keywords in order to download larger sets and combinations of data from PDS. Nodes have consistent APIs so users can point their scripts to download the variety of data needed. For larger files, browsers delivery via scripts allows downloading to occur in the background.

REFERENCES

[1] *Special Issue: The Planetary Data System*, Planetary and Space Science, European Geophysical Society, ISSN 0032-0633, Volume 44, Number 1, January, 1996.

[2] R. Arvidson, Ed., *Issues and Recommendations Associated with Distributed Computation and Data Management Systems for the Space Sciences*, Committee on Data Management and Computation (CODMAC), National Academy Press, 1986.

[3] *IPDA Progress Report*, 2018

[4] Crichton, D. J., Hughes, J. S., et al, *A Scalable Planetary Science Information Architecture for Big Science Data*, Published in IEEE 10th International Conference on e-Science, 2014.

[5] National Research Council (U.S.). *Frontiers in Massive Data Analysis. 2013.*

[6] *PDS4 Architecture Specification,* 2013.

[7] *Extensible Markup Language (XML) 1.0 (Fifth Edition),* W3C Recommendation, 26 November 2008.

[8] *Open Archival Information System (OAIS) Reference Model (ISO 14721)*, 2003.

[9] *Metadata Registry Specification (ISO/IEC 11179),* 2004.

[10] *PDS Search API,* 2014.

[11] Wilkinson, M, Dumontier, M, *The FAIR Guiding Principles for Scientific Data Management and Stewardship*, Nature, March 15, 2016.

[12] NASA ad hoc committee on Big Data, *6th and Final Report of the Big Data Task Force*, November 28, 2017

[13] *SPASE: Space Physics Archive Search and Extract*, http://space-group.org , May 31, 2018.

[14] *PDS Level 1/2/3 Requirements*, 2017.

[15] *PDS4 Information Model, V1.12.0.0*, 2019.

ACKNOWLEDGEMENTS

Chuck Acton, NASA Jet Propulsion Laboratory
Ray Arvidson, Washington University St Louis
David Baker-Hollibaugh, NASA Goddard Space Flight Center
Maria Banks, NASA Goddard Space Flight Center
Gerbs Bauer, University of Maryland
Reta Beebe, New Mexico State University
Nancy Chanover, New Mexico State University
Dan Crichton, NASA Jet Propulsion Laboratory (Author)
Steve Hughes, NASA Jet Propulsion Laboratory
Lisa Gaddis, USGS
Mitch Gordon, SETI Institute
Ed Guinness, Washington University St Louis
Emily Law, NASA Jet Propulsion Laboratory
Tim Mcclanahan, NASA Goddard Space Flight Center (Author)
Tom Morgan, NASA Goddard Space Flight Center
Jordan Padams, NASA Jet Propulsion Laboratory
Paul Ramirez, NASA Jet Propulsion Laboratory
Mark Showalter, SETI Institute
Tom Stein, Washington University St Louis
Ray Walker, University of California, Los Angeles