DATA MANAGEMENT, PRESERVATION AND THE FUTURE OF PDS

Reta Beebe - New Mexico State University, Las Cruces NM Telephone: 575-646-1938
Email: rbeebe@nmsu.edu

Co-Authors
Acton, Charles - Jet Propulsion Laboratory, Pasadena CA
Arvidson, Raymond - Washington University, St Louis MO
Bell, Jim -Cornell University, Ithaca NY
Boice, Dan - Southwest Research Institute, San Antonio TX
Bolton, Scott - Southwest Research Institute, San Antonio TX
Bougher, Steven -University of Michigan, Ann Arbor MI
Boynton, William -University of Arizona, Tucson AZ
Britt, Daniel -University of Central Florida, Orlando FL
Buie, Marc - Southwest Research Institute, Boulder CO
Burns, Joseph - Cornell University, Ithaca NY
Capria, Maria Teresa – IASF-INAF-Roma/Past chair of IPDA, Rome IT
Coradini, Angioletta - IFSI-Roma/PI of Juno/JIRAM, Rosetta/VIRTIS & DAWN/VIR Rome IT
Crichton, Daniel - Jet Propulsion Laboratory, Pasadena CA
Ford, Peter - Massachusetts Institute of Technology, Cambridge MA
French, Richard - Wellesley College, Wellesley MA
Gaddis, Lisa - U.S. Geological Survey, Flagstaff AZ
Gierasch, Peter - Cornell University, Ithaca NY
Gladstone, Randy - Southwest Research Institute, San Antonio TX
Gordon, Mitch - SETI Institute, Mountain View CA
Greeley, Ronald - Arizona State University, Tempe AZ
Hansen, Kenneth - University of Michigan, Ann Arbor MI
Jakosky, Bruce - University of Colorado, Boulder CO
Kasaba, Yasumara - Tohoku University/Current Chair of IPDA, Sendai City, JP
Khurana, Krishan - University of California Los Angeles, Los Angeles CA
Kurth, William - University of Iowa, Iowa City IA
Law, Emily - Jet Propulsion Laboratory, Pasadena CA
Lorenz, Ralph - JHU Applied Physics Lab, Baltimore MD
Nixon, Conor - Goddard/Univ. of Maryland, Greenbelt. College Park MD
Paranicus, Chris - JHU Applied Physics Lab, Baltimore MD
Pryor, Wayne - Central Arizona College. Coolidge AZ
Roatsch, Thomas - DLR Institute of Planetary Research, Berlin DE
Russell, Chris - University of California Los Angeles, Los Angeles CA
Schwehm, Gerhard - European Space Agency, Villa Franca ES
Simpson, Richard - Stanford University, Palo Alto CA
Sykes, Mark - Planetary Science Institute, Tucson AZ
Tholen, Dave - University of Hawaii Oahu, HI
Walker, Raymond - University of California Los Angeles, Los Angeles CA
Withers, Paul - Boston University, Boston MA
Zender, Joseph - European Space Agency, Noordwijk NL

PREFACE

Efficient, effective archiving and distribution of data is an integral part of planetary science research. We strongly encourage the decadal committee to support the concentrated effort currently underway to evolve the Planetary Data System (PDS) from an archiving facility to an effective on-line resource for the NASA and International communities (the PDS2010 project).

We urge the committee to incorporate the following three issues in the final report:

- Identify as a high priority, the need for broad emphasis from the NASA Planetary Science Division to assure that its policies and procedures guarantee adequate, consistent support for data analysis within the missions and the community and to enable effective archiving.

- Strongly recommend that future NASA Planetary Science Division NRAs and AOs include specific requirements that in addition to raw data, missions and instruments provide data in physical units. Archive planning should be an integral part of the proposal planning, and funding should be identified in the award to ensure teams have adequate resources to meet this additional obligation.

- Strongly recommend that the NASA Planetary Science Division support the upgrade of PDS including leveraging modern data base and Web 2.0 technologies in order to ensure improved data standards and efficient, effective storage, search, retrieval and distribution of scientifically useful planetary data in the coming decades.

Background and Justification

1. Introduction

Planetary exploration by spacecraft represents significant national investments that cannot be easily repeated. Return visits by more capable spacecraft are rare and depend on results obtained from precursors. Events observed by Earth-based or *in situ* instruments often unfold slowly and are not repeatable. Scientists need access to original data to verify reported results, to test new insights and theories, to carry out time-dependent studies, and to assess limitations in our knowledge so that future observations can be planned.

If an archive is comprehensive, readily accessible, and usable, it can meet the needs listed above — it can serve as a virtual reflight of missions and observing campaigns preserved in its contents, at a cost which is minuscule compared with acquiring the original data. However, creating and maintaining a high-quality archive requires commitment from the funding agency, the data providers, and the users—a point which was recognized by the National Academy of Science in its mid-decadal "report card" (ref.: Grading NASA's Solar System Exploration Program: A Mid term report – Co-chairs W. Huntress and N. Noonan (2008) ISBN:0-309-11493-4).

In the remainder of this paper we discuss the present state of data management and archiving within the Planetary Science Division and our recommendations for improvement within the PDS 2010 framework.

2. Background

Although NASA had been including language in contracts for several years that required data from planetary missions be submitted to the National Space Science Data Center (NSSDC), it was clear by the 1980s that a more methodical process with better user access was needed. The NSSDC collection was important as a deep archive, but, because of the lack of a process that provided direct interaction of mission teams and qualified scientists for assessing development of the products, its contents were highly variable in terms of both quality and content. Viking and Voyager stimulated interest both in 'data mining' (searching acquired but previously unexamined data) and reanalysis (seeking new discoveries from previously studied data) and the demand for direct access to mission products increased.

After a study and a prototype phase, the former Solar System Exploration Division (SSED) at NASA established the PDS in 1989. The PDS was a distributed system, with a central node (incorporating both management and engineering functions), supporting nodes (imaging and the Navigation and Ancillary Information Facility (NAIF)) and discipline nodes (DNs) responsible for science data at home institutions that qualified as 'centers of excellence' in atmospheres, geosciences, particles and fields, rings and small bodies. Creation and structure of PDS were based on recommendations from the National Academy of Science (NAS) Committee on Data Management and Computing (CODMAC) (1982, 1986 and 1988) that archives should be housed with science expertise (and, within the context of dealing with multiple short-lived missions this recommendation has proven to be a workable solution for integrating science discipline expertise into the archived products). It was established that PDS would explicitly serve the SSED-funded community and NSSDC would receive copies of PDS data sets for permanent archiving and distribution to non-NASA researchers, international scientists, educators, and the general public. Early data transfers were by magnetic tape and later by CD and DVD physical media, which led to the use of ISO9660 compatible structure and naming conventions that are still in use.

PDS formation was roughly coincident with the birth of the world wide web. In the ensuing two decades, improved technologies produced ever increasing data complexity and data volume while network communication transformed both how PDS did its business and how it interacted with both data providers and data users. The 'distributed' system that was designed for dataset exchange via tape or CDs was integrated so that queries for data could be submitted not just from home institutions but from personal computers from homes, hotel rooms and foreign shores. Instrument teams began delivering terabytes of raw and partially processed data; calibration files were continually being revised, leading to new versions of higher-level products. And users began asking for not only more support, but more sophisticated support—*could PDS provide all of the atmospheric temperature-pressure profiles over the PHOENIX landing site, could images from the Shoemaker-Levy 9 Jupiter encounter be recalibrated, and did near-infrared spectra exist (from any source) of asteroid 4370 Dickens?*

The PDS evolution initiated by the web revolution has been uneven and strongly constrained by a limited budget. With the significant increases in data volumes (Figure 1), the challenge of capturing and protecting the bits themselves is daunting. On the other hand, the process for assuring the long-term integrity of the data has become more generally tractable – PDS data are now monitored and distributed with checksums, replicated at mirror sites and each node has a backup and recovery plan.

While modern web services and the more complex needs of the science user communities have produced substantial increases in expectations with regard to high granularity access and highly processed data, leverage to assure uniform delivery of data from instrument teams has been severely limited. For example, some instrument teams want to deliver data in a range of processing states; others are content to make their raw binary files public and let others generate higher-level products. PDS representatives work with instrument teams, often with widely differing approaches to data archiving, to ensure as much uniformity as possible in presentation at each processing level and to produce documentation that is understandable to users familiar with the field. The situation has been hampered by NASA's earlier inattention to enforcing delivery of calibrated products and to assuring that adequate funding was reserved by the missions to allow the teams to produce standard products, especially at the highest processing levels, which are in greatest demand.

Another issue that the PDS confronts is calibration. For some instruments, the calibration of data is an ongoing process that can take years or longer. The reasons for this vary and include such diverse issues as the accumulation of sufficient data to draw proper conclusions to working with flight spares to revisit calibration issues. Data analysis is a process. This is why raw data is often not useful and calibrated data can often be eclipsed. The resolution of this problem is not to offer nothing, as has been the case in the past, but provide cautions and to offer intermediate products which are the best products that can be reasonably provided at a given time, and to provide for a final set of calibrated products once the calibration process has been stabilized or at end of mission.

In 2005 NASA reorganized the PDS, moving the management to Goddard Space Flight Center and reorganizing the JPL-based engineering node. At the same time NASA became more vigilant in requiring that missions plan and budget for data analysis and archiving.

3. The Diversity of Planetary Science

Planetary data are acquired with flyby and orbiting spacecraft making both remote and in situ measurements, surface stations, rovers and sample return missions. Mission lifetimes range from months to decades. Archiving this diverse reservoir as well as supporting ground-based observations, laboratory data and spacecraft radio tracking and engineering information is challenging. The need to apply standards that assure long-term preservation and data integrity imposes additional constraints on PDS policies and procedures. PDS differs from a facility such as the Space Telescope Science Institute (STScI), which deals with an accumulating archive from a few very specific instruments. HST has developed a data pipeline and can provide an on-the-fly calibration service based on the latest and best calibrations requested by users. However, in the course of its nearly two decades of operation, HST has obtained data from 18 remote

sensing instruments. In contrast, PDS works with a much more diverse set of instruments and teams where virtually all planetary exploration 'observatories' are born and die on time scales that are short when compared to HST's operating lifetime.

During the first half of 2009, PDS ingested data from 82 instruments on 17 spacecraft ranging from three-dimensional *in situ* magnetometer data to gigabyte image strips from push-broom cameras. By 2015 the PDS is projected to house data from 511 instruments from 70 spacecraft. This will result in an estimated data volume of 245 terabytes from the individual data sets included in Figure 1.

Each planetary mission defines its observations, collects its data, and deposits its results in the archive within a few years. Funding disappears before calibration on many instruments is fully mature. In addition, lack of oversight and mission funding to produce higher-level products from the wide range of instrumentation, and divergent community practices among disciplines have led to PDS data sets that are not easily compared with each other and (in some cases) poorly understood except by those who were involved in the data acquisition. The reality is that PDS must support many different data pipelines each optimized for its mission and instrument.

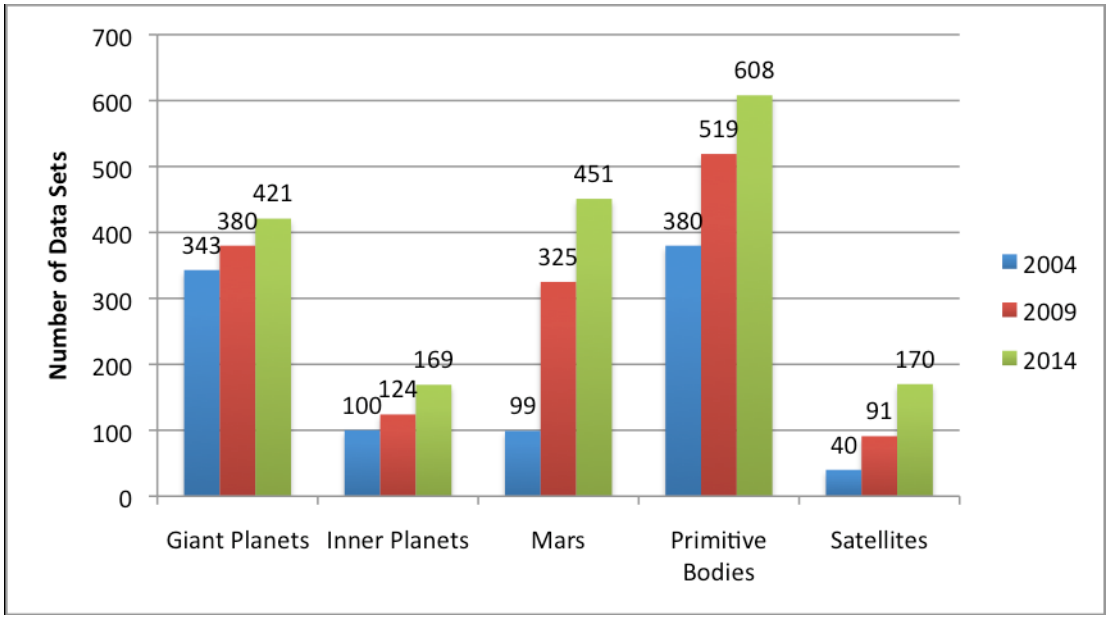

Figure 1. The Estimated Number of Accumulated Data Sets Per Decadal Category. Note: Lunar missions are included in the satellites category. Possible contributions from the following missions have not been included due to lack of information or because they are scheduled beyond 2014 -- **Satellites:** Grunt (Russia) Phobos, Yinghuo-1 (China) Phobos, Chang'e 1 (China) Lunar, Chang'e 2 (China) Lunar, Kaguya (Selene) Lunar, SELENE 2 (Japan) Lunar, Chandrayaan 1 (M3, MINI-RF) Lunar, Chandrayaan 2 (India) Lunar, Luna-Glob (Russia) Lunar, Lunar mini-Landers Lunar, MoonNEXT (ESA) Lunar, MoonLITE (UK) Lunar, Smart-1 (ESA) Lunar, LAPLACE (Ganymede)(ESA) – **Mars:** ExoMars (ESA) Orbiter, ExoMars (ESA) Lander, MarsNEXT (ESA), Mars Sample Return – **Inner Planets:** Bepi Colombo (ESA), Venus Express, Venus Climate Orbiter (JAXA) – **Giant Planets:** Outer Planet Flagship (launch 2016) – **Miscellaneous:** Discovery AO-2008/9, Discovery AO-2010, New Frontiers 3.

To access a summary of timelines for missions in operation and under development see http://atmos.nmsu.edu/pub/download/NASA_Mission_Summary_special_022409.xls

4. International Implications

The Decadal Survey represents an opportunity to raise awareness of the rapidly changing data management and archiving requirements for the next decade(s) and to recognize the growing trend for international cooperation on missions and in data sharing. Along with the growing internationalization of space comes an urgent need both to ensure the preservation of and to provide access to an ever-increasing volume of usable planetary data worldwide. When ESA began plans to establish their Planetary Science Archive in the mid-1980s, influenced by the need for rapid progress and the fact that there was already an experienced ESA cadre of PDS users, they adopted PDS data standards. Subsequently, when India's ISRO selected NASA and ESA instruments for Chandrayaan-1, they adopted PDS standards for their archive. As a result, ESA and NASA worked together to establish the International Planetary Data Alliance (see http://planetarydata.org) in 2006 as a mechanism to develop international standards for planetary science data archiving and encourage international interoperability. Japan's JAXA has accepted PDS standards and is working through the International Planetary Data Alliance to adopt the interoperability protocol developed by ESA/NASA for access to Venus Express data for their Venus mission, Planet C. At the same time, China's CNSA is developing Chang'e-1, and individuals are working to establish archives that will be PDS compatible. Thus, the PDS standards have become the de facto international standards and, although these agencies are receptive to PDS leadership, it is the responsibility of PDS to strive to produce well-defined standards to sustain the efforts to make archives from international missions available in compliant formats to all users.

Improvement of the PDS and international access can yield significant benefits to the planetary program by ensuring that best use is made of data collected in past and ongoing explorations. A dynamic model for data archiving and management within NASA is an essential component in planning our role in the future of space exploration. Only by supporting continued improvement of the PDS can NASA capitalize on data collected in past and ongoing explorations.

5. Expectations of the PDS2010 Project

In coordination with international archiving agencies through the IPDA, PDS will:

- Develop revised, rigorous but simple archiving standards that are consistent, easy to learn, and easy to use;
- Accept a limited number of archive data formats, which will simplify development of data management, conversion, and manipulation;
- Provide adaptable tools to both mission and ground-based data suppliers for designing archives, preparing and validating data, and optimizing delivery to the PDS;.
- Develop a standard interface with the Solar System Exploration R&A and DAP programs that assures that participating scientists who have proposed to deliver data can do so in the most efficient and effective manner;

- Leverage modern web and computing technologies to support the operations as a fully online, distributed, international data system;
- Provide better access allowing users to identify, transform and obtain selected data quickly from anywhere in the system;
- Work with the instrument teams to develop tutorials for data that are intrinsically difficult to use;
- Flag intermediate data that are to be used with caution while instrument teams do ongoing calibration work;
- House and provide access to models that have been developed by the science community and in common use;
- Provide a highly reliable, scalable computing infrastructure that protects data integrity, links Data Nodes into an integrated data system, and provides the best service to both data providers and users for at least the next decade.

6. Responsibilities

Recent attention at NASA Headquarters to the need for more clearly specified proposal and mission requirements and reorganization of the management structure of PDS has set into motion processes, which are leading to considerable progress in achieving the goals above.

There are many stakeholders and each has responsibilities that must be clearly identified and supported by NASA in order to ensure successful data archiving and access. Those stakeholders include NASA Headquarters, the Planetary Data System, Principal Investigators of PI-led missions or Instrument PIs on Flagship or Other Missions, Ground-based Suppliers of Telescopic and Laboratory Data and the Data End User Communities. We have attempted to enumerate these responsibilities in a draft charter for archiving (see http://atmos.nmsu.edu/pub/download/Planetary_Science_Draft_Charter.pdf**).** However, none of these requirements can be reasonably addressed unless NASA Headquarters assigns sufficient priority to the requirements for archiving and funding to allow teams to analyze the data sufficiently to complete calibration and documentation so that all the stakeholders can meet the requirements our preliminary archiving charter defines.

7 Conclusions

Even though PDS received high marks in "Grading NASA's Solar System Exploration Program –A Mid Term Report", Co-chaired by Norine Noonan and Wesley Huntress, it should be noted that this was a progress grade that was based on current Headquarters approaches and PDS progress since reorganization in 2005. If this momentum is to be sustained at a level that will allow the PDS to transform into the online research support facility that will serve the science community to make optimal use of mission data, the Planetary Science Division must continue to stress the importance of end-to-end management of data acquisition, adequate funding for data analysis within the missions and in data analysis programs, and completion and maintenance of PDS2010 to ensure PDS will meet the solar system exploration challenges of the next decade and continue providing improved user services.