Planetary Data Architecture

Lessons from Terrestrial Remote Sensing

*A White Paper to the National Research Council (NRC)
Planetary Science and Astrobiology Decadal Survey 2023-2032*

Danika Wellington¹

Abstract

Moving into the next decade, NASA should develop a strategy for increasing the science return of existing datasets, promoting the development of derived datasets that may find broad usage within the community, and supporting a community-focused environment for data sharing and data assimilation. This strategy should recognize that for some planetary bodies, such as Mars, the state of the field is entering a situation focused less on acquiring isolated data points, but interrelated, synergistic products capable of supporting multi-dataset based modeling, long-term monitoring, and near-term forecasting, if made available and accessible. Terrestrial organizations, particularly satellite remote sensing data providers, are a potentially invaluable source of insight into establishing best practices and standards within the planetary science community.

¹ KBR (Kellogg, Brown & Root). KBR is a government services provider that contracts with federal agencies, including USGS. These views are the author's alone and do not represent those of the company or USGS. Email: danika.wellington@asu.edu

1. Introduction

Terrestrial satellite remote sensing provides an important analog for the planetary research community in terms of mission architecture, data calibration and availability, and utilization of higher-level derived data products. Continuous terrestrial remote sensing observations pre-date the modern-day period of persistent spacecraft at Mars or the Moon, and as such the terrestrial community of data providers and data users has reached a comparatively more advanced state of maturity. Forward-thinking decision-making with regards to sustained mission and data architecture have enabled significant science return from remote sensing data products, including those from long-retired spacecraft missions, such that terrestrial “lessons learned” should serve as important talking points within the planetary community.

While the highest priority scientific questions are obviously quite different between the terrestrial and planetary communities, the instruments, datasets, and data processing challenges posed in deriving accurate information from large-scale multidimensional datasets are overall similar. Due to the relative ease of access to Earth orbit and the increasing involvement of the commercial sector, data volumes from Earth-orbiting satellites, already quite large, are expected to grow exponentially in the coming decade [1]. Data volumes from beyond the Earth-Moon system are modest at present but can be expected to increase significantly in the next decade as well, due to NASA’s ongoing commitment to planetary exploration plus the growing participation of other space agencies. As such, the planetary community can look to the Earth observation counterparts of NASA and USGS, as well as agencies like NOAA and other governmental agencies and commercial companies based around the world, for insight into structuring the future of planetary data architectures.

2. Data Access, Retrieval, and Sharing

Most planetary data users access data in one of two ways: by retrieving individual observation data files from NASA’s Planetary Data System (PDS), or, for team members involved in active planetary missions, through non-public-access servers managed by research centers or academic groups that store embargoed mission data. NASA’s PDS is designed for long-term archiving and preservation of planetary data, often in predominantly raw format with calibration information provided as ancillary files; it is not well-suited to discoverability or ease of use, nor does it readily adapt to new standards or formats. Difficulties in retrieving data from the PDS and transforming it into a meaningful format are a common experience for junior researchers and other members of the community, which limits their ability to integrate multiple datasets or even efficiently find and extract data from a large, single-instrument archive. These issues may be exacerbated by the move to PDS4, which is likely to hybridize rather than standardize the holdings. Furthermore, higher-level data products developed by individual research groups rarely or belatedly find their way to the PDS, likely due to the significant barriers to submission, limited availability of funding to document and restructure data products for PDS standards, and intervening professional priorities. Some of these issues, and the divergence between the PDS’s scope and funding versus user expectations, are captured in the Planetary Data System Roadmap Study for 2017–2026 [2].

In addition to being available sometimes in only raw format, such planetary data is also naturally therefore in raw sensor space. Efficiently performing spatial queries and retrieving spatial subsets of data products thus becomes difficult, as an orbital imaging observation's spatial information may be as simple as the center longitude of the image swath. Terrestrial data are typically provided in a gridded tiling scheme, defined in a particular spatial projection. This allows users to extract portions of images for analysis, rather than the entire imaging observation, which likely contains significantly more than the user area of interest. Recently, the long-running NASA/USGS Landsat satellite series decided to move portions of their product suites into a cloud architecture as Cloud-Optimized Geotiffs (COGs) [3]. These products support HTTP range requests to efficiently locate and retrieve subsets of raster image files, such that users do not have to download large amounts of data only to clip images to their particular study regions, saving bandwidth, filesystem space, and memory. Depending on which products, over what time scales, the user is attempting to query, efficient access to data products can provide anything from a simple convenience to making large-scale data ingest computationally feasible.

All of this highlights the fact that the planetary science community does not have a widely adopted, user-centric and community-supported spatial data infrastructure to support geospatial analyses. An increasingly large data volume returned from bodies with a long-term spacecraft presence (Mars and the Moon) would be expected to increase the need and desire for efficient data discovery and retrieval to support processing such as is already evident within major terrestrial projects. This lack of infrastructure is highlighted by other white papers and publications [4, 5, 6], and is a major aspect of the MAPSIT roadmap [7]. Some institutions and agencies maintain tools that do permit some spatial querying, although these function primarily as search capabilities on available PDS holdings. For example, the PDS Geosciences Node maintains a web form and API (the Orbital Data Explorer) to search for data products by latitude/longitude bounding boxes and product metadata [8]. Some datasets are not discoverable in this way; for instance, MRO/MARCI images are not accessible and cannot be efficiently retrieved for processing by any means.

Furthermore, data sharing among researchers in planetary science, when desired or required by the funding source, often necessitates identifying a repository for derived data. For data that are meant to be easily accessible and usable by others in the field, NASA's PDS is probably not the best option -- the effort required to transform data into an archival format may be overly burdensome for small-scale research projects. Separating archival from community products may allow more flexibility by users and preserve the PDS's primary purpose as safeguards of raw mission datasets. At present, data repositories for planetary datasets outside the PDS are limited to only a few options, including repositories that are hosted by journal publishing companies that may have a dubious commitment to long-term open access.

3. Data Usability

Terrestrial satellite imaging providers, and the Landsat program in particular, have popularized the term "analysis-ready data" to refer to data that have been pre-processed such that they can be immediately ingested by user analysis pipelines [9]. Landsat ARD include derived data products, such as surface reflectance and surface temperature, as well as pixel

quality bands that indicate the presence of clouds or cloud shadows. These data support higher-level mapping efforts and are highly valued by those communities. Users do not need to individually carry out these pre-processing steps themselves, allowing them more time to focus on their own research questions.

By contrast, planetary data archives typically take the approach of providing the minimum information such that a researcher could make use of instrument response coefficients, published sensor bandpasses, and available SPICE kernels, and generate their own derived products. There are reasons why users might prefer to do these steps themselves -- perhaps they would like to implement their own calibration adjustments, handle topographic or photometric corrections in a particular way, or otherwise take ownership of particular portions of the data pre-processing. Furthermore, with respect to orbital data in particular, the data provider may not be able to predict the user's preferred map-projection, which will likely vary depending on the latitude region and other specifics of the use-case, and re-projection would result in a loss of fidelity. In reality, though, pre-processing and map-projection are a significant burden on users, and result in substantial duplication of effort within the field. As data volumes increase, users interested in running an analysis across a large dataset may face significant computational challenges if they must also pre-process all input data.

4. Open Data Policies

NASA's Earth Science data and information policy states that "There will be no period of exclusive access to NASA Earth science data" [10]. This is in stark contrast to most planetary mission datasets, in which science data are generally subject to an embargo period of several months or more. While this embargo period is intended to allow researchers who have expended significant amounts of professional time in the pre-planning, calibration, and mission operations work the rights of first publication on high-value science data, it also has the effect of limiting more widespread involvement in active mission data analysis and delaying potentially important scientific results. It even inhibits participation by interested members of the general public, who may (e.g., for rover missions) have access only to JPEG images that have often been further compressed for the web.

While these delays may not necessarily negatively affect the results than can be inferred from the observations themselves, there are at least two potential scenarios in which such embargos can have serious consequences: 1) too slow recognition of important underlying features in the data that could have been subject to follow-on science observations, and 2) a severe hindrance in the ability of the community to react to present-day, real-time dynamics in the martian climate system. With regards to (1), while the time-pressure for follow-on observations is comparatively small for orbital observations, it is inevitably quite large for rover missions. Embargoing these data products effectively severs the wider planetary community from providing feedback or developing rapid analysis tools that could aid overall science return. As an example of (2), the occurrence of the 2018 martian global dust storm prompted a large interest in near-real-time MARCI imaging data, which was still under embargo; thus, interested community members were dependent on the instrument team to assemble and distribute information on the storm's evolution.

5. Software Development and Availability

Processing and analysis pipelines depend on underlying software routines that may or may not be written, maintained, documented, and placed under source-control in alignment with generally understood software best-practices. Unfortunately, academic research groups rarely have the funding to support development teams, and thus often have little ability to consistently maintain important analysis tools and pipelines. This also disincentivizes code sharing, since this same lack of support limits the ability of researchers to respond to queries and technical issues that may arise from potential users. In the worst case scenario, a high profile instance of code sharing by researchers (in a different field of study) that was not in line with outside expectations led to negative reactions by the general public and loss of trust in the results [11].

Thus, while it may be of little utility for research groups to attempt an unstructured release of ad-hoc analysis code, particularly valuable processing or modeling software should be prioritized for cleanup and made openly available. Some research groups maintain open source analysis and visualization tools for the community, although the longevity of these tools against future funding shortfalls must be a concern, and the community needs mechanisms to provide sustained investment in software development. Finally, the persistent widespread use of proprietary programming languages and software within the community (such as IDL/ENVI and ArcGIS) limits the use of tools, extensions, and workflows to only those users with access to the appropriate license. Further funding for educational workshops may encourage groups to migrate away from these, although this will necessarily require some additional investment of time and resources.

6. Data Harmonization and Assimilation

In the terrestrial remote sensing community, the strong interest in dynamic processes places a high importance on both dense temporal sampling and long time series observations that together can capture both rapid-scale surface changes (such as logging, wildfires, and disaster areas) and long-term, multidecadal variations (such as climate change, urbanization, and patterns of natural resource use). As such, not only have consistent decisions been made with respect to bandpasses and spatial resolution of long-running Earth observation missions (such as the NASA/USGS Landsat satellites), a high priority is placed on data harmonization -- the creation of derived products from different instrument sensors that can be treated as one larger dataset. Landsat data products have been harmonized across multiple generations of Landsat imaging sensors [12] as well as with Sentinel-2 data operated by ESA [13]. Harmonization efforts, together with accurate geospatial registration, may be especially valuable to studies of long-timescale surface dynamics, which Mars is known to experience in its distribution of surface dust [14]. Furthermore, future orbital mission planning can and should incorporate the potential for one instrument's data set to effectively extend the temporal range of another, preferably with some overlap to allow accurate cross-calibration. Regularly produced data products with accurate geospatial registration and inter-instrument harmonization can provide a useful data source for assimilation into martian climate models, and provide

up-to-date information on the present state of the surface and atmosphere as weather satellite data do for forecasting on Earth.

7. Data Interoperability

There are two ways that data interoperability may be considered: in one sense, that similar data products are produced with similar metadata layout and format such that any process wishing to ingest both does not need to invoke special handling for each individual dataset, and secondly, that data products work easily and naturally across a range of higher-level software tools, including GUI applications. The PDS has enabled international cooperation in planetary data formats through the International Planetary Data Alliance (IPDA). However, the PDS archival format is generally not natively understood by common software tools (without special extensions). Browse products are, but these products may be either lossily compressed or subject to a reduction of bit depth with respect to the original data file, and not suitable for analysis. Generating such products (i.e., in commonly used formats) may be viable within a well-supported user community, although probably outside the scope of the PDS's duties.

8. Data Calibration

Planetary mission data is generally first provided to the PDS in a raw, uncalibrated format. Depending on the mission and instrument, additional calibration information or calibrated data files may also be supplied, although often these initial calibration data are superseded by future work that may develop more accurate models of the instrument in-flight response or derived products based on improved atmospheric corrections. However, these products are often significantly delayed from the data acquisition date. The highly anticipated release of CRISM MTRDR products is one such example; pipeline development for these products began in 2012 [15], although the first release was not until 2016. MER Pancam and MSL Mastcam in-flight calibrated I/F products are another example; an initial release of the MER Pancam I/F products on the PDS is dated to late 2014, with a second update only recently in 2020. MSL Mastcam data, by comparison, have not yet been released to the highest fidelity in-flight calibration state, even eight years after the beginning of the surface mission (although such work is in progress). Production of higher level products is often dependent on the highly specific technical expertise of mission science team members and their ability to carve out sufficient time and resources to devote to public release products, leading to significant delays.

Earth science data, by contrast, are valued for their timeliness, otherwise (except for long term studies) they may become obsolete with respect to the needs of policy makers and other users dependent on up-to-date information. NASA's Earth Observing System Data and Information System (EOSDIS) comprises Distributed Active Archive Centers (DAACs) that process, archive, and distribute data, making it available soon after acquisition. In addition, USGS/EROS hosts the EROS Cal/Val Center of Excellence (ECCOE) that continually researches and improves upon the calibration accuracy of Landsat data products. This establishes in-house expertise and continuity for sensor calibration, and provides support for continual improvement and harmonization. Such an investment by the planetary community

would not only provide fantastic scientific benefits, it would also enable an avenue of support for early-career scientists who may be uninterested in academic pathways.

9. Earth Science: A Case Example

A recent project by the USGS/EROS center was effectively made possible by quality data architecture and a multidecadal investment in harmonized data acquisition. The USGS LCMAP (Land Change Monitoring, Assessment, and Projection) initiative recently released a product suite that provides multispectral surface reflectance modeling, break detection, and classification across the conterminous United States, at Landsat 30x30m scale (~8 billion pixels). This project was able to treat thirty-three years of the Landsat archive (assembled across multiple generations of sensors) as a radiometrically seamless, spatially co-registered and gridded dataset. This was made possible by Landsat ARD, considered by [16] to be “foundational” to the project.

10. Mars Science: An Example of Putative Benefits

Mars is a dynamic planet under continual orbital observation. Thus, the state of its surface dust cover and atmosphere is constantly refreshed with new observations, although these data cannot be immediately accessed by the community. The ability of the atmosphere to spawn regional and global-scale storms is of interest not only scientifically, but operationally as well, particularly for surface missions. Theoretically, research groups and mission teams could constantly pull newer data products not only to build their own early-warning systems for atmospheric opacity increases, but also to adjust targeted orbital observations to take advantage of surface dust movement that may provide a new window for spectral observations of the underlying bedrock. Furthermore, much the same way that Earth-observing weather satellites provide forecasting on Earth, models able to retrieve timely information from orbital datasets could attempt to evaluate risks from surface dust reservoir replenishment and forecast the evolution of nascent storm systems. These use cases require the data to be quickly accessible and easily usable.

References

- [1] NASA, “Earthdata Cloud Evolution” NASA Earth Observing System Data and Information System (EOSDIS) website, accessed Sept-2020.
<https://earthdata.nasa.gov/eosdis/cloud-evolution>
- [2] NASA (2017) “Planetary Data System Roadmap Study for 2017-2026”, available online at https://pds.nasa.gov/home/about/PlanetaryDataSystemRMS17-26_20jun17.pdf
- [3] USGS (2020) “Landsat Data Moving to Public Cloud in Early 2020”, accessed Sept-2020.
<https://www.usgs.gov/news/landsat-data-moving-public-cloud-early-2020>.
- [4] Radebaugh, J. et al. (2020) “Maximizing the value of solar system data through Planetary Spatial Data Infrastructures”. Planetary Science and Astrobiology Decadal Survey 2023-2032, white paper.

- [5] Williams, D. et al. (2020) “RPIFs to PDUCs: New Planetary Data Utilization Centers to support NASA’s Planetary Data Ecosystem”. Planetary Science and Astrobiology Decadal Survey 2023-2032, white paper.
- [6] Laura, J. et al. (2017) “Towards a Planetary Spatial Data Infrastructure.” ISPRS Int. J. Geo-Inf. 6(6), 181; <https://doi.org/10.3390/ijgi6060181>.
- [7] Mapping and Planetary Spatial Infrastructure Team (2019) “Mapping and Planetary Spatial Data Infrastructure Roadmap 2019-2023”, accessed Sept-2020.
<https://www.lpi.usra.edu/mapsit/roadmap/>
- [8] Wang et al. (2017) “Accessing planetary data using PDS Geosciences Node’s Orbital Data Explorer.” The 3rd Planetary Data Workshop, LPI Contribution No. 1986, 7026.
- [9] Dwyer et al. (2018) “Analysis Ready Data: Enabling Analysis of the Landsat Archive” Remote Sens. 2018, 10(9), 1363; <https://doi.org/10.3390/rs10091363>
- [10] NASA EOSDIS Data and Information Policy. Available online at <https://earthdata.nasa.gov/collaborate/open-data-services-and-software/data-information-policy>
- [11] Chawla, Dalmeet S. (2020) “Critiqued coronavirus simulation gets thumbs up from code-checking efforts.” Nature 582, 323-324. doi: 10.1038/d41586-020-01685-y
- [12] USGS, “Harmonizing the Landsat Archive”, available at <https://www.usgs.gov/core-science-systems/nli/landsat/harmonizing-landsat-archive>
- [13] Claverie, M., et al. (2018). “The Harmonized Landsat and Sentinel-2 surface reflectance data set.” *Remote Sensing of Environment*, 219, 145-161.
- [14] Newman, C. (2020) “Toward More Realistic Simulation and Prediction of Dust Storms on Mars”. Planetary Science and Astrobiology Decadal Survey 2023-2032, white paper.
- [15] Seelos et al. (2012) “CRISM Map Projected Targeted Reduced Data Records (MTRDRs) – High Level Analysis and Visualization Data Products” Planetary Data Workshop, June 2012.
- [16] Brown, J. et al. (2020) “Lessons learned implementing an operational continuous United States national land change monitoring capability: The Land Change Monitoring, Assessment, and Projection (LCMAP) approach”. *Remote Sensing of Environment*, 238 (111356) <https://doi.org/10.1016/j.rse.2019.111356>.