New and Old Data: Archiving with PDS – One Philosophy

Lynn D.V. Neakrase, Kathrine Sweebe, Lyle Huber, Zena Stevenson, Nancy Chanover, Joni J. Johnson, Reta Beebe
NASA PDS Atmospheres Node, Department of Astronomy, New Mexico State University, Las Cruces, NM 88003-8001

Introduction. The Research and Analysis (R&A) programs within NASA's Planetary Science Division now require archiving of resultant data with the Planetary Data System (PDS) or an equivalent archive [1]. With the advent of the PDS4 archiving standard, data providers have more flexibility than ever for archiving current data, migrating past data, and co-developing with the PDS the type of archive providers and users want. The PDS Atmospheres Node (ATM) is developing an online environment for assisting data providers with this task. Streamlining the process for both new data providers and migrators of old data shows that the processes are parallel and are components of the same design process. The new online environment will inform potential users to manage and produce PDS4 Bundles. More than a simple PDS4 label editor, the Educational Labeling System for Atmospheres (ELSA) will assist providers in designing their archive by walking them through the process of creating label templates used in generating PDS4 archive bundles for their data. Integration of the Python/Django code for creating new PDS4 bundles with existing Python code for internal migration of PDS3 to PDS4 provides a unified, parallel system for handling metadata and creating the bundle structures needed for a good PDS4 archive.

One Philosophy. Common to both migration and creation of new PDS4 products is the need for consistent, unique identifiers used for cross-referencing within the PDS4 system. *Logical Identifiers* (LIDs) are constructed via Uniform Resource Names (URNs), which are heavily dependent on set patterns for effective and unique identification of various parts of a PDS4 bundle.

urn:nasa:pds:<bundle>:<collection>:<product>

Figure 1. An example of the format of the URN used in PDS4 internal references. The hierarchical structure of the PDS4 standard is evident in the identifiers for the bundle, collection, and subsequent products (files).

ATM has been working with data providers of various R&A programs to assess the best way to design PDS4 LIDs for their various projects. Over the past two years, ATM has moved to a '*top-down*' approach of working with data providers. Data providers that were encouraged to start with designing the patterns for their LIDs ultimately produced more complete and usable PDS4 archive products with less assistance from node personnel, suggesting a better understanding of the archiving task required by their program award.

Along with internal references for the Bundle/Collection/Product structure within PDS4, URNs are also used to reference the Context Products, or rather, the links to the descriptions of spacecraft, facilities, instruments, and targets. Context products are 'official' listings of information about the missions/spacecraft, their instruments and targets, and/or supporting lab/field/observatory facility work and these are officially managed by the PDS Engineering Node. Because these are managed by a centralized organization, they have a particular format that can be difficult to use without PDS help.

Most of the perceived difficulties with the transition to PDS4 for data providers can be alleviated with adequate work with node personnel. However, scheduling of these interactions can be difficult for both the data providers and the node. To compensate for this, the Atmospheres Node is considering ways to streamline the interactions with data providers by providing them ready access to help in setting up their PDS4 archives. Similarly, internal PDS pressures to migrate PDS3 holdings to PDS4 have independently driven efforts to produce more automated ways of creating PDS4 archive products. Parallel development of nearly identical processes for new data providers and migration of old data have led to development of an integrated approach to PDS4 archive creation at ATM.

Practical Applications. Development of practical online support tools is one of the requirements of the PDS, and with the release of the PDS4 Standard this requirement is especially important for encouraging widespread usage of the new system. ATM has been developing an online environment for boosting efficiency and accuracy of our interactions with data providers and users since the completion of the first PDS4-native missions (LADEE and MAVEN*; **Still actively producing data*). PDS3-PDS4 migration testing at ATM led to considering the 'whole' or 'complete' effort for data providers with the goal of producing complete PDS4 bundles with PDS4 labels for all parts of the bundle. The internal referencing thus led to the top-down approach. As stated above, the top-down approach was also independently developed resulting from interactions with R&A archiving projects.

The Educational Labeling System for Atmospheres (ELSA) was initiated and sought to streamline the creation of internal referencing and selection of context referencing needed for PDS4 bundle creation. Automation was key and ELSA development aimed to help data providers learn PDS4 while automating the more difficult and important aspects of the internal PDS4 referencing system. The end product of ELSA was to create usable, complete PDS4 templates that data providers could use to construct their bundles for submission to ATM with significantly less time spent on scheduling phone conversations or exchanging emails with node personnel. The goal was to take many of the common questions and common setup issues for new PDS4 providers and automate the interface. Automation will allow better communication with ATM in the long term by removing common stumbling blocks of the initial setup of PDS4 label templates.

Currently ATM is hosting the ELSA tool publicly for beta testing of simple cases. (https://atmos.nmsu.edu/elsa/) ELSA allows users to create an account with ATM and to begin creating bundles automatically with an easy-to-use online interface. ELSA walks users through a series of questions initially to set up the patterns that will be used for the internal referencing for the bundle. ELSA also prompts users to select which collections they will be using, and ultimately the types of products within those collections. Once initial bundle setup is complete, users have the ability to add context references to their bundles via pull-down selectors for facilities and missions and their associated instruments and targets. ELSA then populates the label template stubs with all the references that have been designed or selected. Users can then select the types of products for their collections and do some limited editing/setup for those templates. Currently ELSA works with document files and simple delimited tables as a proof-of-concept step in the tool development. At any time during this process ELSA allows users to see the labels and download them for use or further offline development. The intent of developing ELSA is to provide help where it is needed and allow easy access to the finished (or partially finished) templates for further communication with the node or offline pipelining for label creation.

Future Complexity. Currently ELSA is being tested for basic functionality, but ATM will continue to add capabilities. Current prioritization is driven by usefulness to the planetary atmospheres community. ATM recognizes how ELSA may be useful to the more general PDS community, hence a modular approach to ELSA's development through Python and Django online implementation. ATM is in the process of designing an ELSA interface for our PDS3 migration tool with the hopes of using the infrastructure we've designed for new data to be useful for easier migration of our older data. The modular design of the code for ELSA should facilitate easy expansion to other common data formats for the atmospheres community and in turn for PDS in general for more complex data types. For example, common to astronomy and planetary atmosphere studies is the FITS data format which essentially consists of tables and arrays/images linked together in a single file separated by headers. The approach to handling this sort of complex product with ELSA would be to use pieces designed for the component parts building on pieces that already work for the 'simple' products. This lends itself to expansion through layering functionality within ELSA's framework.

ATM plans to continue to develop ELSA with more functionality, working closely with data providers to produce a useful tool to streamline the process of submitting data to the PDS. The new PDS4 standard lends itself particularly well to web automation through the use of the centralized Information Model that governs the rules behind PDS4 and the implementation of the labels through a common XML format. Automation of certain parts of the archiving process allows consistency and accuracy of the internal referencing but provides flexibility to handle a wide variety of products. The potential of a system such as ELSA to provide expandability over time to more data products beyond the usage at ATM is great and could lend itself to better integration across all of PDS.

References: [1] NASA ROSES-2017 https://nspires.nasaprs.com/external/solicitations/summary.do?method=init&solId=%7BE757EF32-60E6-76AE-A276-21A1F8BA96BB%7D&path=open; [2] Stevenson et al. (2017), *Proc. Fall DPS*, Abs. #218.03; [3] Neakrase et al., *Proc. Fall 2017 DPS*, Abs. #218.04.