Why can't I leave PDS archiving until the end of my funding?

It is vital for potential DAP and PDART archivers to realize that allocating a specific amount of effort (e.g., six weeks) to archiving starting in the last quarter of funding is far from an ideal approach to producing archivable products by end-of-funding; conversely, the same amount of effort distributed over three years is highly likely to succeed. There are practical reasons for this, noted below.

1. Planning makes archiving easier.

PDS4 is a metadata-rich system, and the PDS4 data structure set is very constrained[1]. As future archivers first consider their data development process, the following questions should be part of the process:

A. Is my data structure PDS4-compliant? If not, how will I convert it?

Conversion programs may already exist, or the archive developer may have to write them. If it is not clear whether the data structures being considered are compliant with the PDS4 constraints, an early consultation with the PDS node ultimately receiving the data will provide much useful information. Having a data sample available for this discussion, even if it's just mock-up or sample data, can be a great aide to understanding.

B. What metadata do I need to supply for each data product?

Not only is there prescribed PDS4 metadata to consider, there are also specific discipline metadata requirements for archiving images, spectra, observational geometry, etc. A sample file and telecon with the PDS node will quickly identify additional metadata PDS will require for archiving.

Metadata specific to a particular project will have to be defined by the archiver in a local data dictionary. PDS has tools, training, and examples available to assist those creating local dictionaries.

Collecting all label metadata in a database or spreadsheet where label writers (human or machine) can access it will make generating labels much easier when the time comes.

[1] For purposes of this discussion, *metadata* is everything in a PDS4 label, *data structure* is the physical format of the data in the data file.

N.B.: The level of effort required to archive doesn't scale with the number of products or size of the data to be archived in any meaningful way. It more closely scales with the complexity of the metadata that must be used and defined, and to a lesser degree the amount of additional documentation required to support the data. Starting design of these elements of the archive early on is critical to keeping archive development on track.

2. Frequent checking makes for small corrections rather than large perturbations.

There is both method and technique involved in designing good PDS4 labels. Starting small with basic metadata and periodically providing sample data and early draft labels to the PDS node for feedback creates opportunities for the new archiver to develop skills and find the tools that work early in the data development process. It also generates opportunities for the PDS node to offer examples or solutions that have worked for similar data.

3. PDS will be scrupulous about standards validation prior to review.

External peer reviewers cannot review data that they cannot read and visualize with standard PDS4 tools, so standards verification is an absolute requirement for data going out to review. The canonical validator that PDS uses to verify standards compliance can be installed on the data preparer's local system and incorporated into the production process to ensure that the labels are complete and syntactically correct. PDS will happily assist with installation and training for these tools. PDS will reject data delivered for review that contains errors that would prevent a reviewer from being able to review the data.

4. PDS Nodes review schedules are generally fixed half a year or more in advance.

N.B.: *A Principal Investigator's archiving obligations are not met until PDS officially accepts the lien-resolved data for archiving.*

All archive data must pass an external peer review before it can be accepted for archiving, and PDS nodes cannot schedule external peer reviews on demand. The main reasons for this are:

1. The primary mandate from NASA to PDS is to archive NASA mission data, thus it is mission schedules that drive the main peer review schedule at any given node. PDS is also keenly aware of its obligations to its discipline

communities, and keeps a close eye on data posting deadlines to insure that datasets are publicly available in time to be eligible for the appropriate DAP and PDART proposals when scheduling reviews.

2. Running an external peer review requires identifying at least two reviewers for each data set, getting the data to them, allowing time for reviewers to analyze the data, meeting to discuss the issues raised, and development of the "liens list" of both science and standards-compliance liens that must be addressed (i.e., a new version of the data set without these issues must be generated and validated) prior to archiving. It is time-consuming, calendar-consuming, and expensive. PDS generally reviews clusters of related data in reviews scheduled months in advance, as often as four times a year at some nodes.

3. When a review is scheduled, the time span from delivery of the review-ready dataset through external review, lien resolution, and acceptance of the final, edited data for archiving, is typically 2-3 calendar months.

N.B.: The most time-consuming part of lien resolution is nearly always editing or creating documentation or metadata as required by the reviewers. Rerunning label generating code is typically the least time-consuming part of the lien-resolution process

Consequently, an archiver cannot ask for an external peer review and expect to have a liens list in hand two weeks later. Neither can PDS present a liens list to the archiver and expect the final delivery within a week. Most likely, the candidate archive data will be added to a review already scheduled for similar or related data; and PDS will work with the archiver to set a final delivery schedule that is considered achievable by the archiver. Waiting until the last half-year of funding before discussing review scheduling with the node may well result in no review slot being available until after funding ends.

-Anne Raugh, Small Bodies Node, with thanks to Sheridan Ackiss.