Proposing a PDS4 data archive for the NASA PDS Imaging Node
June 25, 2015

So you've decided to archive data as part of your proposal to a NASA research program. The information presented here will help you to plan the preparation of your proposed PDS4-compliant data archive. PDS4 is the current PDS archive implementation, and it is an integrated system designed to improve access to PDS data across the nodes (see https://pds.nasa.gov/pds4/about). PDS4 uses the Extensible Markup Language (XML) to create product labels, and templates for a wide variety of product types have been developed to ensure adherence to PDS standards on product formatting and use of keywords. There is extensive information for data providers available at the PDS Home page (see https://pds.nasa.gov/pds4/about/portal.shtml). Before you prepare a data archive plan, please be sure to review the information in the PDS4 Concepts Document and the (PDS4) Data Providers Handbook. This document echoes and supplements that information and will help you to estimate the effort (and cost) of preparing your archive.

First, we outline the current basic requirements for preparing a PDS4-compliant archive. Note that a PDS4 archive has an encompassing structure in which the file hierarchy is uniform across all of the PDS. Basic elements of this structure are listed below.

- Data are organized into Bundles
  - Bundles consist of similar project data
    - Mission Instrument Bundles include all data for a particular instrument, OR all lab data from a single lab or lab project, OR field data from a single field campaign, etc.
- Bundles have Collections
  - Collections are grouped files all having a similar origin
    - Document Collection – Contains all documentation necessary for understanding and using the data
    - Data Collection(s) – Contains all of the data, logically grouped and could consist of multiple directories
    - Calibration Collection – Could contain calibration data separate from the data collection(s) and/or more detailed information about complex calibration efforts
- Everything is a Product and Products belong in Collections
  - All data are considered products and should have PDS4 XML metadata labels
  - Products are not limited to only data
    - All documents, context files, XML schema, and even delivery websites are considered products and will have labels.
    - Labels require logical identifier strings that ensure registration in the PDS4 system
  - Data providers are responsible for preparing the full archive with help from the Imaging Node.

You can find examples of PDS4-compliant products (comprised of four basic data structures: arrays, tables, parsable byte streams, and encoded byte streams), collections, bundles and packages that illustrate the design concepts and goals of the PDS4 system. Look for these examples on the PDS Home page under PDS4, Information for Providers (see https://pds.nasa.gov/pds4/doc/examples/). If you don't find examples that you think are relevant to your proposed archive, ask us for help. Also, feel free to use the available PDS4 software, including ***Generate***, ***Validate*** and ***Transform*** tools as well as the PDS4-specific tools (see https://pds.jpl.nasa.gov/pds4/software/index.shtml). (Note that these PDS programs were developed under Oracle Java (currently version 1.6), and they are intended to run on any platform where Java is supported.) Finally, you may also need to refer to the PDS4 Schema (see https://pds.jpl.nasa.gov/pds4/schema/index.shtml) to understand how your archive will fit into the PDS4 system.

- Contacts in the PDS Imaging Node (PDS-IMG) for help with PDS4 archiving:
  Lisa Gaddis, Imaging Node Principal Investigator and Science Lead, Email: lgaddis@usgs.gov, Phone: (928) 556-7053
  Sue LaVoie, Imaging Node Co-Investigator and Technical Lead, Email: slavoie@jpl.nasa.gov, Phone: (818) 354-5677
  Chris Isbell, Imaging Node Data Manager/Data Curator and PDS4 Lead, Email: cisbell@usgs.gov, Phone: (928) 556-7211
  Jordan Padams, Imaging Node Data Manager/Data Curator and PDS4 Programmer, Email: Jordan.H.Padams@jpl.nasa.gov, Phone: (818) 354-3130

We recognize that it takes some effort to learn how to produce XML labels and conform to PDS4 standards, and we will help you with this. Please feel free to communicate with us as often as needed to ensure that you approach this in an efficient manner. Let us develop templates for labels and help you organize the bundle structure and identify the components for your particular data set. Also, please note that archiving with PDS is not just a matter of providing data to the appropriate node. All data providers are required to see the process through to the end including allotting enough time for iterative work on the label creation process throughout, and including enough time for the peer review and lien resolution at the end of the grant period.

To plan your archive in a proposal to NASA, you should be sure to discuss these elements of the required proposal archive plan and include time estimates for each element:

- Be aware of all of the tasks involved in creating an archive:
  - Summarize the products you intend to archive, including the data products, documentation, ancillary information (e.g., information on how the data were obtained, processed, calibrated---include anything that a user would need to know to use your data as a scientific product), etc.

        - Be sure to characterize the full scope of the archive, including all versions of products (i.e., whether you intend to include interim products, multiple mission phases, etc.) to be included.
        - Look for representative examples of PDS4 products in PDS to serve as examples of your products and archive, and if you don't find any, ask PDS-IMG personnel.
    - Outline the design and directory layout of your complete archive, including the bundle(s) and collection(s) as well as any needed browse images.
    - Describe the basic contents of the labels for each of these
        - If you want to really wow the reviewers, prepare preliminary XML labels for your products
- Summarize the full expected volume and nature of products in your archive
    - Include volume information for all archive components, including images, documents, labels, etc.
    - Describe the complexity of the archive, including the number and nature of the components, file and image formats, any unusual aspects, etc.
- Describe the processing you plan to do to update, revise, reformat or otherwise restore the products you will archive
    - Explain whether the processing is complicated or simple, and discuss how long will it take you to develop procedures and/or scripts to do the processing. (If procedures and scripts already exist, be sure to mention that.)
    - Describe the computing resources to do the work proposed, including data storage space, processing capacity and speed, etc. and explain whether you have them or have a plan to get them.
- Describe your plan to work with PDS-IMG to ensure PDS4-compliance for your products
    - Expect to communicate often and to iterate with us on archive design and product formats, labels, etc.
    - Expect to validate your archive against existing PDS4 products and using PDS-supplied software
    - Be prepared to provide sample data for and to participate in a peer review of your archive, with the support of PDS-IMG personnel
    - Be prepared to follow the archiving process through to the end in partnership with PDS-IMG and to commit to resolving any liens noted during peer review
- Describe a schedule for development, testing, validation and delivery of your products, and be sure to include time required for participating in all steps.

Note that if you intend to submit new (possibly higher-order) products or reprocessed data already in the PDS, you will need to do some additional tasks to ensure your new products will be PDS4 compliant. For example, some of the linking documents and mission documentation from a PDS3 dataset might require collation into a PDS4 document collection under the original mission. PDS-IMG will do most of the work related to the creation of PDS4 linkages that your project may require,

although you, as the data provider, will still be responsible for PDS4 compliant structures within your own project.

Our experience indicates that preparation of a simple archive takes one person between 1 and 3 months (i.e., up to3 months for data providers not familiar with archiving and PDS4 requirements).

We understand that not all data providers have strong backgrounds in XML or PDS organization. PDS-IMG staff will commit to helping you with these tasks to help ensure PDS4-compliance and successful delivery to PDS. Once you are selected for funding by NASA to create a PDS4-compliant archive, the interaction between you, the data provider, and PDS-IMG personnel should be something like this:

- **Data Provider** contacts **PDS-IMG** about selection and starting a new archive
- **PDS-IMG** acknowledges this new project and requests sample data products
- **Data Provider** provides information and sample products
- **PDS-IMG** provides XML Label Templates for all parts of the archive
- **Data Provider** and **PDS-IMG** iterate this process to include all appropriate metadata fields are in labels
- **Data Provider** produces documentation describing the data, including any calibration and further processing (e.g., higher level derived data) and any references as to where the data came from
- **Data Provider** (with **PDS-IMG** help) produces all other files within the bundle structure including
    - Bundle file – the label for the bundle that describes the contents of the entire bundle and links the entire bundle to the PDS4 registry system
    - Collection files and inventories (depending on the number of collections) – the label and inventory list of the collection's contents (manifest)
- **Data Provider** and **PDS-IMG** validate all Bundle contents
- **PDS-IMG** organizes and holds a peer review of the completed bundle
- **Data Provider** responds to this review and resolves any identified liens on the archive products
- **PDS-IMG** hosts the data on the web and declares the data certified (capable of being used in future proposals) when liens are successfully resolved

PDS archiving efforts require that all data submissions from data providers be peer-reviewed before being designated as certified data. Certified Data have the added benefit of being recognized for use in future proposal efforts across planetary science. Data can be hosted on PDS websites before certification is complete but cannot be used in future proposals until they have attained Certified status. Before the archive bundle goes out for review, PDS-IMG will be responsible for finalizing the bundle with other required contents to complete the effort, which will include:

- **Context Collection** – Contains references for the PDS4 system and includes official system references for instruments, targets, spacecraft, etc.
- **XML_Schema Collection** – Contains references or mirrored copies of XML schemas (blueprints) for the labels used in this bundle (includes references to the version of the system model and local data dictionaries, etc.) used for system continuity and traceability