Preparing a PDS4 data archive for the NASA Planetary Data System
Cartography and Imaging Sciences ("Imaging") Node

Congratulations on being selected for funding by a NASA research program! You've successfully proposed a data archiving plan and you've been selected to proceed with your work plan, including archiving data in the NASA Planetary Data System. As part of the NASA proposal process and following selection for funding, you have committed to fully supporting the development and validation, delivery, review and lien resolution for your archive. As you likely already know, your data archive must be compliant with the current archive implementation of PDS, the so-called PDS4. We in the PDS Imaging Node (IMG) will help you with preparing your archive, including helping to draft labels, reviewing your archive plan, validating your delivery, organizing a peer review of your archive, and ultimately accepting your finalized data into the PDS. As soon as you are selected, we will assign a Point-of-Contact (POC) with whom you'll primarily work. Please feel free to communicate with that POC and others in IMG as often as needed to ensure that you approach archiving efficiently.

PDS4 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 XML classes and related attributes. 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 an archive, 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 get you started on the typical workflow of developing an archive, planning and scheduling archive activities, and estimating the effort and cost of preparing, reviewing and certifying your archive prior to acceptance into PDS. ***IMG will work with you every step of the way!***

Archiving Workflow

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 workflow between you (the data provider) and the PDS-IMG POC 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 such that all appropriate metadata fields are included in labels

- **Data Provider** produces documentation describing the data including any calibration and further processing (e.g., higher level derived data) and any references regarding data source and origins
- **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 each collection's content (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 bundle content (archive products, documentation, etc.)
- **PDS-IMG** hosts the data on the web and declares the data certified (capable of being used in future proposals) once liens are successfully resolved

Archive Planning

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***. Your Point-of-Contact (POC) will communicate with you often and will bring in other PDS experts as needed. As outlined in the workflow below, we will develop templates for labels and help you organize the structure of your archive and identify the components for your particular data set. We 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 with your POC 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 to be archived, including 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 include the full scope of the archive, including all versions of products (e.g., any interim products, data from multiple mission phases, etc.).
    - 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 and work with IMG to prepare preliminary XML labels for your products
- Summarize the full file size (digital storage requirements) and nature of products in your archive

  - Include data size 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.
- Develop software and scripts for any processing you expect to do to update, revise, reformat or otherwise restore the products you will archive
  - Assess whether the required processing is complicated or simple, and discuss time required for you to develop procedures and/or scripts to do the processing.
  - Assess the computing resources to do the work proposed, including data storage space, processing capacity and speed, etc. and ensure that they are available for the archiving work.
- 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, 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 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.

Estimation of Effort and Cost

Even if you are a seasoned PDS veteran, it's unlikely that you can simply deliver an archive-ready package to PDS at the end of your performance period. Your POC will help you define appropriate data structures, design PDS labels, design the components of the archive (data, documentation, index tables, supporting materials), and prepare for the peer review. Discuss the scope of the work needed to archive your data with your POC to accurately estimate the effort required in your proposal. Our experience indicates that preparation of a simple archive takes one person between 1 and 3 months (i.e., up to 3 months for data providers not familiar with archiving and PDS4 requirements).

Basic Elements of a PDS4 Archive

Once you've planned your archive development with the IMG POC, you're ready to begin creating the PDS4 archive as part of the workflow described above. Outlined here are the current basic elements of 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:

- 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 that all have 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 possibly consisting of multiple directories
    - Calibration Collection – Could contain calibration data separate from the data collection(s) and/or more detailed information about complex calibration efforts
    - Other Collection(s) as required
- 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.

Peer Review

PDS archiving efforts require that all data submissions be peer-reviewed before being designated as certified data and accepted into the PDS. The review committee consists of a

small number of scientists who have relevant expertise and typically several PDS representatives who have PDS4 technical expertise. The peer review process results in a list of recommendations called liens that must be addressed before the data set can be accepted.

Certification

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. These other contents include the following:

- **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, rules) 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

Imaging Node Contact Information

You've likely already been in touch with us, but here is a reminder of our contact information---please don't hesitate to contact us 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