PDS4 Build-A-Bundle Exercise Worksheet

PDS4 Training, 2017 Fall AGU Meeting, New Orleans, LA, December 12-15, 2017

Objectives

1. Develop a basic understanding of the key components of a PDS4 archive.
2. Develop a basic understanding of the structure of a PDS4 label.
3. Learn how to use the PDS4 documentation for determining PDS4 label content and syntax.
4. Produce a set of valid PDS4 archive products.

Step 1: Archive Bundle Organization Design

Normally at this stage you would want to think about the various types of data files that are to be archived, and think about the most logical way of organizing them into bundles and collections. There is no hard and fast rule governing how a PDS4 archive is to be organized. However, data providers may find it useful to consider the following questions:

- What organization makes sense for the data?
- What organization are other data providers on the same project planning on using?
- What are data users likely to find the most useful organization?

Typical criteria for archive organization include data purpose, data processing level, mission phase, observation target, etc.

As there is only a single data product for this exercise there will be a single bundle with separate collections for the data and the document products.

Step 2: Define Bundle and Collection Identifiers

Once the archive organization has be completed, a unique set of identifiers must be defined for each entity. In order to help provide a unique id, we suggest an ordered approach.

Bundle Identifiers

Bundle identifiers must be unique across the entire PDS archive. We try not to include researcher’s names in the bundle id. For clarity we try to include information about mission, instrument, and type of science to round out the Bundle ID and create a unique identifier. Individual nodes may have specific rules for this so it is always a good idea to be in regular communication with your node representative.

Examples:

Mars Pathfinder Atmospheric Opacity Data	urn:nasa:pds:mpf_opacity
Voyager Calibrated IRIS Data	urn:nasa:pds:vgr_iris_calibrated

MAVEN SWEA Calibrated Data urn:nasa:pds:maven-swea-calibrated

Bundle LIDs have the form:

urn:nasa:pds:*bundle_id*

- Allowed characters include: lowercase letters, digits, dash, period, and underscore.

Enter the logical identifier for your bundle:

urn:nasa:pds:____

Collection Identifiers

Collection identifiers must be unique with the bundle. The first element of the collection LID is always the collection type (i.e. data, document, browse, calibration, etc.). Collections are typically subdivided by purpose or processing level. **Remember, collection LIDs are a combination of the bundle LID + an extra segment used as the collection ID.**

Examples:

LADEE UVS Document Collection *urn:nasa:pds:ladee_uvs:document*

Phoenix MET Raw Data Collection *urn:nasa:pds:phx_met:data_raw*

MAVEN SWEA Calibrated Data Collection
urn:nasa:pds:maven.swea.calibrated:data

Collection LIDs have the form:

urn:nasa:pds:*bundle_id*:*collection_id*

- The collection LID includes the bundle LID.
- The collection_id must begin with the collection type.
- Allowed characters include: lowercase letters, digits, dash, period, and underscore.

Enter the logical identifiers for your collections here:

urn:nasa:pds:_________________________:______________________________

urn:nasa:pds:_________________________:______________________________

Basic product LID formation rule

While you would normally not want to list all of the individual product LIDs at this point in archive design, you should think about the convention used to form them. Basic product LIDs must be unique within the collection. Frequently, the file base name (file name with the file extension removed) converted to lowercase is used for the product_id portion of the LID. The file version number (if present) should also be removed to insure that subsequent versions of the same product have the same LID. **Remember, basic product LIDs are a combination of the collection LID + an extra segment used as the product ID.**

Describe the basic product LID formation rule here:

Step 3: Generate Document and Document Collection Products

When preparing individual basic product labels, it is often best to start with documents. Documents will often contain information that is key to using the data files. As a result, it is often desirable to include references (by LID or LIDVID) to key documents throughout the archive. Furthermore, document labels are often once-off products not requiring the same level of automation required for data set containing many individual data products.

Document File Label Generation

- Open the document template file (document_template_1900.xml) in a text editor.
- Fill in the following parameters:

<logical_identifier>	Fill in the document collection LID defined in step 2. Add the product_id segment.
<title>	While this is a title for the product, for documents it is typical to use the title of the document itself.
<product_class>	Always “Product_Document” for document products.
<Citation_Information>	Optional. Contains information about the authors, publication year, with freeform fields for keywords and descriptions of the file. The *description* element is a description of the product, and not a citation description.

<Modification_History>	Optional. This is a place to record when this file gets updated complete with version updates and a description field for describing the changes.
<Context_Area>	Optional. Area to list physical or conceptual entities associated with the document (e.g. mission, spacecraft, instrument, target, etc.). The values for <name>, <type>, and optional <lid_reference> should be extracted from the relevant PDS Context Product where one exists. See optional reading at the end of this document for more information.
<Document>	Contains file-specific metadata: Document name, authors, publication date, document editions (multiple formats/languages could be other editions), file format information, DOI info, etc.
<document_name>	Typically the document title in "Title Case"
<Document_File>	Describes the various "editions" (file formats, languages, etc.) of the document
<local_identifier>	Provides an identifier by which a particular document

- Save the revised template as "*filename*.xml", where *filename* is the name of the document file being labeled.

Document Collection Product Generation

Once the document files are labeled (in our exercise – one document) we have the LID necessary to build the *Collection Inventory File* for this collection.

The collection inventory is a listing of all products considered part of this collection. The inventory consists of a 2-column, comma-separated list. The first column is a designator designed to alert the registry system of whether or not the listed product is new to the PDS4 archive or could be found in some other preregistered place.

"P" *(Primary)* designates that the product is being registered for the first time, and is present in this collection.

"S" *(Secondary)* designates that the product has been already registered and may or may not be physically present in this collection.

The second column designates the ***LID::VID*** of each included product. The LID::VID is a combination of the logical identifier (LID) and the version identifier (VID) separated by a double colon.

So for our example in the Document Collection, the inventory file should consist of a 2-column table with only one entry – the entry for the document file.

- Create a new text file.
- Add the following text:

P,*document_product_LIDVID*

Where “P” is literal, and *document_product_LIDVID* is the LIDVID of the document product that you just created.

- Save the new text file with the name:

  collection_*bundle_collection*.csv

Where *bundle*, and *collection* are the bundle and collection ID’s respectively.

Document Collection Label Generation

- Open the collection template document (collection_template_1900.xml) in a text editor.
- Fill in the following parameters:

<logical_identifier>	Defines the LID for this collection. Fill in the document collection LID defined in step 2.
<title>	Collection title
<product_class>	Always “Product_Collection” for collection products.
<Citation_Information>	Optional. Probably not relevant for document collections.
<Context_Area>	Optional. Should roll-up context information from the products which are members of the collection.
<Collection>	Contains parameters describing the collection.
<collection_type>	Always “Document” for document collections.
<File>	Contains file and statistical information for the collection inventory file that you just created.
<Inventory>	Contains record structural information for the collection inventory file. As the collection inventory file format is fixed, this section is identical for all collection files and has been hardcoded in the template.

- Save the document collection label using the same file basename that was used for the collection inventory file.

Step 4: Generate Data and Data Collection Product Labels

Data (Product Observational) File Label Generation

The next step is to generate the data product label and data collection product. The first step is to select the appropriate data template file for the type of data that you are labeling. The following template files are provided:

table_binary_template_1900.xml

table_character_template_1900.xml

table_delimited_template_1900.xml

- Open the appropriate template file in a text editor.
- Fill in the following parameters:

<Identification_Area>	Consists of the following subclasses:
<logical_identifier>	Fill in the data collection LID defined in step 2. Add the product_id segment.
<title>	Fill in a title for the data file. This title should distinguish this data file from others in the same collection.
<product_class>	Always “Product_Observational” for data products.
<Citation_Information>	Optional. See document label for details.
<Modification_History>	Optional. See document label for details.

The contents of <Observation_Area> for observational product labels is identical to that of <Context_Area> in other types of labels. However, unlike <Context_Area>, <Observation_Area> and many of its subclasses are required rather than optional.

- Fill in the following parameters:

<Observation_Area>	Consists of the subclasses listed below. Values for <name>, <type>, and optional <lid_reference> should be extracted from the relevant PDS Context Product where one exists for all of the <Observation_Area> subclasses. A more detailed discussion is provided in the optional reading at the end of this document.
<Time_Coordinates>	Contains start and stop times for the observation.
<Primary_Results_Summary>	Optional. Contains parameters describing the data, including <purpose> (Science, Calibration, etc.), <processing_level> (Raw, Calibrated, Derived, etc.), <discipline_name> (Atmospheres, Fields, Particles, Imaging, etc.), and facet parameters indicating the scientific content of the observation (Color (Imaging),

<Investigation_Area>	Identifies the investigation (i.e. mission, observing campaign, etc.) of which the data are part
<Observing_System>	Identifies the components which are responsible for capturing the data (i.e. spacecraft, instrument, etc.). The <Observing_System_Component> subclass should be repeated for each component.
<Target_Identification>	Identifies the subject (target body) of the observation (e.g. Mars, Enceladus, Solar Wind, etc.)

<Reference_List> is used to provide references to related documents, data products, browse products, etc. For this exercise, <Reference_List> should reference the document product created in step 3.

- Fill in the following area:

<Reference_List>	Identifies related products in the PDS archive, or external publications.

<File_Area_Observational> describes the physical characteristics (name, size, MD5 checksum, etc.) and the internal structure of the data file. The subclass(es) selected will depend upon the data file being labeled.

- Fill in the following area:

<File_Area_Observational>	Describes the associated data file. The structure of this area will vary depending upon the type of data files that you are labeling.

Data Collection Product Generation

Once the data files are labeled (in our exercise – one data file) we have the LID necessary to build the *Collection Inventory File* for this collection.

The process for creating a data collection inventory is identical to that for creating one for the document collection. Please refer to step 3 for details.

Data Collection Label Generation

The process for creating a data collection label is similar to that for creating one for the document collection.

- Open the collection template document (collection_template_1900.xml) in a text editor.
- Fill in the following parameters:

<logical_identifier>	Defines the LID for this collection. Fill in the document collection LID defined in step 2.
<title>	Collection title
<product_class>	Always “Product_Collection” for collection products.
<Citation_Information>	While this is still optional for data collections, it is strong recommended that you include <Citation_Information> for your data collections. This will provide information that can be used in publications to identify relevant data sets. PDS is working on producing DOI’s for its collections and bundles that will help simplify this process.
<Context_Area>	Optional. It’s strongly recommended that the context information from the individual member products be rolled-up and included in data collection labels.
<Collection>	Contains parameters describing the collection.
<collection_type>	Always “Data” for data collections.
<File>	Contains file and statistical information for the collection inventory file that you just created.
<Inventory>	Contains record structural information for the collection inventory file. As the collection inventory file format is fixed, this section is identical for all collection files and has been hardcoded in the template.

Step 5: Generate Bundle Readme and Label Files

The final step is to produce the bundle product. The bundle product consists of a label file, and (optionally) a “readme” file. While we recommend that you consider producing a readme file for your archive bundles, we will not create one as part of this exercise. More on the type of information that it might be useful to include in a readme file is provided in the optional reading section below.

Unlike collections, there is not separate bundle inventory file. Bundle members are identified through the <Bundle_Member_Entry> subclass.

Bundle Product Label Generation

- Open the bundle template document (bundle_template_1900.xml) in a text editor.
- Fill in the following parameters:

<logical_identifier>	Defines the LID for this bundle. Fill in the bundle LID defined in step 2.
<title>	Bundle title
<product_class>	Always “Product_Bundle” for bundle products.
<Citation_Information>	Required for bundles. This will provide information that can be used in publications to identify relevant data sets. PDS is working on producing DOI’s for its collections and bundles that will help simplify this process.
<Context_Area>	Optional. Should roll-up context information from the products which are members of the collection.
<Bundle>	Contains parameters describing the collection.
<File_Area_Text>	Contains file and statistical information for the readme file, if one is included.
<Bundle_Member_Entry>	Provides a list of the collections which are members of this bundle. A separate <Bundle_Member_Entry> object must be included for each collection.
<lidvid_reference>	LIDVID for the member collection
<member_status>	Either “Primary” or “Secondary” for this exercise both collections may be considered primary members.
<reference_type>	Takes the form “bundle_has_*type*_collection”, where *type* is the collection type. This will either be: bundle_has_data_collection, or bundle_has_document_collection.

Optional Reading

Context_Area and Observation_Area

Context objects are physical or conceptual entities which may be associated with products in the PDS archive. Context objects include missions, spacecraft, instruments, and targets. Each product label file contains a section where context object references may be provided. In observational (data) products this section is called <Observation_Area>, in other types of products it is called <Context_Area>. References provided in these areas will associate the labeled product with the referenced context object (mission, spacecraft, etc.). These references are designed to be used to support product search.

Standardized values for context object name, type, and LID are defined in PDS4 context products. When a context product exists for a particular object, the values defined in that product should be used in all references to that object. If not context product exists for a particular object, you may request that one be created through you PDS node representative. Context products are managed by the PDS Engineering Node (EN).

For this exercise lookup sheets are provided containing the information that you will need to fill out the Context_Area and Observation_Area subclasses.

Reference_List

Because PDS is committed to providing usable data to the public, we ***require*** data providers to include relevant documentation on how submitted data are to be used and how they were generated. The <Reference_List> section should be used to provide references to documentation that are key to understanding and using the archive. There is no limitation on the number of files that could appear in this section, but most often this will be a handful of important informative documents including spacecraft and instrument references, Software Interface Specifications (SISs) or User's Guides, and/or published refereed manuscripts.

Many document formats are acceptable in a PDS archive, however at least one copy in plain ASCII/UTF-8 or PDF/A must be included.

Bundle Readme Files

A bundle readme file is an optional part of the bundle product. If provided, the readme file could include the types of information listed below. Note that these are only recommendations. Please feel free to include or not include any given item, or to include additional information as desired.

Overview	A brief description of the bundle.
Contents	A list and description of the bundle's member collections.
Data Processing Notes	Caveats, errata, and other information which may affect data usage and that are important for data users to understand.
Version History	A list of all versions of the bundle, with a brief description of each.
Key Documents	A list of documents key to understanding and using the data, including the document location (LID) if they are part of this or another PDS4 bundle.
Contact Information	A list of important contacts which could include: the data provider (PI, archivist, etc.), the PDS curating node personnel familiar with the archive, PDS EN node personnel, etc.