Creating a Collection

Collections are the primary means for organizing related PDS4 products. (Collections are themselves organized into *bundles*.) The member products of a collection have IDs based on the collection ID.

What Goes Into a Collection?

Typically, all the products in a collection will be of the same basic type (observational, document, etc.). Observational collections will also usually contain products all from the same instrument, mission phase, observational target, and/or calibration level. Data preparers may opt to use criteria like review cycle or publishing deadline to assign collection membership in order to facilitate bookkeeping their data deliveries.

Organizing the Data

PDS makes no requirement on physical organization of the data, although data preparers will need to agree on an organization for data transfer to their consulting PDS node. (A typical organization is described in section 2B of the *PDS4 Standards Reference*.)

For collections with a very small (<~6) number of products, everything can go into a single directory. For larger collections, any reasonable directory hierarchy can be used; in which case the collection product itself - the inventory file and the label - will be located in the root directory of that hierarchy.

Describing the Collection

There are two important levels of description needed for collections. First, there's the summary information provided as attributes and short description fields in the collection label. These values are used primarily to support the registry search for data across large sections of the PDS archive.

Second, there is the longer prose description that is needed by users who intend to make use of the data files from the collection either individually or collectively.

The Collection Label

The primary description for the collection is, of course, its product label. Some things to remember for the collection label:

- The <description> in the <Citation_Information> should be a brief abstract of the collection contents. If you are creating a new major version of a previous collection by updating all (or nearly all) of the products in the collection, the abstract should usually include a mention of the major change(s) in the new version.
- The <Modification_Area> should have at least one new entry for each new version of the collection, to indicate what has changed.
- For observational data collections, the <Observing_System> of the <Context_Area> can and should be used to tie the collection as a whole to things like spacecraft and/or instrument wherever appropriate.

The Overview Document

In addition to the brief description and common attributes documented in the collection label, there is typically more to say about a collection. Every non-trivial collection should be accompanied by, for example, an overview of the content, a description of unusual observing circumstances, a quality assessment of the results, or any other aspect of the collection contents of which a user should be

aware. Those familiar with PDS3 should recognize this level of description as the content of the `DESCRIPTION` field in PDS3 data set catalog files.

While the *PDS4 Standards Reference* does not require this additional description, it is strongly recommended data preparers follow this procedure to provide the descriptive information:

- Provide the additional description as either a simple ASCII or UTF-8 text file, or a PDF/A file. Name the file "overview.txt" or "overview.pdf".
- Label that file as a *Product_Document*. This product will be a primary member of the collection it describes, so give it a LID consisting of the collection LID with "overview" as the last element. So, for example, for the collection with this LID:

  `urn:nasa:pds:sample-bundle:sample-collection`

  the LID of the overview file will be:

  `urn:nasa:pds:sample-bundle:sample-collection:overview`
- Place the overview file in the root of the collection directory hierarchy, with the collection label.
- Include the overview product as a primary member of the collection in its inventory table.

Primary vs. Secondary Members

Collections are *aggregate products*. They exist to define relationships between *simple* products like observations or documents. Any collection may contain two types of member products: *primary* and *secondary*.

> **Primary** members have logical identifiers (LIDs) that contain the collection's LID. A simple product must be a primary member of exactly one collection - the one on which the product's LID is based. A simple product is always and inextricably associated with its primary collection. (The PDS sends products to the deep archive in their primary collections.) If a primary member product is updated to a new version, the collection product *must always* be updated as well.
>
> **Secondary** members of a collection are primary members of some other collection. Any simple product may be listed as a secondary member of any number of additional collections. In most cases, when a secondary member of a collection is updated it is not necessary to update the collection.

A collection may contain only primary, only secondary, or both primary and secondary member products.

Compiling the Inventory Table

The *inventory table* identifies all the products - primary or secondary - comprising the collection. The table must be formatted as a comma-delimited table with carriage-return/linefeed carriage control at the end of every line, with one line for each member of the collection. Each line has two fields:

1. The first field is a single character, either "P" or "S", to indicate the status of the member product - *primary* or *secondary*.
2. The second field identifies the member product by logical and version identifiers.

Primary members, indicated by a "P" in the first field, *must* be identified by both logical and version identifiers. The format is:

```
<logical_identifier>::<version_id>
```

The *logical_identifier* (LID) and *version_id* (VID) are both taken from the attributes of the same name in the <*Identification_Area*> of the member product. For primary members that have been updated, only the highest version number of the product is listed in the inventory table.

Secondary members, indicated by an "S" in the first field, may be identified by either LID+VID, or by LID alone. Omitting the VID implies that the latest available version of that member product should be considered a member of the collection. Including a VID implied that only that specific version is considered a member, even if a new version of the product becomes available.

The actual order of the lines in the inventory is not significant; they need not be sorted in any particular order.

Versioning

Every PDS4 product - observation, document, or collection - has its own version number. The product version number tracks changes in both the label and the data files of that product. The version number of a collection product tracks changes in the collection label and in the inventory table.

For the collection, minor version number increments typically indicate small changes in the collection label, while major version number increments indicate changes to the inventory table. A collection inventory table *must* be updated, and major version number *must* be incremented, when any of the following happens:

- A new member product - that is, one with a logical identifier (LID) not previously listed in the inventory table - is added to the collection. The new LID, or LID+VID, is added to the existing inventory table, if any.
- A primary member of the collection is updated and has a new version identifier (VID). In this case the line in the inventory table referring to the old version is replaced with a line referring to the new version of the member product.
- A secondary member identified by LID+VID is updated, and the new version should be referenced by this collection rather than the older version. This case is treated the same way as an updated version of a primary member product - the new LID+VID replaces the old LID+VID.
- A member product, primary or secondary, is dropped from the collection. The line referring to the dropped product is simply removed from the inventory table, and the collection label updated accordingly. This doesn't happen very often, so when it does a specific remark should be included in the <*Modification_History*> section of the collection label noting that one or more products have been dropped.