Chapter 6. Data Set / Data Set Collection Contents and Naming

The Data Set / Data Set Collection Contents and Naming standard defines the conventions for maintaining consistency in the contents, organization and naming of archive quality data sets.

Data Sets are defined in terms of Data Products, which were introduced in Chapter 4. A data set is an aggregation of data products with a common origin, history, or application. A data set includes primary (observational) data plus the ancillary data, software, and documentation needed to understand and use the observations. Files in a data set share a unique data set name, share a unique data set identifier, and are described by a single DATA_SET catalog object (or equivalent).

Data Set Collections are defined in terms of data sets. A data set collection is an aggregation of several data sets that are related by observation type, discipline, target, or time which are to be treated as a unit; that is, they are intended to be archived and distributed together. Data sets in a data set collection share a unique data set collection name, share a unique data set collection identifier, and are described by a single DATA_SET_COLLECTION object (or equivalent). One of the primary considerations in creating a data set collection is that the collection as a whole provides more utility than the sum of the utilities of the individual data sets.

Figure 6.1 shows the relationships among Data Products, Data Sets, and a Data Set Collection.

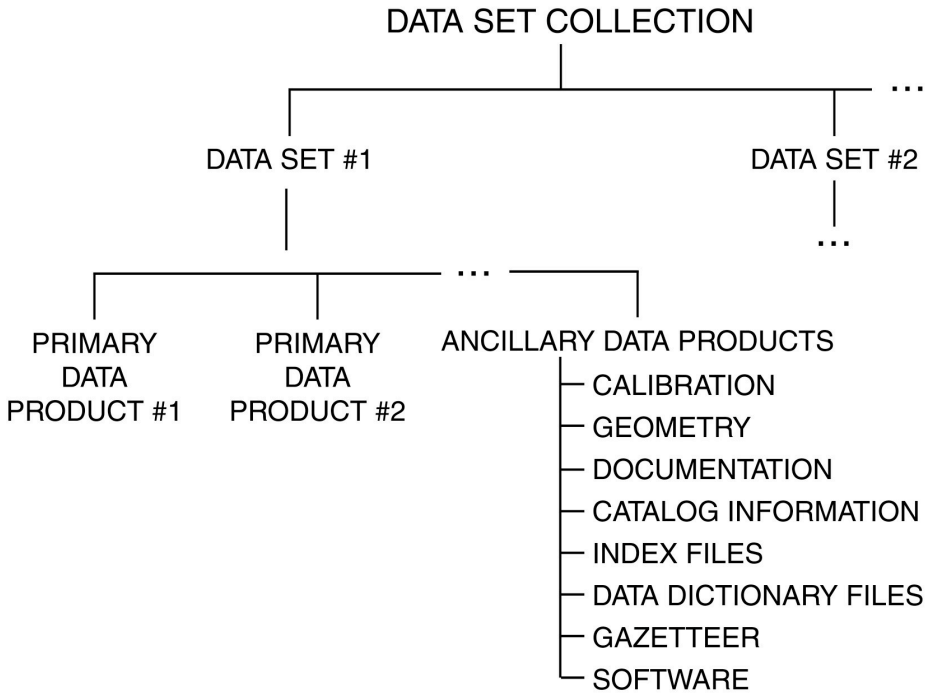

Figure 6.1 Relationships among a Data Set Collection, its Data Sets, and their Data Products.

Note that with respect to Figure 6.1, additional data sets (e.g., Data Set #2) have structure similar to Data Set #1. And, Ancillary Data Products are often organized into directories corresponding to the subject areas shown (see Chapter 19 for a more detailed description of each directory).

Ancillary Data Products may include any or all of the following:

Calibration - Data products used in the conversion of raw measurements to physically meaningful values or data products needed to use the data.

Geometry - Data products needed to describe the observing geometry. Examples include SEDRs and SPICE files.

Documentation - Data products which describe the mission, spacecraft, instrument, and/or data set. These may include references to science papers or the papers themselves.

Catalog Information - Descriptive information about a data set expressed in Object Description Language (ODL) and suitable for loading into a catalog. For more information, see Appendix B.

Index Files - Information that allows a user to locate the data of interest - a table of contents. An example might be a table mapping latitude/longitude ranges to file names.

Data Dictionary Files - An extract of the *Planetary Science Data Dictionary* (PSDD) that is pertinent to the data set and expressed in ODL.

Gazetteer - Information about the named features on a target body associated with the data set.

Software - Software libraries, utilities, and/or application programs to access/process the data products.

6.1 Data Set Naming and Identification

Each PDS data set must have a unique name (DATA_SET_NAME) and a unique identifier (DATA_SET_ID), both formed from up to seven components. The components are listed here; valid assignments for each component are described in Section 6.3:

Instrument host
Target
Instrument
Data processing level number
Data set type (optional)
Description (optional)

Version number

A DATA_SET_NAME must not exceed 60 characters in length. Where the character limitation is not exceeded, the full-length name of each component is used. If the full-length name is too long, an acronym is used to abbreviate components of the name. Where possible, each component of the DATA_SET_NAME should identify and reflect the corresponding (acronym) component used in forming the DATA_SET_ID.

The DATA_SET_ID cannot exceed 40 characters in length. Each component of the DATA_SET_ID is an acronym that identifies and reflects the corresponding (full-name) component used in forming the DATA_SET_NAME. Within the DATA_SET_ID, acronyms are separated by hyphens.

Multiple instrument hosts, instruments, or targets are referenced in a DATA_SET_NAME or DATA_SET_ID by concatenation of the values with a forward slash, "/", which is interpreted as "and." The slash may not be used in any other capacity in a DATA_SET_ID.

6.2 Data Set Collection Naming and Identification

Each PDS data set collection must have a unique name (DATA_SET_COLLECTION_NAME) and a unique identifier (DATA_SET_COLLECTION_ID), both formed from up to six components. A data set collection may contain data sets that cover several targets, be of different processing levels, or have different instrument hosts and instruments. Since the individual data sets will be identified by their own data set names, some of this information need not be repeated at the collection level. Therefore, the DATA_SET_COLLECTION_NAME uses a subset of the DATA_SET_NAME components in addition to a new component, the collection name, which identifies the group of related data sets. The components are listed here; valid assignments for each component are described in Section 6.3:

Collection name
Target
Data processing level number (optional)
Data set type (optional)
Description (optional)
Version number

A DATA_SET_COLLECTION_NAME must not exceed 60 characters in length. Where the character limitation is not exceeded, the full-length name of each component is used. If the full-length name is too long, an acronym should be substituted. Where possible, each component of the DATA_SET_COLLECTION_NAME should identify and reflect the corresponding (acronym) component used in forming the DATA_SET_COLLECTION_ID.

The DATA_SET_COLLECTION_ID must not exceed 40 characters in length. Each component is an acronym that identifies and reflects the corresponding (full-name) component used in forming the DATA_SET_COLLECTION_NAME.

Multiple targets or data processing levels are referenced in the data set collection name or identifier by concatenation of the values with a forward slash (/) which is interpreted as "and."

6.3 Name and ID Components

6.3.1 Restrictions on DATA_SET_ID and DATA_SET_COLLECTION_ID

Within the DATA_SET_ID and DATA_SET_COLLECTION_ID, acronyms are separated by hyphens. The only characters allowed are:

- Uppercase characters, A-Z
- Digits, 0-9
- The hyphen character, "-"
- The forward slash, "/"
- The period character, ".", but only as part of a numeric component (e.g., "V1.0" but not "C.A")

6.3.2 Standard Acronyms, Abbreviations, and Assignments

This section details the standard acronyms and abbreviations required for formulating the DATA_SET_ID and DATA_SET_COLLECTION_ID values. They are also recommended for use, as appropriate, in the formation of other NAME- and ID-class element values. Standard values for data dictionary elements mentioned in the following sections are listed in the PSDD. New values are added to these lists as needed by the PDS data engineers.

1. **Instrument host** name and ID values are selected from the standard value list of the corresponding PSDD entry (INSTRUMENT_HOST_NAME or INSTRUMENT_HOST_ID data element). Note that the acronym EAR has been used for Earth-based data sets without a specific instrument host.

2. **Collection names and IDs** are created as needed by the data preparers in conjunction with the PDS data engineer. Current IDs and their corresponding names include:

GRSFE	Geological Remote Sensing Field Experiment
IHW	International Halley Watch
PREMGN	Pre-Magellan

3. **Target name** values are selected from the standard values listed in the PSDD for the TARGET_NAME element. Target acronyms are selected from the following list:

Target ID	Target Name
A	Asteroid
C	Comet
CAL	Calibration
D	Dust
E	Earth
H	Mercury
J	Jupiter
L	Moon
M	Mars
MET	Meteorite
N	Neptune
P	Pluto
R	Ring
S	Saturn
SA	Satellite
SS	Solar System
U	Uranus
V	Venus
X	Other, (e.g., Checkout)
Y	Sky

NOTE: Satellites or rings are referenced in DATA_SET_NAMEs and DATA_SET_IDs by the concatenation of the satellite or ring identifier with the associated planet identifier; for example:

JR	Jupiter's rings
JSA	Jupiter's satellites

If Jupiter data are also included in the ring and/or satellite data set then only Jupiter ("J") is referenced as the target.

Note that in some cases this component represents the TARGET_TYPE rather than the target name, for example:

A	Asteroid
C	Comet
CAL	Calibration
MET	Meteorite

Valid values for the TARGET_TYPE data element are listed in the PSDD.

4. **Instrument name and ID** values are taken either from the corresponding PSDD element, or from the following list of values designated for certain types of ancillary data:

Names:	INSTRUMENT_NAME data element in the PSDD
IDs:	INSTRUMENT_ID data element in the PSDD
Ancillary Data:	ENG or ENGINEERING for engineering data sets SPICE for SPICE data sets GCM for Global Circulation Model data SEDR for supplemental EDR data POS for positional data

5. **Data processing level number** is the National Research Council (NRC) Committee on Data Management and Computation (CODMAC) data processing level number.

   Normally a data set contains data of one processing level. PDS recommends that data of different processing levels be treated as different data sets. However, if it is not possible to separate the data, then a single data set with multiple processing levels will be accepted. Use the following guidelines when specifying the data processing level number component of the data set identifier and name:

   (a) the processing level number of the largest subset of data or
   (b) the highest processing level number if there is no predominant subset.

Level	Type	Data Processing Level Description
1	Raw Data	Telemetry data with data embedded.
2	Edited Data	Corrected for telemetry errors and split or decommutated into a data set for a given instrument. Sometimes called Experimental Data Record. Data are also tagged with time and location of acquisition. Corresponds to NASA Level 0 data.
3	Calibrated Data	Edited data that are still in units produced by instrument, but that have been corrected so that values are expressed in or are proportional to some physical unit such as radiance. No resampling, so edited data can be reconstructed. NASA Level 1A.
4	Resampled Data	Data that have been resampled in the time or space domains in such a way that the original edited data cannot be reconstructed. Could be calibrated in addition to being resampled. NASA Level lB.
5	Derived Data	Derived results, as maps, reports, graphics, etc. NASA Levels 2 through 5.
6	Ancillary Data	Nonscience data needed to generate calibrated or resampled data sets. Consists of instrument gains, offsets, pointing information for scan platforms, etc.
7	Correlative Data	Other science data needed to interpret space-based data sets. May include ground-based data observations such as soil type or ocean buoy measurements of wind drift.
8	User Description	Description of why the data were required, any peculiarities associated with the data sets, and enough documentation to allow secondary user to extract information from the data.
N	N	Not Applicable

6. **Data set type** provides additional identification if, for example, the CODMAC data processing level component is not sufficient to identify the type or level of data. Following is a list of valid IDs and names that may be used for this component.

NOTE: Several of the values in this table are currently unique to a particular mission (e.g., BIDR and MIDR were used on Magellan). These values may be used on other missions, if deemed appropriate.

ID	Name
ADR	Analyzed Data Record
BIDR	Basic Image Data Record
CDR	Composite Data Record
CK	SPICE CK (Pointing Kernel)
DDR	Derived Data Record (possibly multiple instruments)
DIDR	Digitalized Image Data Record
DLC	Detailed Level Catalog
EDC	Existing Data Catalog
EDR	Experiment Data Record
EK	SPICE EK (Event Kernel)
FK	SPICE FK (Frames Kernel)
GDR	Global Data Record
IDR	Intermediate Data Record
IK	SPICE IK (Instrument Kernel)
LSK	SPICE LSK (Leap Second Kernel)
MDR	Master Data Record
MIDR	Mosaicked Image Data Record
ODR	Original Data Record
PCK	SPICE PCK (Planetary Constants Kernel)
PGDR	Photograph Data Record
RDR	Reduced Data Record
REFDR	Reformatted Data Record
SDR	System Data Record
SEDR	Supplementary Experiment Data Record
SPK	SPICE SPK (Ephemeris Kernel)
SUMM	Summary (data) (to be used in the browse function)
SAMP	Sample data from a data set (not subsampled data)

7. **Description** is optional, but allows the data provider to describe the data set better – for example, to identify a specific comet or asteroid. Following is a list of example values (both IDs and names) that can be used for this component.

ID	Name
ALT/RAD	Altimetry and Radiometry
BR	Browse
CLOUD	Cloud
ELE	Electron
ETA-AQUAR	Eta-Aquarid Meteors
FULL-RES	Full Resolution
GIACOBIN-ZIN	Comet P/Giacobini-Zinner
HALLEY	Comet P/Halley
ION	Ion
LOS	Line of Sight Gravity
MOM	Moment
PAR	Parameter
SA	Spectrum Analyzer
SA-4.0SEC	Spectrum Analyzer 4.0 second
SA-48.0SEC	Spectrum Analyzer 48.0 second

8. **Version number** is determined as follows:

   (a) If there is not a previous version of the PDS data set/data set collection, then use Version 1.0.

   (b) If a previous version exists, then PDS recommends the following:

      i. If the data sets/data set collections contain the same set of data, but use a different medium (e.g., CD-ROM), then no new version number is required (i.e., no new data set identifier). The inventory system will handle the different media for the same data set.

      ii. If the data sets/data set collections contain the same set of data, but have minor corrections or improvements such as a change in descriptive labeling, then the version number is incremented by a tenth. For example, V1.0 becomes V1.1.

      iii. If a data set/data set collection has been reprocessed, using, for example, a new processing algorithm or different calibration data, then the version number is incremented by one (V1.0 would become V2.0). Also, if one data set/data set collection contains a subset, is a proper subset, or is a superset of another, then the version number is incremented by one.

6.4 Examples

For a data set containing the first version of Mars Cloud Data derived from the Mariner 9, Viking Orbiter 1, and Viking Orbiter 2 imaging subsystems, the data set name and identifier would be:

```
DATA_SET_NAME = "MR9/V01/V02 MARS ISS/VIS 5 CLOUD V1.0"
DATA_SET_ID   = "MR9/V01/V02-M-ISS/VIS-5-CLOUD-V1.0"
```

In this example the optional data set type is not used. The other components are:

- Instrument hosts are Mariner 9, Viking Orbiter 1 and Viking Orbiter 2
- Target is Mars
- Instruments are the Imaging Science Subsystem and Visual Imaging Subsystem
- Data Processing Level number is 5
- Description is CLOUD
- Version number is V1.0

Note that the individual components in the DATA_SET_ID closely match the corresponding components used in the DATA_SET_NAME.

The Pre-Magellan Data Set Collection contains radar and gravity data similar to the kinds of data that Magellan collected and was used for pre-Magellan analyses of Venus and for comparisons to actual Magellan data. In conversation the data set might be described as Pre-Magellan Earth, Moon, Mercury, Mars, and Venus Resampled and Derived Radar and Gravity Data Version 1.0. The data set collection name and ID were:
:

```
DATA_SET_COLLECTION_NAME = "PRE-MAGELLAN E/L/H/M/V 4/5 RADAR/GRAVITY
                           DATA V1.0"

DATA_SET_COLLECTION_ID   = "PREMGN-E/L/H/M/V-4/5-RAD/GRAV-V1.0"
```

(This page intentionally left blank.)