White Paper

PDS4 Archiving of CDF Files

August 23, 2013

Todd King, Joseph Mafi

Overview

The Common Data Format (CDF) is a widely used data format for storing data array that is most commonly used in the Heliophysics domain and for data obtained from ground based observatories. It is also being used in the planetary domain and some international space agencies (i.e. JAXA) and crossover projects (i.e. MAVEN) have adopted CDF as the preferred data format for data storage. As with all acquired data there is a need to preserve (archive) the data for future use. NASA's Planetary Data System (PDS) has defined a set of requirements for archiving which defines allowable data structure and requirement metadata. Since CDF is a general and flexible format, files can be constructed in a variety of ways and not all forms of CDF meet PDS archiving requirements. However, adopting a set of "best practices" for constructing a CDF will enable archiving of the data with PDS and reliable generation of alternate forms (i.e. ASCII Tables) of the data. This white paper describes the best practices and the rational for them.

PDS4 Archive Requirements

The Planetary Data System (PDS) prefers transparent, non-proprietary formats when archiving data. Data should be in a form which can be read using the accompanying plain text metadata. Once read, the data should be usable without additional processing. In general, PDS requires data to be in contiguous blocks (or sequences of bytes), but also supports tabular data where field values are interleaved in a uniform and repeated pattern.

CDF Format

The Common Data Format (CDF) is a data format that internally is a linked list set of structured information which can contain metadata and data. The original CDF specification (version 1.0) supported what is called "rVariables". The dimension of an "rVariable" is set globally in a CDF and so the use of "rVariables" is more limiting. Subsequently "zVariables" were introduced which allowed multiple variables with differing dimensions to be included in a single CDF file. For backward compatibility both "rVariables" and "zVariables" are supported, however "rVariables" are considered deprecated and are rarely used. The data portion of variable may be compressed using RLE, HUFF, AHUFF or GZIP algorithms. A variable may also be sparse, that is, only select values are physically stored with values for each index being derived (repeated or interpolated) from the stored values.

A CDF may consist of a single file (combined attributes and variables) or may be composed of multiple files (attributes in one file and each variable in separate files). Typically a CDF is stored as a single file.

One concept central to CDF is that the physical format should be transparent to the user. This is accomplished by accessing the contents of a CDF through a software layer [2]. This is not suitable for archiving since software may not remain functional well into the future. To meet the PDS archiving requirement it must be possible to describe the location and structure of the data using archivable metadata. This is possible with specific forms of CDF files.

The Data in a CDF

The internal structure of a CDF file [1] consists of a blend of metadata and data that is organized into blocks (CDF calls them "records") of structured information. The metadata describes the dimensions and data types of the data. The metadata may also include any number of attributes that provide contextual or semantic information regarding the data. The CDF specification does not define any required attributes, but ISTP/IACG [3] has defined a set of attributes which has been widely adopted. In many instances these attributes can be mapped to equivalent concepts (or attributes) in the PDS information model. When placing a CDF in a PDS archive this mapping should be performed so that the metadata is in a standard form.

The data for each variable in a CDF file is written as an N -dimensional structure in which all the elements have the same data type. This conforms to the archiving requirements of PDS. However, if the variable is written incrementally, that is with multiple writes, the data will be fragmented in the file because each write introduces a metadata header preceding the block of data. This is not allowed when archiving so data associated with a single variable must be written with a single operation to ensure the data is contiguous.

Requirements for Archivable CDF files

To ensure data in a CDF file will be in an archivable form

1. Create CDF compliant with version 3.4 or later.
2. Use single file CDF.
3. No compression (file or variable).
4. No fragmented variables (all data for a variable must be contiguous in the file).
5. Use only "zVariables" (also recommended by the CDF standard)
6. All data records are physical (record variance for data variables is "VARY")

To aid in the generation of PDS metadata it is advisable to include

1. CDF Tool compliant metadata. [2]
2. ISTP/IACG compliant metadata. [3]

Tools and Techniques

Generating compliant CDF files

Creating a CDF file which is compliant with the structural PDS archive requirements can be achieved using the "cdfconvert" tool which is part of the CDF software distribution [4]. The command and options are:

```
cdfconvert  {src.cdf} {dest.cdf} -single  -network -sparseness vars:srecords.no \
        -compression vars:none -zmode 2
```

This will convert {src.cdf} (which can be a multiple file CDF) and write to {dest.cdf} a copy that is written as a single file with network encodings, no sparseness, no compression and that contains only zVariables.

Using igpp-docgen to create a PDS4 label

The igpp-docgen is an application which defines a environment that can process Apache Velocity[5] templates. It can read CDF files (amount other formats) and make information about the content available for use through the Apache Velocity Template Language[6]

```
docgen -f cdf cdf:cdf:cdf/thg_l2_mag_and_20121226_v01.cdf cdf/pds4label.vm
```

Converting Tables to CDF

Tools exist to convert PDS3 labeled table data into CDF files. PDS-CDF converters are available at <http://mgmt.pds.nasa.gov/converters.html>

References

[1] CDF Internal Format Description; Version 3.4, February 28, 2012; Space Physics Data Facility; NASA / Goddard Space Flight Center. http://cdaweb.gsfc.nasa.gov/pub/software/cdf/doc/cdf34/cdf34ifd.pdf

[2] CDF User's Guide; Version 3.4, February 28, 2012; Space Physics Data Facility; NASA / Goddard Space Flight Center. http://cdaweb.gsfc.nasa.gov/pub/software/cdf/doc/cdf34/cdf34ug.pdf

[3] ISTP/IACG guidelines for CDF, http://spdf.gsfc.nasa.gov/istp_guide/istp_guide.html

[4] CDF Web Site, "Download CDF Software" link, http://cdf.gsfc.nasa.gov/

[5] Apache Velocity, http://velocity.apache.org/

[6] Velocity Template Language, http://velocity.apache.org/engine/devel/vtl-reference-guide.html