PDS4 Training Workshop Introduction

Planetary Data Workshop

Flagstaff, Az

June 14, 2017

Training Session Objectives

- Introduce PDS4 concepts and terminology
- Provide overview of available tools for developing PDS4 archives
- Hands-on exercises for
  - Designing bundles and collections; and LIDs
  - Producing PDS4 labels for simple tables and images
- Demo on validating PDS4 products

Agenda

9:00	Introduction to PDS4 concepts and vocabulary	Guinness
9:30	Help those who have not preinstalled Virtual box	Hardman
9:45	Hands-On 1	Guinness
	Design a bundle	
	Identify collections	
	develop LID algorithms	
10:15	Break	
10:30	Oxygen Demo, Eclipse comments	Raugh
11:10	Introduction to PLAID	Crombie
11:20	Q & A	
11:30	Lunch	
1:00	Hands-On 2	Raugh
	Design label for character table - PLAID, oXy, eclps	
2:15	Break	
2:30	Hands-On 3	Raugh
	Design label for basic image - PLAID, oXy, eclps	
3:30	Demo Validate Tool with known error file	Raugh
4:00	Q&A - Topics for Thursday?	

PDS4 Training Group

- Crombie, K.
- Guinness, E.
- Isbell, C.
- Hughes, S.
- Mafi, J.
- Neakrase, L.
- Padams, J.
- Raugh, A.
- Algermissen, S.

What is PDS4?

- PDS4 is a PDS-wide upgrade from PDS version 3 to version 4
- New generation archive standards and online data system based on modern information modeling
- Improves efficiency of data ingestion, tracking, and distribution
  - Uses Extensible Markup Language (XML) for labels
  - Limits allowable formats for data and documentation

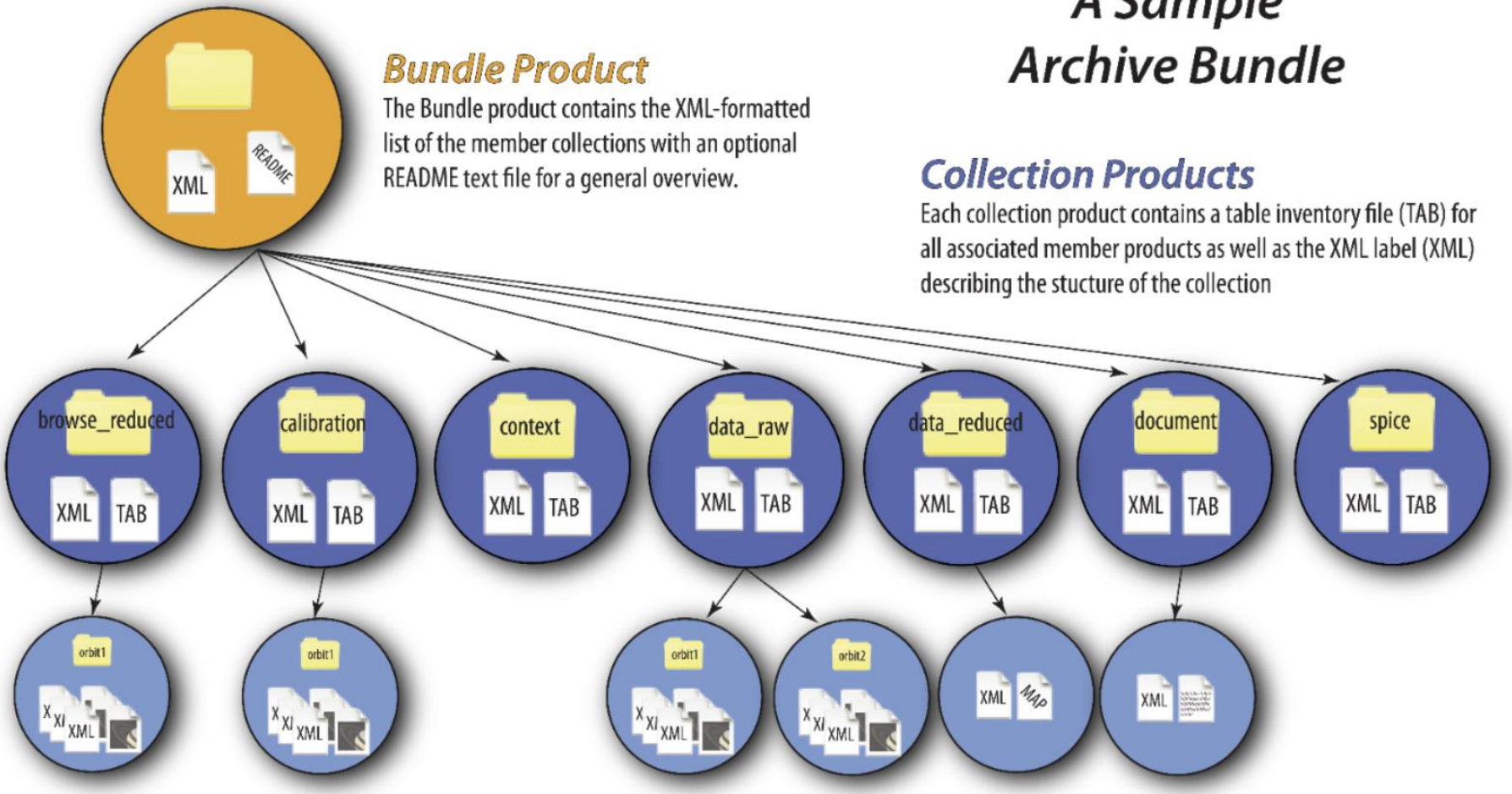
PDS4: The Next Generation PDS

- An explicit PDS4 information model architecture
  - All PDS data tied to a common model to improve validation, discovery, and use
- An explicit software/technical architecture
  - Distributed services both within PDS and at international partners
  - Consistent protocols for access to the data and services
  - Deployment of an open-source registry infrastructure to track and manage every product in PDS

PDS4 Archive Organization (bottom up)

- Basic Product
  - Smallest unit of data registered and tracked in the PDS system (an image, table, document, etc.)
- Collection
  - A group of related basic products of similar type (raw images from an instrument, documents from a mission)
  - Collections are defined by a collection product
- Bundle
  - A group of related collections
  - A bundle is defined by a bundle product and can have an optional readme text file

A Sample Archive Bundle



Bundle Product

The Bundle product contains the XML-formatted list of the member collections with an optional README text file for a general overview.

Collection Products

Each collection product contains a table inventory file (TAB) for all associated member products as well as the XML label (XML) describing the stucture of the collection

Basic Products

Basic products consist of individual or groups of data files (images, headers, documents, tables, etc.) with their associated XML labels that can be placed into logical groupings (collections)

PDS4 Labels: *A single XML label uniquely identifies the product and its component pieces, describes their structure and relative locations, lists related metadata, and provides linkages (references) to related products.*

PDS4 Data Types

- Array
  - A homogeneous n-dimensional array of scalers (images, spectral cubes)
- Table
  - A set of repeating heterogeneous records of scalars (binary or character fixed-length tables)
- Parsable Byte Stream
  - Bytes formatted with standard parsing rules (text files, XML files, CSV tables)
- Encoded Byte Stream
  - Special software needed to decode the bytes (PDF files, JPEG images) – Not used for science observations

PDS4 Information Model

- Defines explicit relationships between major entities of PDS
- Establishes governance for PDS4 metadata for the core dictionary and is extended to discipline and mission level dictionaries
- Provides consistency in PDS labels across many instruments and observation types
- Provides a single authoritative source for data standards
- Based on international standard for data dictionaries

PDS4 Product Labels - XML

- PDS4 information model is implemented in XML for PDS labels
- Each PDS4 product consists of one or more data/document files with a separate XML label file
- The label file is an XML document with a structure defined by the PDS4 common schema
- The label has defined locations for attributes defined by disciplines and missions
- XML supported by 3rd party and open source libraries

PDS3 vs PDS4 label

PDS3

```
PDS_VERSION_ID               = PDS3
LABEL_REVISION_NOTE          = "2006-10-27, Initial; 2008-10-30"

/* File characteristics */
RECORD_TYPE                  = FIXED_LENGTH
RECORD_BYTES                 = 736
LABEL_RECORDS                = 12
FILE_RECORDS                 = 23

/* Pointers to object in file */
^TECP_TABLE                   = 8833   <BYTES>

/* Identification */
DATA_SET_ID                  = "PHX-M-MECA-2-NIEDR-V1.0"
DESCRIPTION                  = "This MECA EDR was generated by the MECA team."
PRODUCT_ID                   = "PS150EM7_00_002BC20848000J1"
PRODUCT_VERSION_ID           = "V1.0 D-22850"
PRODUCT_TYPE                 = "MECA-EM7"
RELEASE_ID                   = "0003"
INSTRUMENT_HOST_NAME         = "PHOENIX"
INSTRUMENT_HOST_ID           = PHX
INSTRUMENT_NAME              = "MECA THERMAL AND ELECTRICAL CONDUCTIVITY PROBE"
INSTRUMENT_ID                = "MECA_TECP"
INSTRUMENT_MODE_ID           = "N/A"
MISSION_NAME                 = "PHOENIX"
...
OBJECT                       = TECP_TABLE
   INTERCHANGE_FORMAT        = BINARY
   COLUMNS                   = 21
   ROWS                      = 11
   ROW_BYTES                 = 736
   OBJECT                    = COLUMN
      COLUMN_NUMBER          = 1
      NAME                   = "CMDTIME WHOLE SECONDS"
      DATA_TYPE              = MSB_UNSIGNED_INTEGER
      START_BYTE             = 1
      BYTES                  = 4
      UNIT                   = SECOND
      DESCRIPTION            = "Spacecraft command receipt time,
                               whole seconds portion"
   END_OBJECT                = COLUMN
   OBJECT                    = COLUMN
      COLUMN_NUMBER          = 2
      NAME                   = "CMDTIME FRACTION"
      DATA_TYPE              = MSB_UNSIGNED_INTEGER
      START_BYTE             = 5
      BYTES                  = 4
       ...
END_OBJECT                   = TECP_TABLE
END
```

PDS4

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model href="http://pds.nasa.gov/pds4/pds/v1/PDS4_PDS_1500.sch"?>
<?xml-model href="http://pds.nasa.gov/pds4/mission/insight/v1/PDS4_INSIGHT_1000.sch"?>
<Product_Observational
    xmlns="http://pds.nasa.gov/pds4/pds/v1"
    xmlns:insight="http://pds.nasa.gov/pds4/mission/insight/v1"
    xmlns:pds="http://pds.nasa.gov/pds4/pds/v1"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://pds.nasa.gov/pds4/pds/v1
                        http://pds.nasa.gov/pds4/pds/v1/PDS4_PDS_1500.xsd
                        http://pds.nasa.gov/pds4/mission/insight/v1
                        http://pds.nasa.gov/pds4/mission/insight/v1/PDS4_INSIGHT_1000.xsd">
    <Identification_Area>

<logical_identifier>urn:nasa:pds:insight_hp3_tem:data_statil_raw:hp3_tlm_raw_0029_20150922_120000</logical_identifier>
        <version_id>1.0</version_id>
        <title>InSight HP3 Tether Length Monitor Raw Product: 1.0</title>
        <information_model_version>1.5.0.0</information_model_version>
        <product_class>Product_Observational</product_class>
    </Identification_Area>
...
<File_Area_Observational>
        <File>
            <file_name>hp3_tlm_raw_0029_20160922_120000.tab</file_name>
            <creation_date_time>2016-09-22T11:45:29.217Z</creation_date_time>
        </File>
        <Table_Character>
            <name>HP3-TLM Raw</name>
            <offset unit="byte">0</offset>
            <records>496</records>
            <description>One row of this table contains raw measurements of the radiometer instrument.</description>
            <record_delimiter>Carriage-Return Line-Feed</record_delimiter>
            <Record_Character>
                <fields>7</fields>
                <groups>0</groups>
                <record_length unit="byte">38</record_length>
                <Field_Character>
                   <name>Spacecraft Clock Time</name>
                   <field_number>1</field_number>
                   <field_location unit="byte">1</field_location>
                   <data_type>ASCII_Real</data_type>
                   <field_length unit="byte">15</field_length>
                   <description>Spacecraft clock time at which measurement occurred</description>
                </Field_Character>
                ...
            </Record_Character>
        </Table_Character>
    </File_Area_Observational>
</Product_Observational>
```

XML Schema and Schematron

- XML schema used to define the attributes, their order and constrain their content
  - Defines data types
  - Defines structure of classes and attributes
  - Specifies required and optional classes and attributes
- Schematron provides additional rules for constraining label content
  - Used for standard value lists
  - Enforces context-dependent constraints (“If this, then that”; “Either this or that, but not both”)
- The schema and schematron are used to validate labels
- The PDS4 core schema is derived from the information model

PDS4 Terminology

- A PDS4 *attribute* is equivalent to a PDS3 keyword
  - Each instance of an attribute has one value
  - Some attributes may be repeated if allowed by the schema
- A *class* is a group of attributes and/or subclasses
  - Again, some classes may be repeated if allowed by the schema
- Each PDS4 product is identified by a *logical identifier* (LID) – globally unique
- Each PDS4 product has a unique *version identifier* (VID) – typically of form M.n
- A *LIDVID* is the combination of a LID and VID – has form LID::VID

More PDS4 Terminology

- A *dictionary* defines attributes, classes and their relationships within a *namespace*.
  - Each namespace has a unique abbreviation and steward.
- A *discipline dictionary* is created to address a particular type of data product or discipline (images, spectra, geometry, cartography, etc.).
- A *mission dictionary* is created to support a specific mission or investigation.

For label preparation and validation, all dictionaries are presented as XML *schema* files that are referenced by the product labels.

Creating and Editing PDS4 Labels

- Any text editor can be used to create or edit a PDS4 label. However …
- An XML-aware editor will make the task much easier
- A schema-aware XML editor is even better.
  - Helps guide you in what classes/attributes are allowed and where they occur in the label
  - Can do validation on the fly
- Several schema-aware editors will be demonstrated in today's workshop

Logical Identifiers (LID) Formation

- PDS4 LIDs must globally unique
  - Therefore there are some guidelines for creating LIDS to meet this requirement
  - LID formation schemes for an archive should be done early in the design phase because LIDs will be used for cross-referencing in the archive
- Constructed using four (bundle), five (collection) or six (basic product) fields
- Fields separated by colons. Colons may not be used within a field
- LIDs must be lower case

Logical Identifiers (LID) Formation Continued

- For PDS submissions, the first three fields are urn:nasa:pds
  - Other agencies will have their own first three fields (urn:esa:psa or urn:jaxa:jaxa)
- Basic product LIDs constructed by appending an id to the parent collection's LID; collection LIDs constructed by appending an id to the parent bundle LID
  - urn:nasa:pds:<bundle_id>
  - urn:nasa:pds:<bundle_id>:<collection_id>
  - urn:nasa:pds:<bundle_id>:<collection_id>:<product_id>
- Bundle LID must be unique among all PDS4 bundles

Resources for More Information

- PDS4 Introduction
  - https://pds.nasa.gov/pds4/about/
- PDS4 Documentation
  - https://pds.nasa.gov/pds4/doc
- SBN Wiki about PDS4
  - http://sbndev.astro.umd.edu/wiki/SBN_PDS4_Wiki
- PDS4 Schema
  - https://pds.nasa.gov/pds4/schema
- PDS4 Software
  - https://pds.nasa.gov/pds4/software