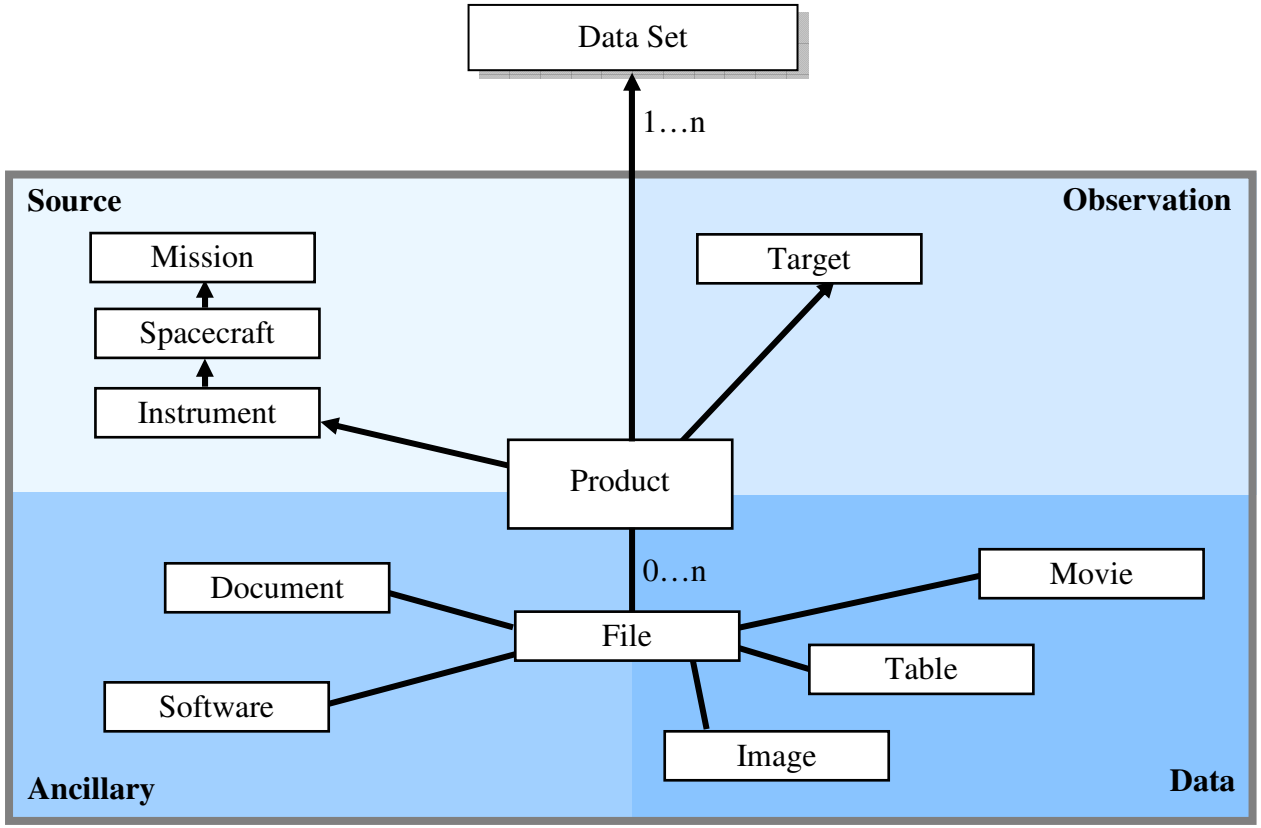
PDS4 Information Model

Proposed Design

Todd King

Information Model - Overview



Information Model Layers

- "Catalog" information (Semantics)
  - Descriptions of instruments, spacecraft, personnel, documents, etc
  - Context for observations (data)
- "Data Object" information (Structure)
  - Descriptions of the organization and storage of data.

Starting Points – PDS3

If we begin with the premise that every byte in a data resource must be described by metadata then we have PDS3.

- To improve PDS3 we would need to do:
  - Redefine some standard value list to be more orthogonal (i.e. product type, instrument type, …)
  - Define new (or expanded) set of primary data objects (we need a new Table object, perhaps image)
  - Add more semantic information (scientific relevance)
  - Compartmentalize mission specific terms.
- In the end we will have a re-formulated version of PDS3 which won't look or function much differently

Start Points – Some Place New

If we allow data to be archived in formats with well-documented, publicly available specifications.

- Then the metadata describes resources at a more conceptual (semantic) level.
  - Data Objects are "external" (defined by format spec.)
  - Resources are associated by reference.
  - Well defined resource types.
  - We need orthogonal lists of attribute values.
  - Allow local (mission specific) extensions.
- This is a real departure from the current PDS approach.

The SPASE Approach

- Developing a new conceptual model is costly (time and money)
- SPASE has developed such a model (Not perfect, but reasonable)
  - Used by
    - NASA's Heliophysics Virtual Observatories exchange information with SPASE.
    - EuroPlanet reached consensus that SPASE will be its exchange model.
    - Cluster Data Archive is based on SPASE.
    - CANOPUS is implementing SPASE compatibility.

What is the SPASE Information Model

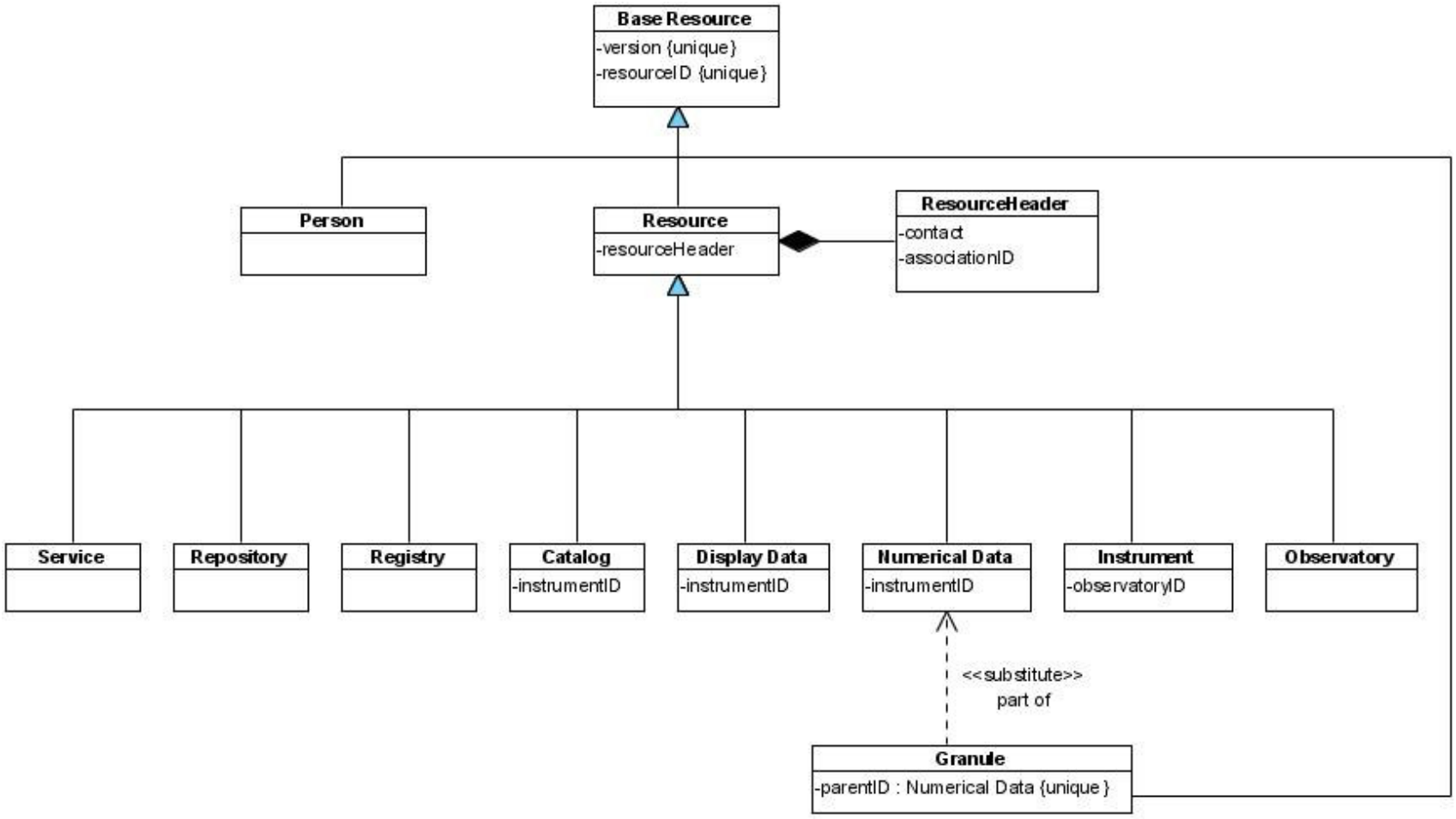
- An information model designed to share resources at the "catalog" layer.
- Science semantics (observed region, measurement type, etc)
- Expects data to be self documented
  - Parameters are "accessible" through keys (field names).
  - Supports a wide range of self-documented formats (table, image, movie, document)
- Numerical Data resources can be composed of multiple granules.
  - Equivalent to a PDS dataset with a single type of data.
- Granule descriptions contain only variant information and inherit attributes of parent Numerical Data resource.

SPASE Information Model - More Details

- Reference to “data” by URL (paths are allowed)
- Standard for unique resource identifiers (URI) that permits local autonomy.
- Allows local extensions for every resource (coming in version 1.3.0)

- Tools for SPASE XML
  - Parsers
  - Validators
  - Editors
  - Harvesters
  - … and more

SPASE Class Diagram



What SPASE does not offer

- Ability to describe the detailed data
  - Data not stored in a self-documented format can not be described and shared.
- Guidelines or requirements for archiving.
  - Which formats are archive quality?
- All generally used resource types.
  - For example, missing Document, Software.

What PDS has always done well

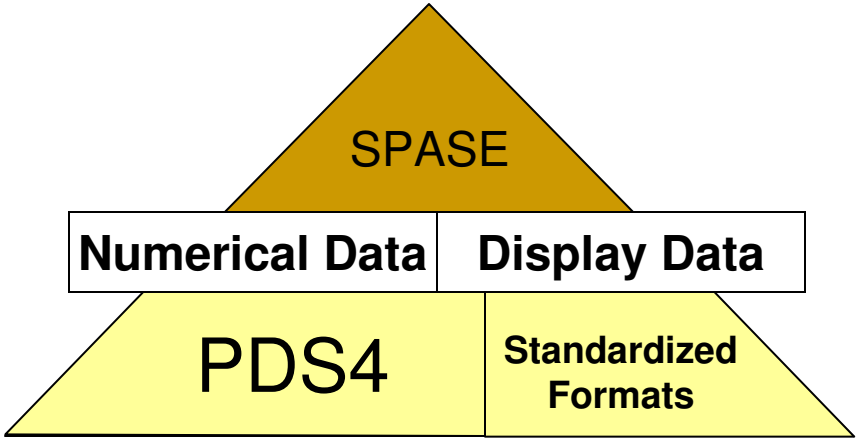
- Requirements for archive quality collections.
- Ability to describe raw data.
  - Even self-documented data must be described the same as "raw"
- Catalog information at the right level.

Some Limitations

- Very limited number of formats with well-documented, publicly available specifications are allowed for archiving (JPEG2000)

What to do?

- Blend the best
- Use SPASE for catalog layer.
- Define PDS data layer to provide access to well-formed data.



SPASE-PDS Equivalency

SPASE Resource	PDS Catalog
Observatory	Mission.cat Insthost.cat
Instrument	Inst.cat
NumericalData DisplayData Catalog	Dataset.cat (single product type)
Person	Person.cat
Registry	Inventory.cat (less detail than SPASE)
Repository	Node.cat
Granule	Product (Label + data)
Service	

Missing Resource Types in SPASE

- Document
  - Recognized need
- Software
  - Recognized need
- Model (and simulation data)
  - Actively being worked
- Collection (Dataset)
  - New
  - More natural than “virtual” numerical data resource.

PDS4 Data Layer

- Define a set of general purpose classes for describing well-formed data.
- Which means that PDS4 is a format specification.
  - Include “PDS4” as a Format in SPASE.
  - PDS4 would be one of the (preferred) formats for archival data.
- Could be PDS3 Data Objects
  - TABLE, SPREADSHEET, IMAGE, SERIES, SPECTRUM, QUBE
  - Or a more limited set of objects to complement allowed formats.

PDS4 Classes

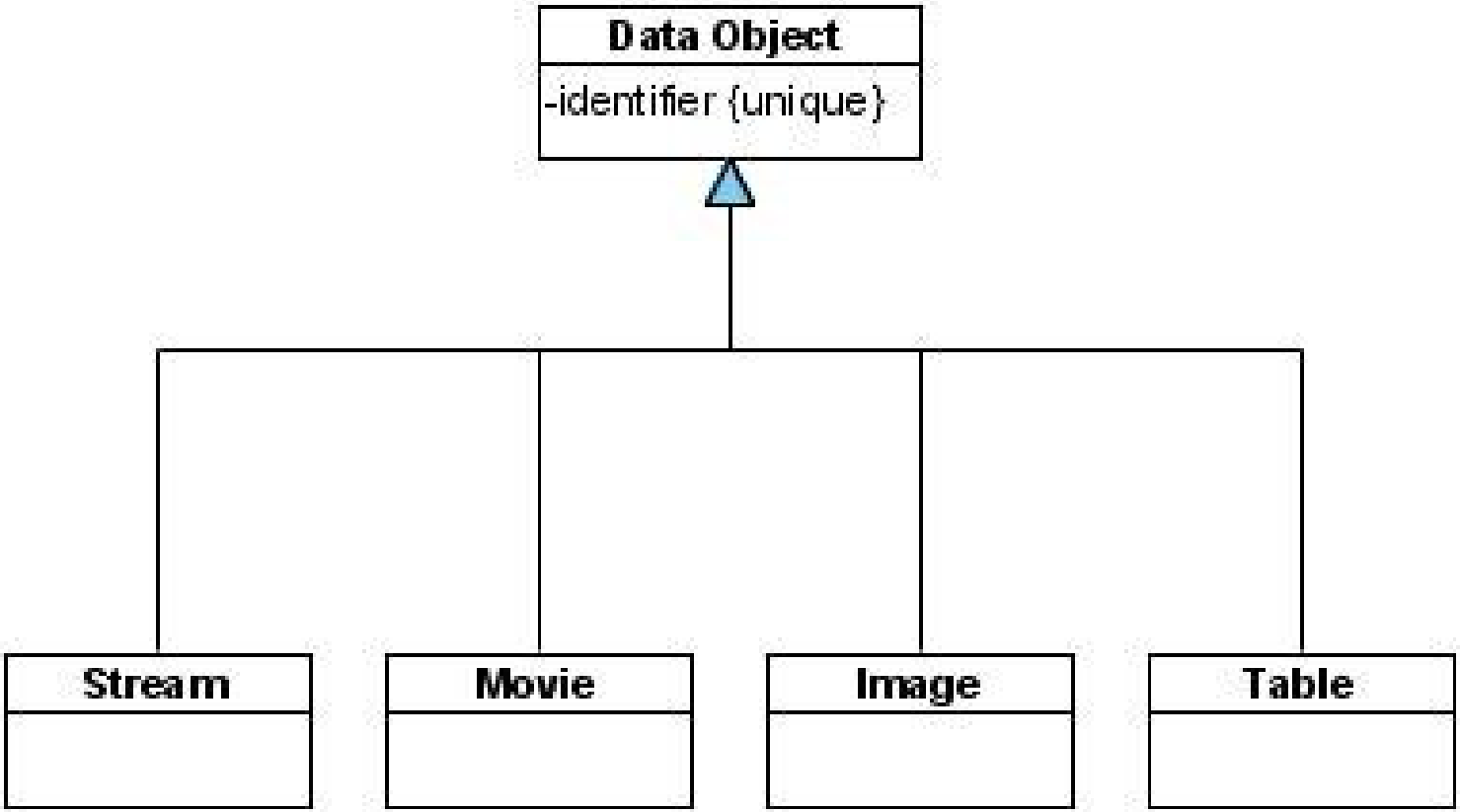


Table Class

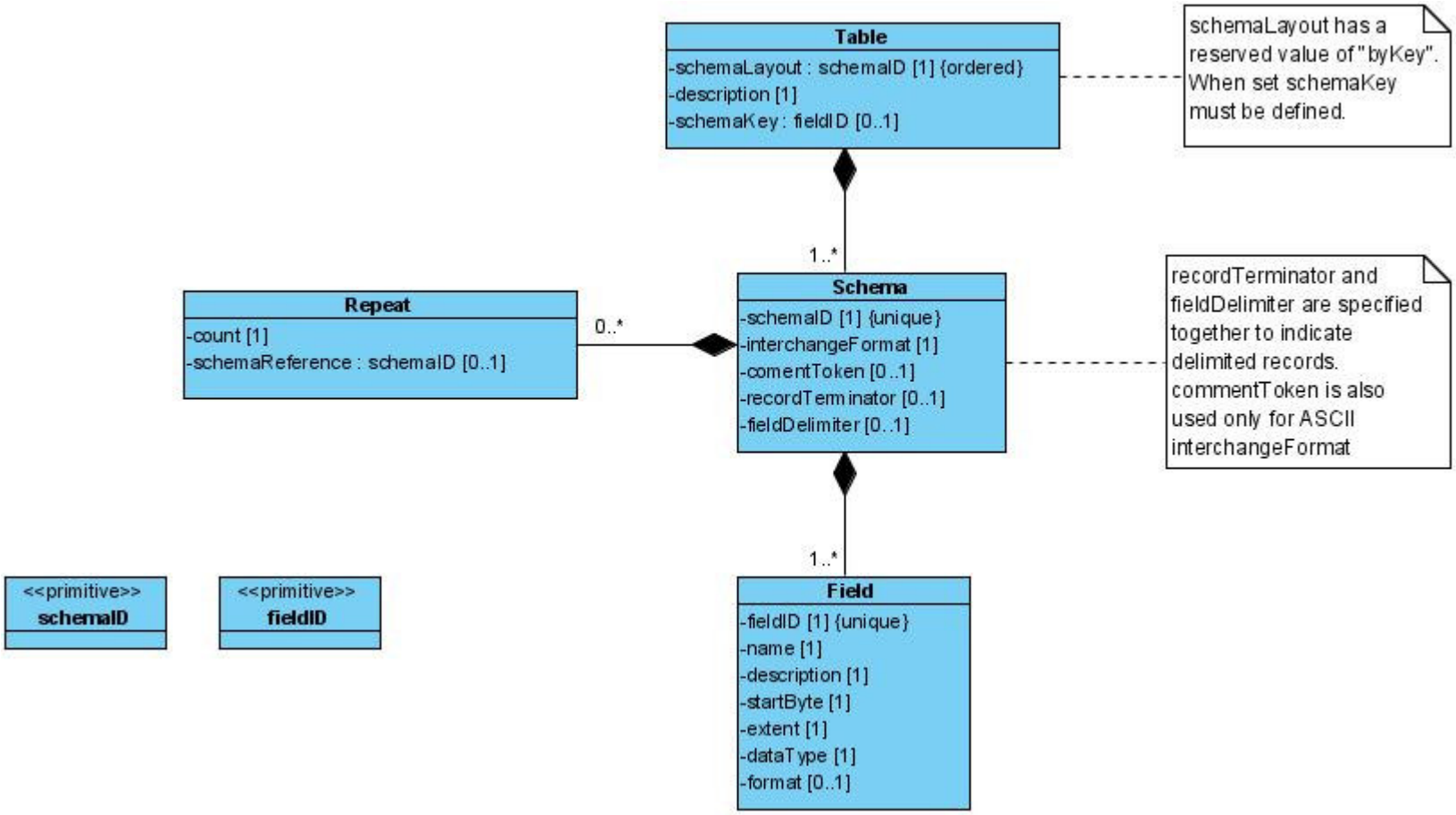


Table
-schemaLayout : schemaID [1] {ordered}
-description [1]
-schemaKey : fieldID [0..1]
schemaLayout has a reserved value of "byKey". When set schemaKey must be defined.
1..*
Repeat
-count [1]
-schemaReference : schemaID [0..1]
0..*
Schema
-schemaID [1] {unique}
-interchangeFormat [1]
-comentToken [0..1]
-recordTerminator [0..1]
-fieldDelimiter [0..1]
recordTerminator and fieldDelimiter are specified together to indicate delimited records. commentToken is also used only for ASCII interchangeFormat
1..*
<<primitive>>
schemaID
<<primitive>>
fieldID
Field
-fieldID [1] {unique}
-name [1]
-description [1]
-startByte [1]
-extent [1]
-dataType [1]
-format [0..1]

Other Classes

- Image: multi-plane, bitmaps.
  - Need input from image experts.
  - What is the minimal set of attributes?
  - AGU acceptable archive formats:
    - vector formats: EPS, Shapefile
    - Raster: JPEG, PNG, TIFF, GeoTIFF
- Movie: sequence of images
  - Wait for results of Movie object working group?
  - Are SPASE supported self-documented formats sufficient? (MPEG, AVI, QuickTime)
  - AGU acceptable archive formats:
    - MPEG-1, MPEG-2, MPEG-4/DivX, VRML

How to Deal With Legacy Products?

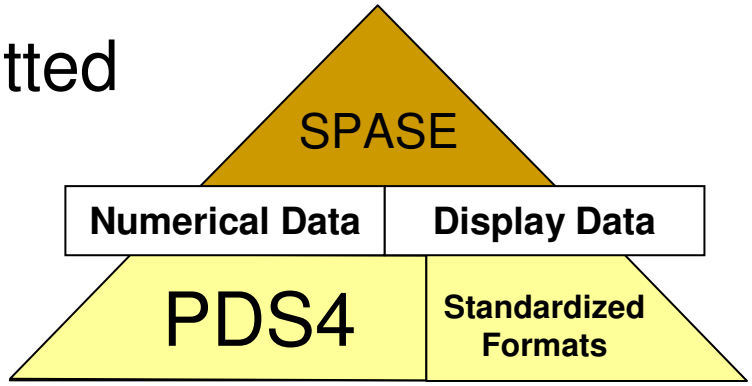
- PDS3 catalog resources can be mapped to SPASE resource description.
- PDS3 labels could be considered a “format” and included in the SPASE list of acceptable formats.

Recommendations

- Adopt a suitable modified SPASE for “catalog” level descriptions.
- Join SPASE and participate in enhancing the data model for PDS needs.
- Define data object information model that can be used to document “raw” data.
  - Base classes are: Table, Image, Movie.
- Define SPASE usage requirements for building archives.
  - What associations are required.
  - Content of collections.
  - What formats are allowed (not all that SPASE permits)
    - Vet and select acceptable archival formats

Benefits of SPASE+PDS4

- PDS can share resources with international and NASA inter-division partners.
- Changes to SPASE layer are vetted by the international community.
- PDS Archiving standards will be useful for a much broader community.



- PDS still defines archiving standards for NASA's planetary community.
- PDS development can benefit more of NASA's science enterprise.