PDS4 Data Architecture & Design (Part II)

March 22, 2010

PDS4 Data Design Working Group

Topics

- Overview and Context
- Information Model
- Data Dictionary
- Grammar
- Standards Management
- Planning and Resoures

Data Dictionary Concepts

- A data dictionary contains a set of "data elements"
- Each data element is defined by a set of attributes
- The Data Dictionary expresses the information model
  - Structural Metadata
  - Descriptive Metadata
- Data elements have associated governance

PDS3 Dictionary Issues

- Current structure is simplistic and is limited by the ODL grammar.
- There is no constraint language.
  - Required and Optional constraints are not enough.
- Less than half of modern data dictionary requirements are met by the current structure.
- Namespaces for local governance were a late add-on.
- Data element creation and maintenance is difficult since PDS wide consensus is often required.

Design Decisions

- Adopt a standard data dictionary model
  - ISO/IEC 11179 – Metadata Registry Specification
  - Provides a standard structure
  - Provides a standard way to define data elements
  - Provides a common understanding of data definition within and across organizations, including international.

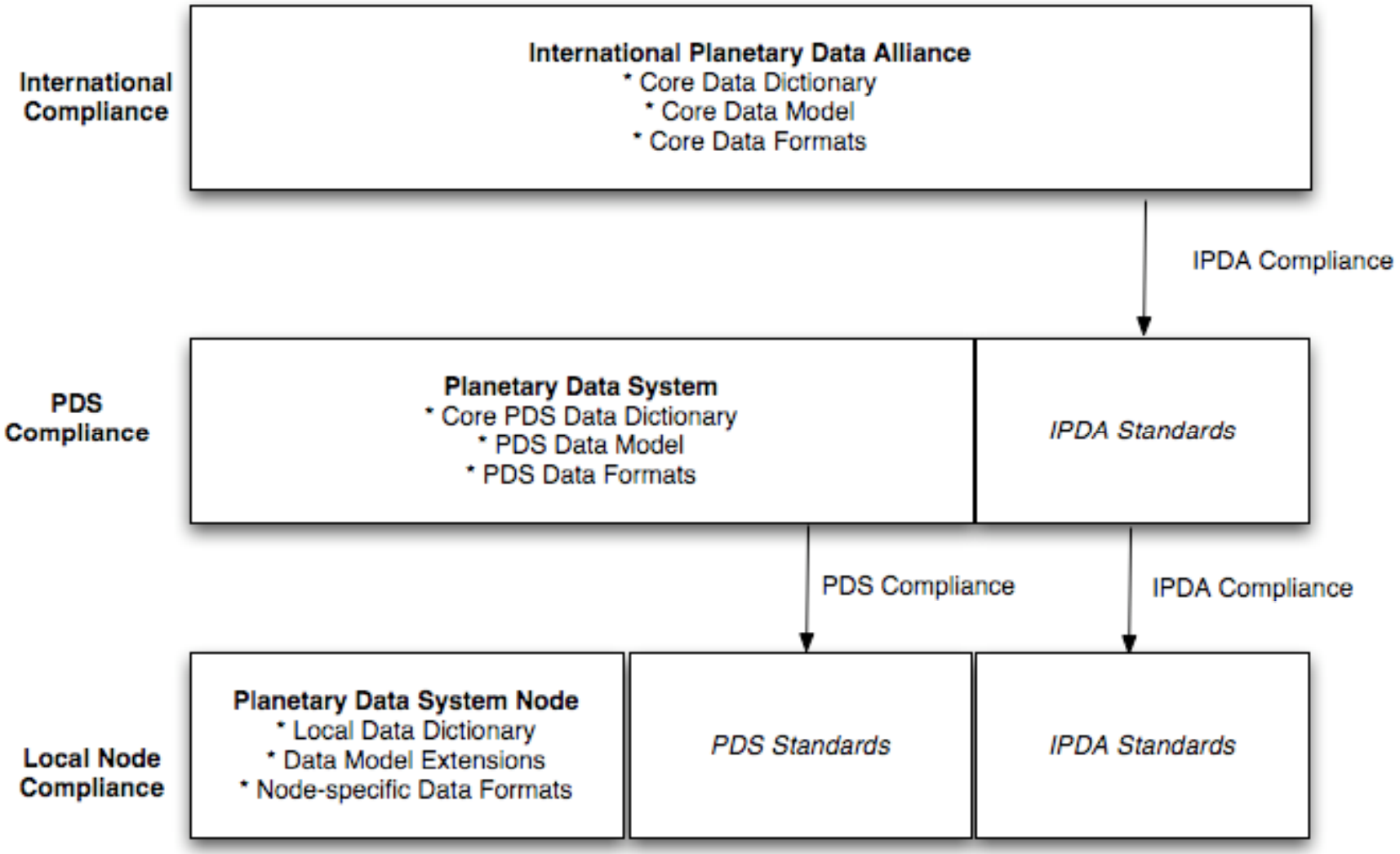
Data Dictionary Model

- **Data Element**
  - Name
  - Submitter, Steward
  - Definition
  - Namespace
  - Source of definition
  - Change log
  - Version
  - Concept
  - Alternate Names
  - Definition in multiple natural languages
  - Classification
  - Unit of measurement
  - Effective Dates
- **Object**
  - Data Elements
- **Valid Value**
  - Value
  - Submitter, Steward
  - Definition
  - Cardinality
  - Source of definition
  - Change log
  - Version
  - Concept
  - Character Set
  - Representation
  - Minimum and Maximum Value
  - Minimum and Maximum Length
  - Alternate encodings
  - Effective Dates

PDS4 Data Dictionary

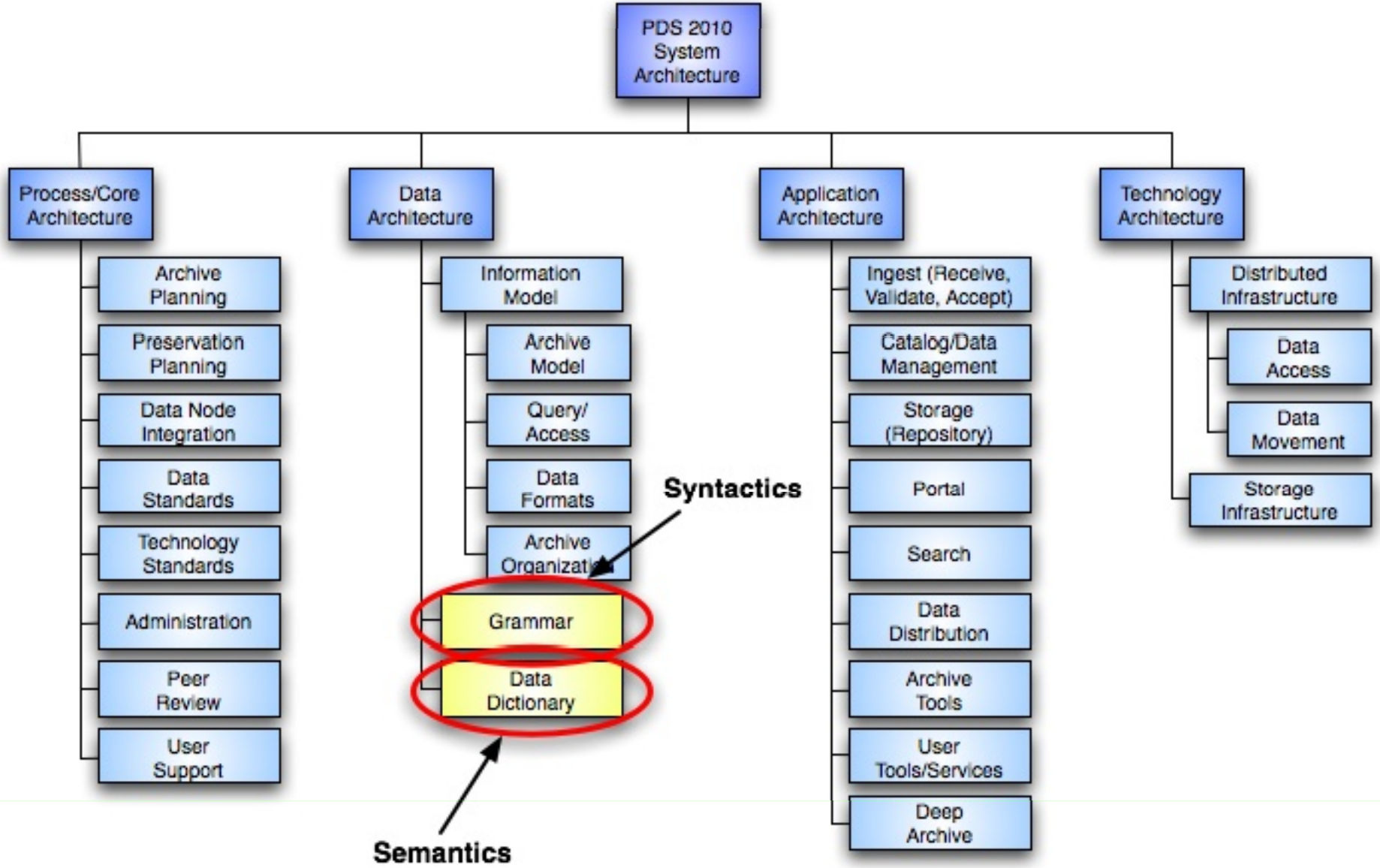
- The data dictionary content is tightly coupled with the information model.
  - Each attribute in the model is defined as a data element using the ISO/IEC 11179 model.
  - Each attribute is assigned a steward
    - Stewards are responsible for the definition and maintenance of an attribute
    - Identifies local governance and localizes changes
  - When implemented in XML Schema each data element becomes an XML element.
  - A steward can assign one or more XML namespaces to group their attributes
  - A Registration Authority is responsible for all attributes in one model
  - Classes are managed similarly.

Standards Hierarchy



Topics

- Overview and Context
- Information Model
- Data Dictionary
- Grammar
- Standards Management
- Planning and Resources


PDS 2010 System Architecture
Process/Core Architecture
Data Architecture
Application Architecture
Technology Architecture
Archive Planning
Preservation Planning
Data Node Integration
Data Standards
Technology Standards
Administration
Peer Review
User Support
Information Model
Archive Model
Query/ Access
Data Formats
Archive Organizati
Grammar
Data Dictionary
Syntactics
Semantics
Ingest (Receive, Validate, Accept)
Catalog/Data Management
Storage (Repository)
Portal
Search
Data Distribution
Archive Tools
User Tools/Services
Deep Archive
Distributed Infrastructure
Data Access
Data Movement
Storage Infrastructure

What is the PDS Grammar?

- The PDS Grammar provides a standard syntax for capturing PDS Labels
  - It is a critical part of the data standards and architecture as it defines the format to exchange information
- We use the term "grammar" because in computer science it describes *formal* language for defining the syntactic rules and constraints for capturing and organizing data from computer languages to metadata
- Grammars generally provide syntactic rules, not semantic rules.
- The decision to use XML will be explained in the system technical design presentation.

Grammar Implementation

- W3C XML and XML Schema have been adopted.
  - Allows for transformation to multiple formats (e.g., XML->PVL)
  - More on this in the system design discussion
- Generic XML Schema for products are generated directly from the information model
- Specific XML Schema are derived for a mission's data products.
- XML Data Product labels are generated from and validated by the specific schema.

XML Integration

- Data Dictionary Service – manages dictionaries using ISO 11179 Model. Exports dictionaries to a XML Schema
- PDS Schema – Captures the types, elements, and structures within the PDS namespace
- Local Schema – Captures the types, elements, and structures for some mission, node, etc. Builds on and inherits from PDS Schema
- Label Schema – Builds on schema from dictionary service to further refine content of a label
- XML Labels – W3C recommendation and a multitude of libraries to read and write XML

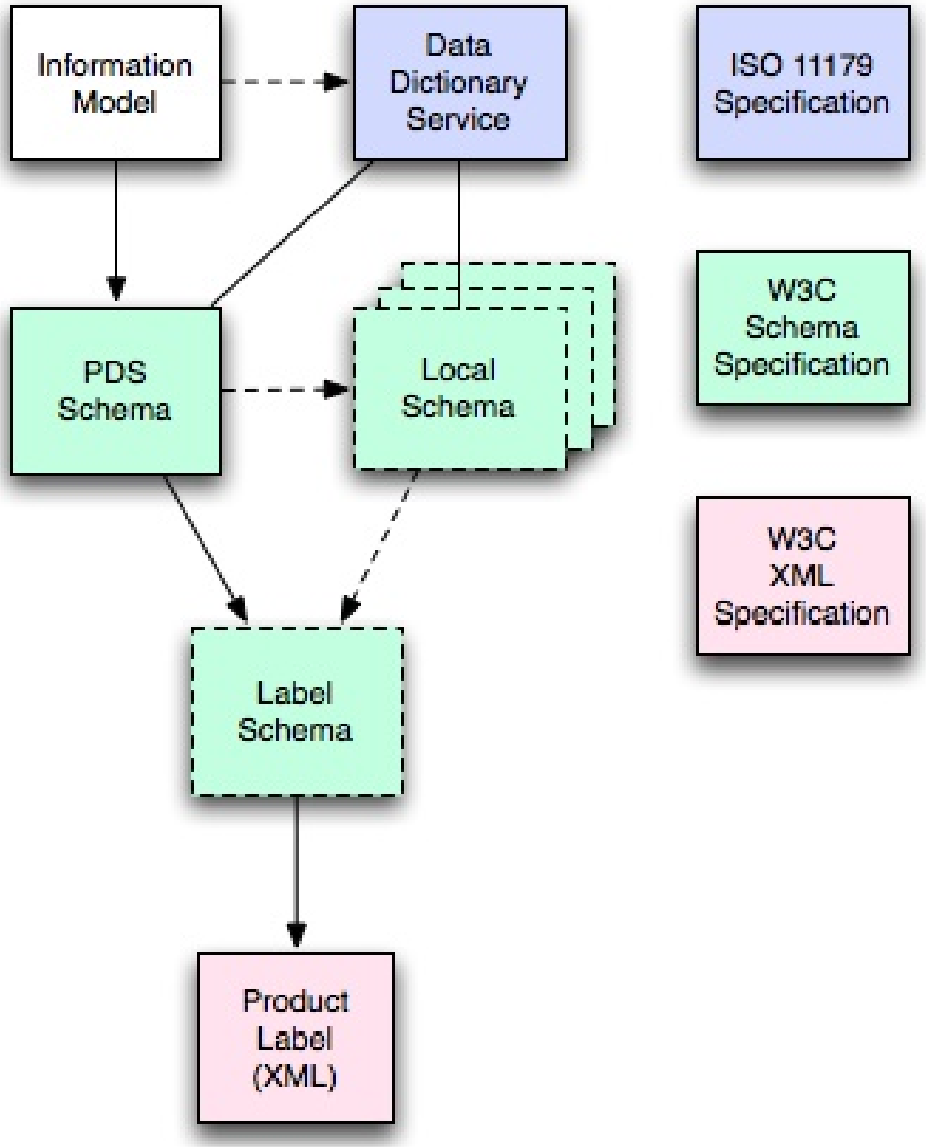
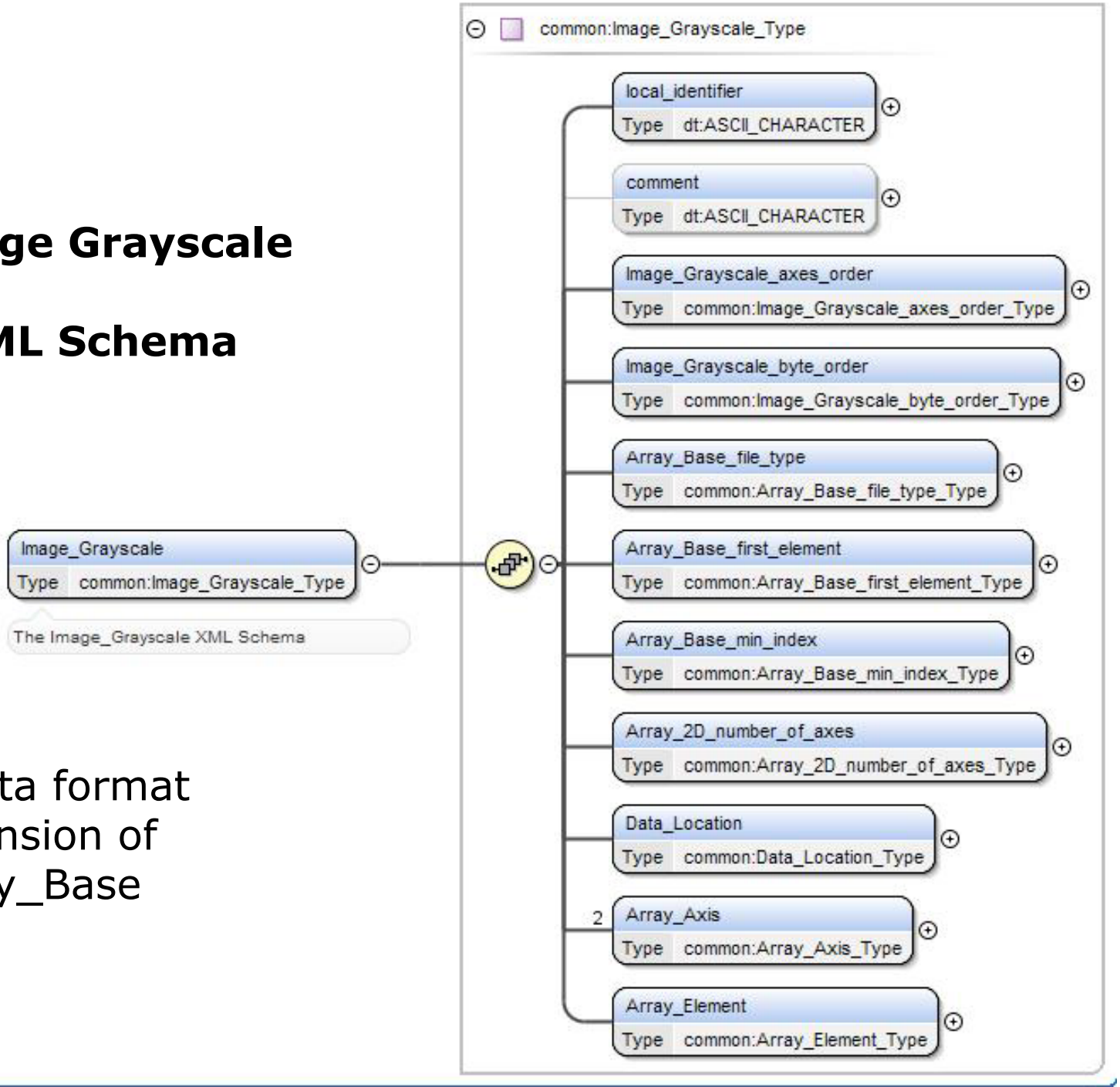


Image Grayscale

XML Schema



A data format extension of Array_Base

PDS4 XML Product Label

```
<?xml version="1.0" encoding="UTF-8"?>
<Product_Table_Character xmlns:ns1="http://pds.nasa.gov/schema/pds4/
  <Identification_Area>
    <logical_identifier>URN:NASA:PDS:PHX-M-TT-5-WIND-VEL-DIR:PDS4_...</logical_identifier>
    <product_id>TELLTALE_91_151</product_id>
    <version_id>v1.0</version_id>
    <title>PHOENIX Mars Wind Experiment</title>
    <collection_lid_reference>URN:NASA:PDS:COLLECTION_PHX-M-TT-…</collection_lid_reference>
  </Identification_Area>
  <Product_Cross_Reference_Area>
    <mission_reference>URN:NASA:PDS:PHOENIX::v1.0</mission_reference>
    <target_reference>URN:NASA:PDS:MARS::v1.0</target_reference>
  </Product_Cross_Reference_Area>
  <Observation_Area>
    <start_time>2008-08-26T20:36:36.856</start_time>
    <stop_time>2008-10-27T15:32:50.952</stop_time>
  </Observation_Area>
  <Area>
    <File>
      <local_identifier>FILE1</local_identifier>
      <directory_path_name>data\MGS-M-RSS-5-TPS-V1.0\SimpleTable…</directory_path_name>
      <name>SimpleTableCharacter_20100201.tab</name>
      <size>111</size>
    </File>
  </Area>
```

PDS4 XML Product Label

```
<Data_Area_Table_Character>
    <Table_Character>
      <number_of_fields>10</number_of_fields>
      <number_of_records>3727</number_of_records>
      <Data_Location>
        <local_identifier>FILE1</local_identifier>
        <offset>1</offset>
      </Data_Location>
      <Table_Record_Character>
        <Table_Character_Field>
          <field_name>SOL</field_name>
          <field_number>1</field_number>
          <field_data_type>ASCII_INTEGER</field_data_type>
          <field_location>1</field_location>
          <field_length>3</field_length>
          <field_format>I3</field_format>
          <field_min_physical>91.0</field_min_physical>
          <field_max_physical>151.0</field_max_physical>
          <field_unit>N/A</field_unit>
          <field_description>PHOENIX Sol number</field_description>
        </Table_Character_Field>
          ...
      </Table_Record_Character>
    </Table_Character>
  </Data_Area_Table_Character>
</Product_Table_Character>
```

Topics

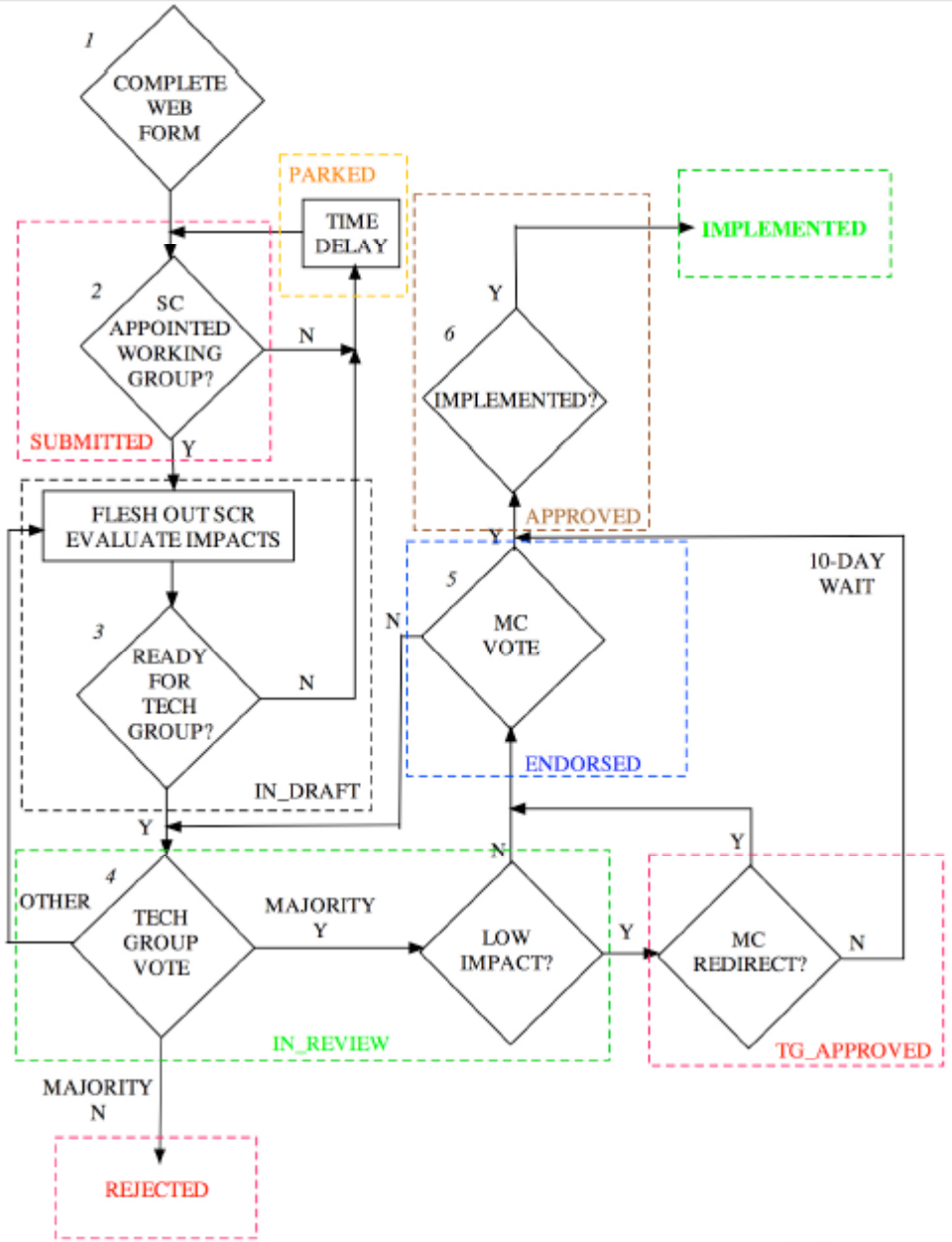
- Overview and Context
- Information Model
- Data Dictionary
- Grammar
- Standards Management
- Planning and Resources

PDS4 Standards Management

- Data Standards
  - The PDS Standards Change Control Process controls changes to any element of the data standards.
  - PDS Node representatives have a role in this process as representatives of the community.
  - A level of local governance is delegated to assigned Steward(s) and the Registration Authority(s) - ISO/IEC 11179

- Information Model / Data Dictionary
  - During development both are managed as metadata databases.
  - They are versioned and configuration managed.
  - They will be deployed as system components and services for operations.

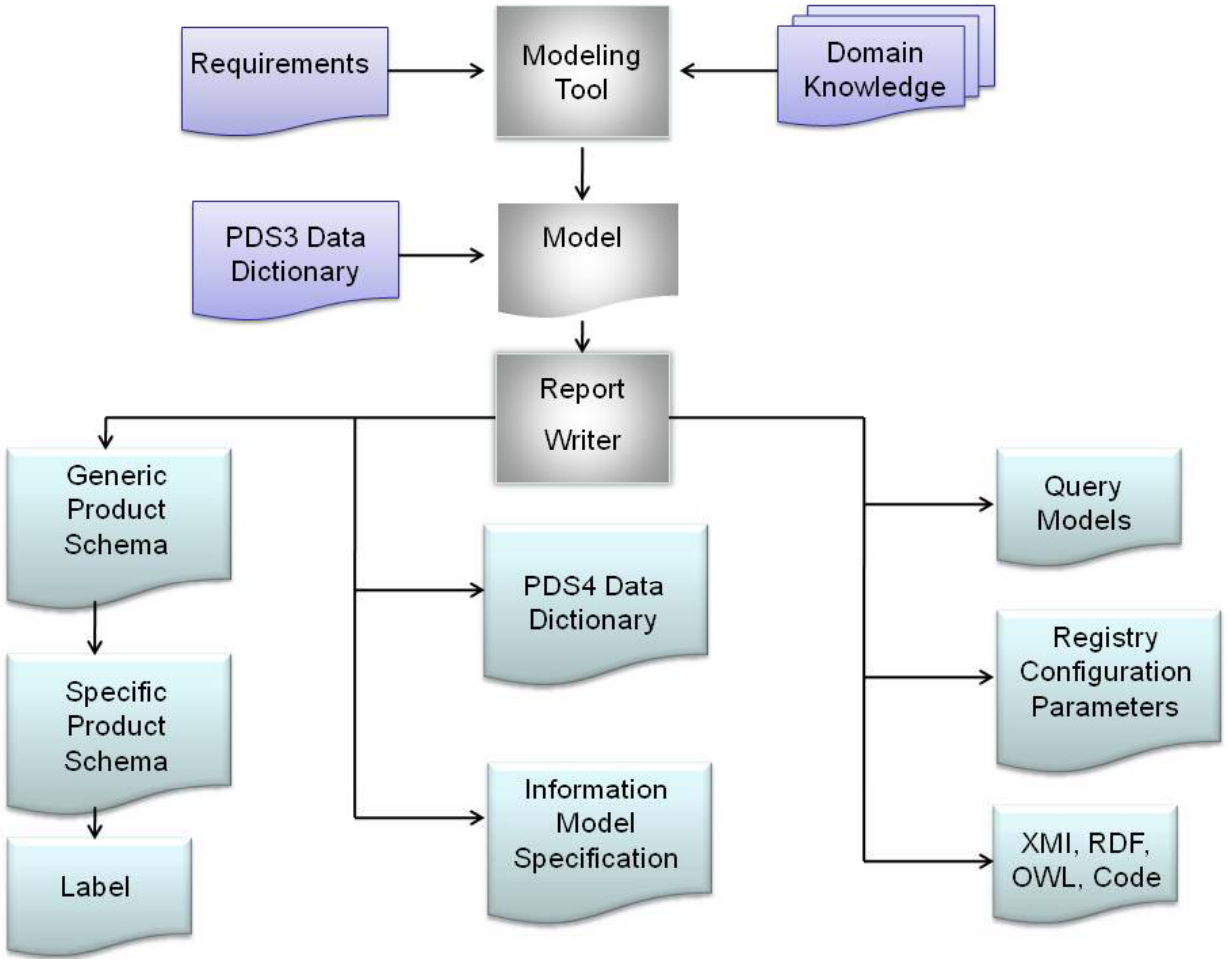
PDS Standards Change Process



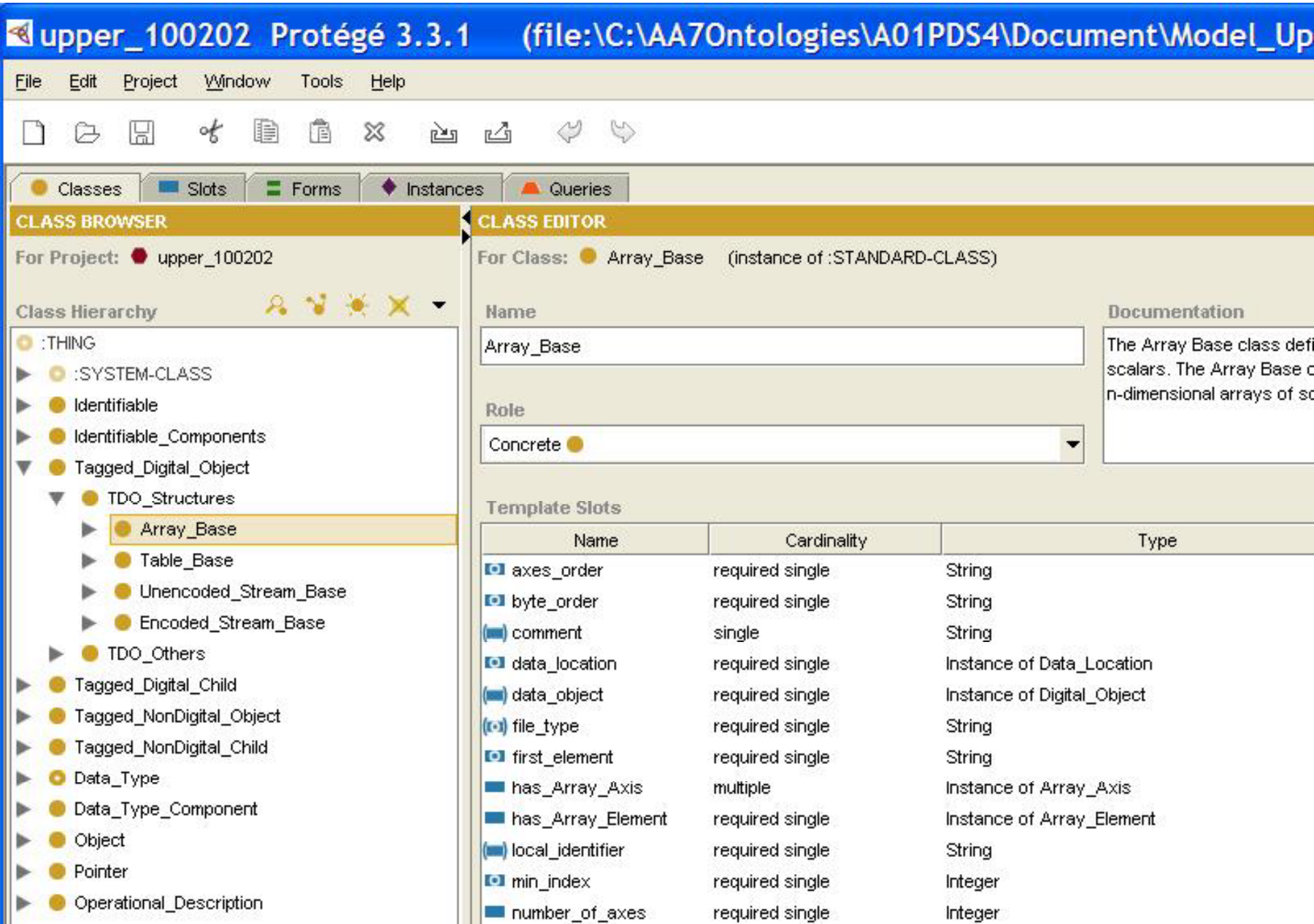
Standards Process

The Model Driven Process

- The model is updated frequently to reflect design decisions.
- The operational files and supporting documents are regenerated for use and testing.
- The current version of the model and the generated artifacts as a whole are an implementation-ready set of data standards.



Protégé Ontology – Array_Base



Information Model Specification

- The Information Model Specification is the formal definition of the PDS4 information model.
  - Generated directly from the ontology modeling database.
  - Model is illustrated using several notations
    - UML Class Diagrams
    - Concept Maps
    - Class definitions in a table format (HTML)

Topics

- Overview and Context
- Information Model
- Data Dictionary
- Grammar
- Standards Management
- Planning and Resources

Document Resources

- PDS4 Standards Reference
  - The definitive source for PDS4 standards
- PDS4 Information Model Specification
  - A formal definition of the information model
- PDS4 Data Dictionary Document
  - An electronic, distributable, and human-readable version of the data dictionary content
- Data Providers Handbook
  - Narrative and examples on how to design, generate, and validate data product labels
- Project Wiki
  - http://oodt.jpl.nasa.gov/wiki/display/pdscollaboration/Data+Design+Working+Group
- Project Website
  - http://pds-engineering.jpl.nasa.gov/index.cfm?pid=5&cid=127

Personnel Resources through Sep 2010

- Anne Raugh (.3)
- Mitch Gordon (.3)
- Lyle Huber (.3)
- Ed Guinness (.3)
- Steve Joy/Joe Mafi (.3)
- Boris Semenov (.1)
- Dick Simpson (.1)
- Elizabeth Rye (.8) (.4 EN; .4 IMG)
- Steve Hughes (.5)
- Ron Joyner (.5)
- EN: 1.4 FTE; DN: 2.1 FTE

- +.2 support/node for PDS4 burst activities per node

NOTE: Resources will continue in FY11, but expect some support on the software-side

Schedule and Progress Chart

Data Dictionary Model
ISO/IEC 11179 adopted

Generic Product Model
Designed and in testing

Fundamental Structures
Designed and in testing

Data Formats
Initial set designed and in testing

Data Element Nomenclature
Rules drafted

Data Dictionary
Clean-up started

Context Model
Design started

XML/Schema
Designed and in testing

Discipline Models
Initial set designed; More needed

PDS4 Standards Reference, Tutorials
Concept of Operations, DPH
In progress; dependent on model

Migration

- Legacy data sets remains permanently in archive.
  - System design incorporates support to integrate PDS3 data with distribution services (e.g. Proxy PDS4 Labels for registering PDS3 products.)
- Migrate data on-demand (as needed)
  - High demand data sets can be migrated first
    - Compliant PDS3 Data Formats only require metadata conversion.
- Schedules can be developed to sync with funding much like we do with restorations

Note: This will be discussed further in the migration/transition planning session

Summary: Major Design Decisions

- Use a formal modeling process
  - Data-Driven Design where entire standards process is driven by the model.
  - Object-Oriented Paradigm
- Design standard data structures for PDS data products
- Implement in XML and XML Schema
  - Support modern languages
- Use standard data dictionary model
- Support federated registries and both conventional and semantic search

From Loosely Described to Formally Defined

[1] **Product Model**
- PDS3 – No Defined Product Model
- PDS4 - Formally Defined

[2] **Data Formats**
- PDS3 – No Defined Structure
- PDS4 – Formally Defined

[3] **Data Dictionary**
- PDS3 – Home Grown
- PDS4 – ISO Standard

[4] **Context Model**
- PDS3 – Catalog/Data Set Centric
- PDS4 – Model/Data Product Centric

[5] **Query Models**
- PDS3 – Ad Hoc
- PDS4 – Derived from archive models

Industry Standards* Referenced and Controlling

- ISO/IEC 11179:3 Registry Metamodel and Basic Attributes specification - Adopted for the data dictionary schema.
- ISO/IEC 11404:2007(E) - Provides the specification for language-independent data types.
- Reference Architecture for Space Information Management (RASIM) - CCSDS 312-0.G-1 – Provides the overarching architectural principles.
- Open Archival Information System (OAIS) Reference Model - Provides a standard for information objects.
- W3C XML (Extensible Markup Language) - Rules for encoding documents electronically.
- W3C XML schema - Type description language for XML documents.
- Electronic Business XML (ebXML) federated registry/repository information model – Provides a standard to support federated registry/repository functions
- RDF/RDFS/XML - RDF is a standard model for data interchange on the Web.

* Not a complete list

Backup

Dictionary Requirements (1 of 2)

1. The Data Dictionary shall define data elements in compliance with the data dictionary model
2. The Data Dictionary shall define a "units-of-measurement" value set in compliance with the data dictionary model.
3. The Data Dictionary shall define a "special-values" value set in compliance with the data dictionary model.
4. The Data Dictionary shall define a "data-element-value-alias" value set in compliance with the data dictionary model.
5. The Data Dictionary shall define a "data-element-value-formation-rule" value set in compliance with the data dictionary model.
6. The Data Dictionary shall define a "standard-value" value set in compliance with the data dictionary model.
7. The Data Dictionary shall define a "namespace" value set in compliance with the data dictionary model.
8. The Data Dictionary shall define linked-in-kind data element relationships in compliance with the data dictionary model.
9. The Data Dictionary shall have a naming standard for the data element title (common name).
10. The Data Dictionary shall provide a namespace attribute for indicating control authorities for groups of data elements.
11. The Data Dictionary shall provide a general data type attribute for classifying a data element according to a non-implementation-specific list of data types.
12. The Data Dictionary shall provide a general classification type attribute for classifying groups of data elements according to common characteristics
13. The Data Dictionary shall provide a system classification type attribute for classifying groups of data elements according to the data system that uses it

Dictionary Requirements (2 of 2)

1. The Data Dictionary shall provide an alias attribute for specifying one or more aliases that are applicable to the referenced data element.
2. The Data Dictionary shall provide a standard value type attribute for specifying the type of standard value that is appropriate for the referenced data element.
3. The Data Dictionary shall provide a minimum and maximum column value attribute for specifying the minimum and maximum numeric values that are applicable to the referenced data element.
4. The Data Dictionary shall provide a minimum and maximum length value attribute for specifying the minimum and maximum permissible length of the character values that are applicable to the referenced data element.
5. The Data Dictionary shall provide two identifier attributes for specifying the unique instance of the data element and a locally defined instance of the referenced data element.
6. The Data Dictionary shall provide a textual-description attribute for defining the referenced data element.
7. The Data Dictionary shall provide a data-element-formation-rule attribute that supplies a rule that is to be applied during the creation of a value for the data element (e.g., the values supplied for reference_key_id must conform to the rules used by a specific professional journal for referencing citations).
8. The Data Dictionary shall provide a series-set attribute for specifying if it is permissible or not permissible to specify values in a series or set.
9. The Data Dictionary shall provide a data-element-partial-label attribute that specifies if it is permissible or not permissible for a data element to exist within a partial-label (e.g., a FMT file).
10. The Data Dictionary shall provide a has-units attribute that specifies if it is permissible or not permissible to associate a unit with the referenced data element.
11. The Data Dictionary shall provide a can-be-locally-defined attribute that specifies if it is permissible or not permissible for a data element to be locally defined (i.e., overwrite the attributes of a data element in the PSDD.
12. The Data Dictionary shall provide a can-take-on-identifiers attribute that specifies if it is permissible or not permissible to pre-pend an identifier to a data element.
13. The Data Dictionary shall provide a can-be-a-pointer attribute that specifies if it is permissible or not permissible for the data element / object to be characterized as one of the three types of pointers (e.g., (1) data location pointer, (2) include pointer, and (3) related information pointer).
14. The Data Dictionary shall provide attributes for defining object classes.

Data Model Response to L3's

Element	Commonality	Extensibility	L3 Req
Data Set	High	Low	1.4.4
Product	Medium	High	1.4.4
Mission	High	Low	1.4.4
Instrument	High	Medium	1.4.4
Host	High	Medium	1.4.4
Target	High	High	1.4.4
Node	High	Low	1.4.4
Person	High	Medium	1.4.4
Reference	High	Low	1.4.4
Document	Medium	Medium	1.4.4
Data Use documentation	Medium	Medium	new
Calibration Information	Medium	Medium	new
Software	Medium	Medium	1.4.4
Identifiable	High	None	3.2.1
Data Object	High	None	1.4.1
Data Structure	High	Medium	1.4.1
Data Interpretation (Image, Table)	Medium	Medium	1.4.1
Data Identification	High	Low	1.4.1
Data Metadata	Low	High	1.4.1
Coordinate System	Medium	Low	3.3.4
Map Projection	Medium	Low	1.4.1
Camera Geometry	Medium	Medium	1.4.1
Volume (Package)	Medium	Medium	1.4.1
Index Table	Medium	Medium	2.6.3
Registry	High	Medium	3.2.1
Repository	High	Medium	3.2.1
Resource	High	High	2.6.2
Manifest	High	Medium	4.1.2
Release	High	Medium	2.6.2
HouseKeeping	High	Low	2.6.2
Data Dictionary	High	Low	1.4.2
Grammar	High	Low	1.4.3

Open/Close Issues

- RFAs
  - Node Exercises and Data Dictionary Review
    - Open/In Progress: 24
    - Preliminary fixes: 18
  - Tech Session
    - Open/In Progress: 44
    - Closed: 25

Registry Configuration Parameters

- Registry Configuration Parameters are generated from the Ontology modeling tool database.
  - The parameters are used to configure the registry for specific product types.
  - Parameters
    - Product type
    - Associations between product types
    - Elements for indexing

Arrays

Homogeneous N-dimensional array of scalars, used for:

- Raster images
- Cube (ISIS core, e.g.)
- 2D spectra

Tables

Repeating set of (scalar) fields (a.k.a. “repeating record structure”, “set of columns”, etc.)

- Comes in both ASCII and binary
- Fixed-width columns and records

Parseable Byte Stream

Bytes must be divided into fields based on syntax rules; bits are not manipulated.

- ASCII or (potentially) binary
- Parsing rules must be known to and accepted by PDS.
- Examples: CSV files, XML files, FITS headers, VICAR headers HTML files, etc.

Encoded File

Bytes must be altered to be understood (bit manipulations, compression, etc.)

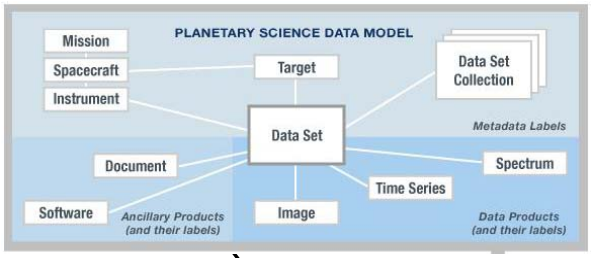
- Must conform to a standard accepted by PDS.
- Likely restricted to documentation (PDF/A, e.g.)

Query Models

- Query models are extracted from the Ontology modeling tool database.
  - Requirements are used to identify desired search parameters in onotology.
    - Product Type
    - Product Attributes
    - Product Relations
  - Parameters are extracted and written to the appropriate notation
    - Relational Schema
    - XML Schema
    - RDF Schema

PDS4 Data Dictionary in Context

Information Model



Used to Create

Label Schema

Validates

Expressed As

Extracted/Specialized

Data Element

has

Class

Planetary Science Data Dictionary

Conforms to the ISO/IEC 11179 Specification

Product

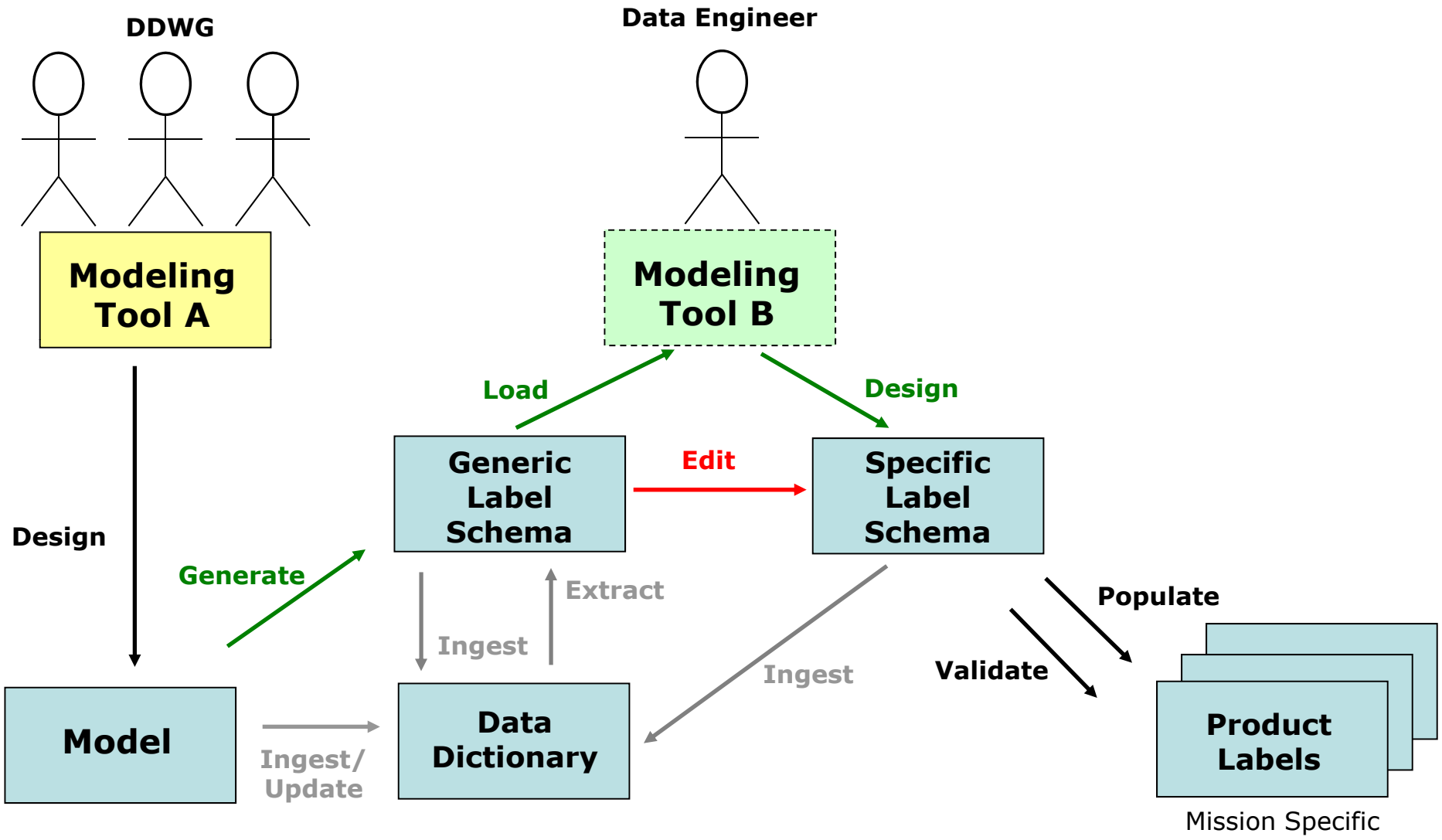
Tagged Data Object (Information Object)

```
OBJECT = IMAGE_GRAYSCALE
  DATA_LOCATION          = ("N2075WE02R.FIT", 0)
  FIRST_ELEMENT          = TOPLEFT
  MIN_INDEX              = 0
  NUMBER_OF_AXES         = 2
  AXES_ORDER             = FAST2SLOW
  BYTE_ORDER             = MSBF
  ELEMENT_BYTES          = 2
  ELEMENT_TYPE           = DECIMAL_INTEGER
  ELEMENT_UNIT           = "DATA NUMBER"
  AXIS_NAME              = ("LINE", SAMPLE")
  AXIS_LENGTH            = (248, 256)
  AXIS_SCALE_TYPE        = ("N/A", "N/A")
  AXIS_UNIT              = ("N/A", "N/A")
END_OBJECT = IMAGE_GRAYSCALE
```

Describes

Data Object

PDS4 Product Label Creation



Current using editor
Current using Oxygen/Future Design Tool
Proposed

PDS4 XML Product Label

```
- <Product_Table_Character xsi:schemaLocation="http://pds.nasa.gov/schema/pds4/common file:/D:/WINWORD/OnlineSTPS-V1.0/Product_Table_Character_2009-06-09p.xsd">
  - <Identification_Area>
    - <puid>
        PDSURN:PHX-M-TT-5-WIND-VEL-DIR:PDS4_ATM_PRODUCT_TABLE_CHAR_ID:1.0
      </puid>
      <Identification_Area_data_set_id>PHX-M-TT-5-WIND-VEL-DIR-V1.0</Identification_Area_data_set_id>
      <product_id>TELLTALE_91_151</product_id>
      <version_id>1.0</version_id>
      <title>PHOENIX Mars Wind Experiment</title>
      <creation_time>2009-01-01T23:34:30</creation_time>
    - <logical_identifier>
        PDSURN:PHX-M-TT-5-WIND-VEL-DIR:PDS4_ATM_PRODUCT_TABLE_CHAR_ID
      </logical_identifier>
    </Identification_Area>
  - <Cross_Reference_Area>
      <instrument_host_puid>PDSURN:PHX-V1.0</instrument_host_puid>
      <instrument_puid>PDSURN:TT-V1.0</instrument_puid>
      <mission_puid>PDSURN:PHOENIX-V1.0</mission_puid>
      <node_puid>PDSURN:ATMOS-V1.0</node_puid>
      <target_puid>PDSURN:MARS-V1.0</target_puid>
    </Cross_Reference_Area>
  - <Observation_Area>
      <comment>comment1</comment>
      <spacecraft_clock_start_count>904250279.448</spacecraft_clock_start_count>
      <spacecraft_clock_stop_count>909588864.598</spacecraft_clock_stop_count>
      <start_time>2008-08-26T20:36:36.856</start_time>
      <stop_time>2008-10-27T15:32:50.952</stop_time>
    </Observation_Area>
  + <File_Area></File_Area>
```

PDS4 XML Product Label

```
−<Data_Area_Table_Character>
  −<Table_Character>
      <local_identifier>PDSURN:PHX_M_TT_TABLE</local_identifier>
      <Table_Base_Character_file_type>CHARACTER</Table_Base_Character_file_type>
      <number_of_fields>10</number_of_fields>
      <number_of_records>3727</number_of_records>
      <record_bytes>88</record_bytes>
    −<Data_Location>
        <file_local_identifier>PDSURN:PDS4_PHX_M_TT_TABLE_FILE_ID</file_local_identifier>
        <offset>1</offset>
      </Data_Location>
    −<Table_Record_Character>
      −<Table_Character_Field>
          <field_name>SOL</field_name>
          <field_number>1</field_number>
          <field_data_type>ASCII_INTEGER</field_data_type>
          <field_location>1</field_location>
          <field_length>3</field_length>
          <field_format>I3</field_format>
          <field_min_physical>91</field_min_physical>
          <field_max_physical>151</field_max_physical>
          <field_unit>N/A</field_unit>
          <field_description>PHOENIX Sol number</field_description>
        </Table_Character_Field>
      +<Table_Character_Field></Table_Character_Field>
      +<Table_Character_Field></Table_Character_Field>
      +<Table_Character_Field></Table_Character_Field>
      +<Table_Character_Field></Table_Character_Field>
      +<Table_Character_Field></Table_Character_Field>
      +<Table_Character_Field></Table_Character_Field>
      +<Table_Character_Field></Table_Character_Field>
      +<Table_Character_Field></Table_Character_Field>
      +<Table_Character_Field></Table_Character_Field>
    </Table_Record_Character>
```