PDS4 Data Architecture & Design (Part I)

March 22, 2010

PDS4 Data Design Working Group

Topics

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

What is PDS4?

- A transition from a 20-year-old collection of data standards to a modern set of data standards constructed using best practices for standards development.
- Fewer, simpler, and more rigorously defined formats for science data products.
- Use of XML, a well-supported international standard, for data product labeling, validation, and searching.
- A data dictionary built to the ISO 11179 standard, designed to increase flexibility, enable complex searches, and make it easier to share data internationally.

Key Deliverables

- Information Model
  - The Information Model defines the information objects* in the PDS archive. These information objects include data structures, data formats, data products, archive collections, documents, software, and investigations.

- Data Dictionary
  - Model - The Data Dictionary Model provides the schema for the data dictionary.
  - Content - The Data Dictionary documents the data elements used in the Information Model.

- Grammar
  - The grammar is used to capture the metadata for the archive.

- PDS Standards Reference V4.0
  - The PDS Standards Reference V4.0 documents the overall standards architecture.

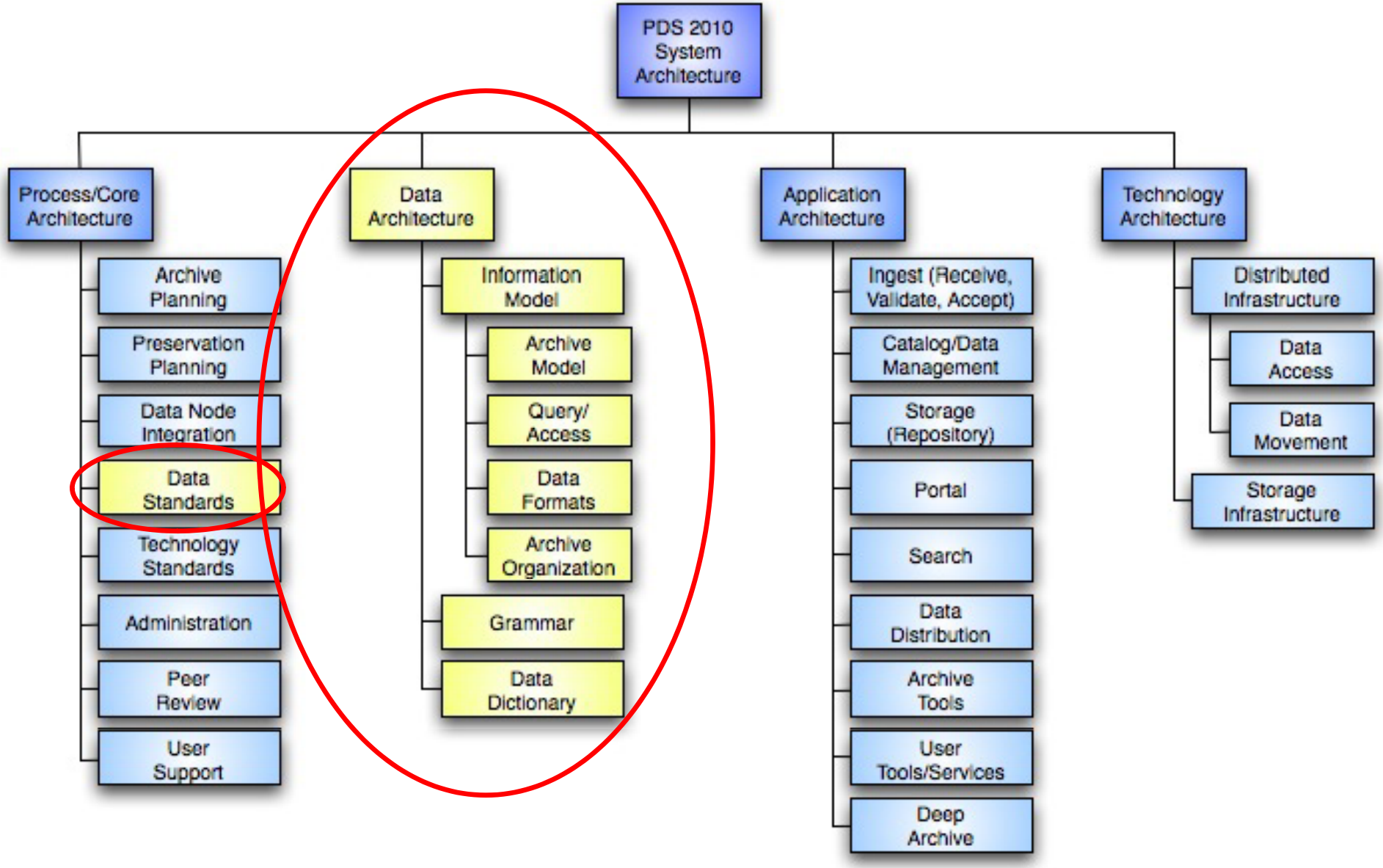
* An information Object is comprised of a data object and descriptive metadata. – OAIS RM

Context

- The PDS 2010 Reference System Architecture has four components.
  - Process Architecture
  - Data Architecture
  - Technology Architecture
  - Application Architecture

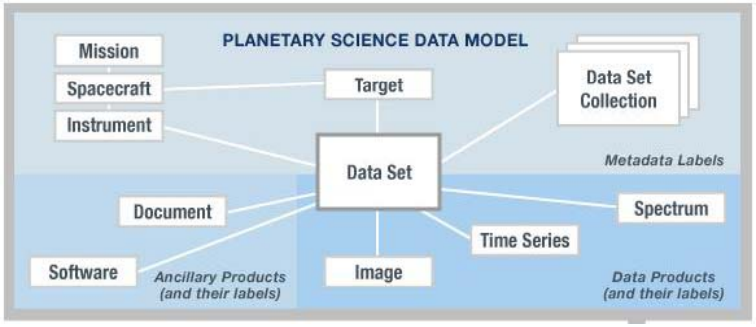
- The Data Architecture is a set of data standards for a planetary science archive data system
  - It guides system design, implementation and operations

PDS 2010 Architecture



Data Architecture Concepts

Information Model



Used to Create

Label Schema

Validates

Expressed As

Extracted/Specialized

Data Element

has

Class

Planetary Science Data Dictionary

Product

Described 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

Data Architecture Drivers

- More Complexity
  - Extensible to handle more complex data and associated information
- More Data
  - Manage, package, and partition large volumes of data
- Greater User Expectations
  - Provide users with well-documented data in a variety of formats
- Create a "System" from the Federation
  - Logically integrate several loosely-couple and geographically distributed systems.
  - Provide location independence
  - Maintain local governance

International

"The data standards within the IPDA, including the data models and derived dictionaries, are based on the NASA Planetary Data System (PDS) standard that is the de-facto standard for all planetary data at the time of the IPDA founding".

Charter of the International Planetary Data Alliance,
3rd Draft, May 24, 2007

PDS3 Issues

- The information model was never formally defined.
- It is riddled with ambiguity and inconsistencies.
  - More than 60 issues were identified during the analysis of the PDS3 standards.
- There are no defined data structures.
- The class hierarchies are not consistent.
- A weak grammar is used to express the model.
- Software solutions are required to fill in the gaps.
- The current data dictionary meets less than half of the new requirements.
- We tried to fix it. There is no simple fix and attempting to will be harder than starting afresh.

Level 2 and 3 Requirements Applicable to Data Standards

1.4 Archiving Standards: PDS will have archiving standards for planetary science data

1.4.1 PDS will define a standard for organizing, formatting, and documenting planetary science data

1.4.2 PDS will maintain a dictionary of terms, values, and relationships for standardized description of planetary science data

1.4.3 PDS will define a standard grammar for describing planetary science data

1.4.4 PDS will establish minimum content requirements for a data set (primary and ancillary data)

1.4.5 PDS will, for each mission or other major data provider, produce a list of the minimum components required for archival data

2.3 Validation: PDS will validate data submissions to ensure compliance with standards.

2.3.1 PDS will develop and publish procedures for determining syntactic and semantic compliance with its standards

2.6 Catalog: PDS will maintain a catalog of accepted archival data sets.

2.6.1 PDS will develop and publish procedures for cataloging archival data

2.6.2 PDS will design and implement a catalog system for managing information about the holdings of the PDS

2.6.3 PDS will integrate the catalog with the system for tracking data throughout the PDS

3.1 Search: PDS will allow and support searches of its archival holdings

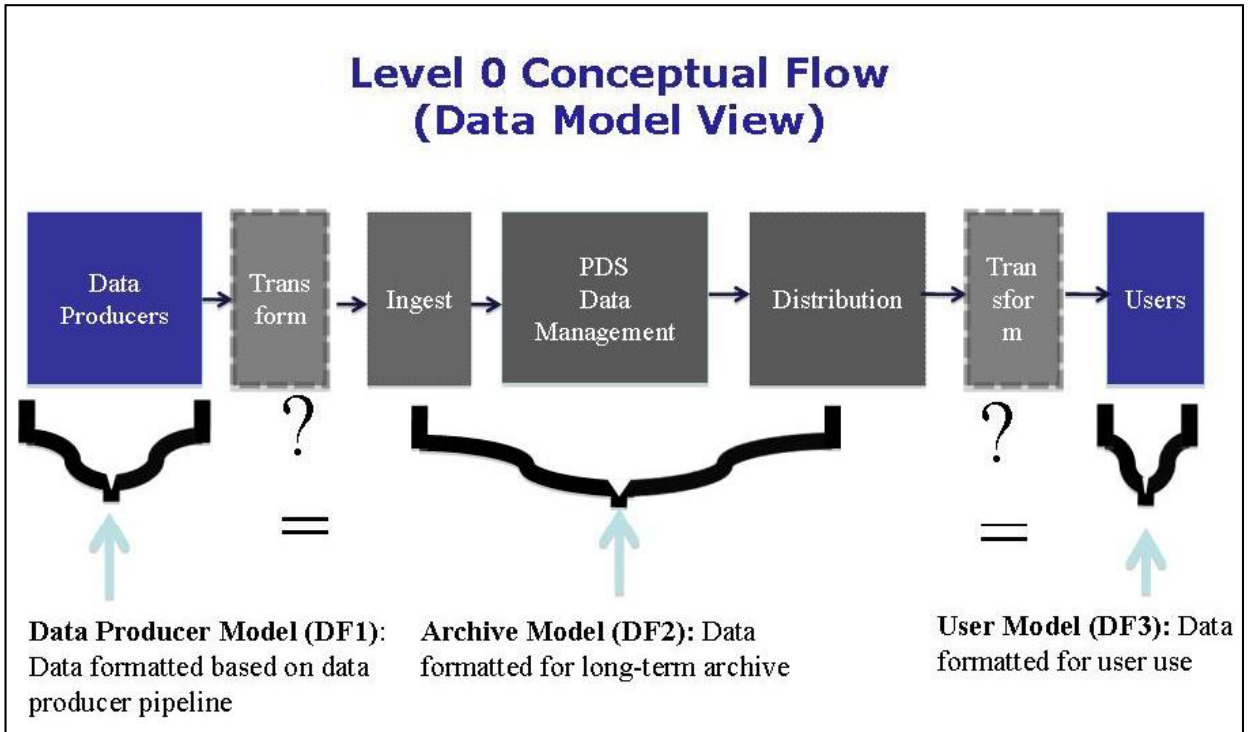
3.1.2 PDS will develop and maintain online interfaces for discipline-specific searching

3.2 Retrieval: PDS will facilitate transfers of its data to users

3.2.1 PDS will develop and maintain online mechanisms allowing users to download portions of the archive PDS4 Data Model Requirements

Objectives 1

- Enable a stable and usable long-term archive.
- Enable more efficient archive preparation for data providers.
- Enable services for the data consumer to find the specific data they need and provide the formats they require.



Objectives 2

- Simplified Data Formats
- Long-term Stability in the Archive
- Efficient Archive Preparation for Data Providers
- Efficient Data Service Development
- Enhanced Data Dictionary

Recommendations to MC (in August 2009)

- Replace PDS3 ad hoc information model with a PDS4 information model that is managed using modern tools
- Replace ad hoc PDS3 product definitions with PDS4 products that are defined in the model
- Require data product formats to be derivations from a core set; Support transformation from the core set
- Replace "homegrown" PDS data dictionary structure with an international standard (ISO 11179 RIM)
- Adopt a modern data language/grammar (XML) where possible for all tool implementations

Principles

- Data Stewardship
  - Define a data architecture that is unambiguous and well-documented
  - Promote well-described, self-contained data sets
  - Provide data formats that are easy to understand and transform
- Data Driven
  - Use the data architecture to guide system development
- Common Vocabulary and Data Definitions
  - Use a standard data dictionary model
  - Federate the data dictionary

Design Approach - Model

- Design and manage the information model in a data modeling tool.
  - The model is formally defined.
  - The model can be validated and tested.
- Define a few simple fundamental data structures.
  - Fundamental data structures may be extended and combined to form more complex data formats
- Use a data driven methodology.
  - Disentangles the model from its implementation.
  - Model can evolve over time as domain changes.
  - Automatic generation of documentation, label schema, and other development artifacts.
- Leverage existing standards.

Design Approach - Dictionary

- Design one dictionary with the authority for each component delegated to a node or a mission.
  - Support for intra-mission cross-correlation
  - Support for intra-node cross-correlation
  - Removes requirement for PDS-wide review of mission-specific keyword desires
- Will integrate with international approach

Topics

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

Design - Key Features of Information Model

- Four base formats for all archived information
- Physical data segments map directly to logical segments
- Documents, software and ancillary data treated as rigorously as observational data
- Keyword content sorted into independent classes
- Data Product Centric
  - Data products exist in a defined context
  - Data products are registry objects

Model Design Decisions

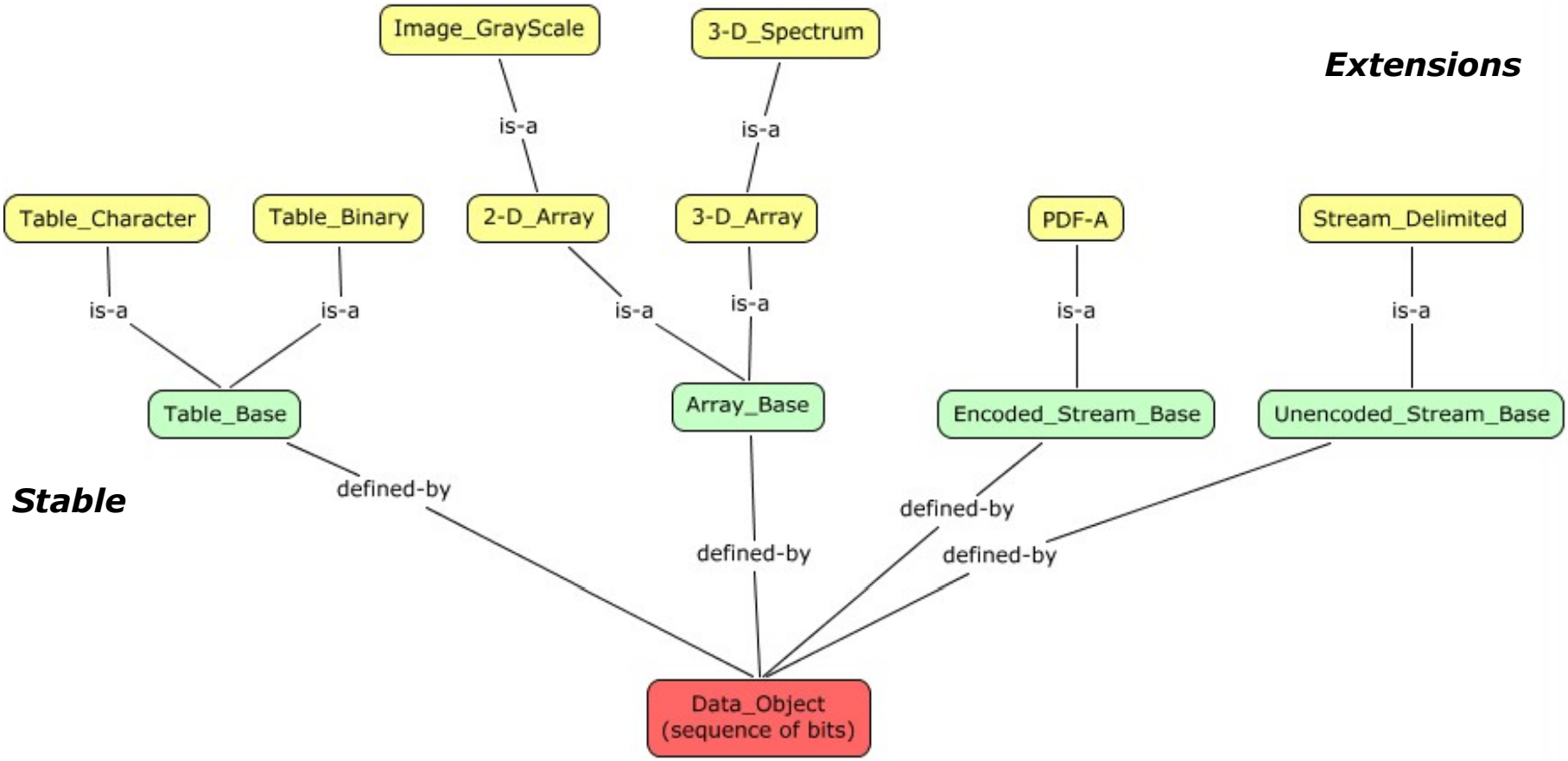
- Use a formal data modeling methodology to define the Data Model
- Rigorously define the data product model
  - Define core data structures
  - Derive new PDS4 "objects"
  - PDS3 "objects" will not be migrated forward
- Make use of class hierarchies for extensibility
- Maintain the Data Model independent of any implementation
- Address issues with the catalog objects and other area of the PDS standards. E.g. Data Set, Volume, Target, Repository, …
- Replace the Object Description Language (ODL)

Base Formats

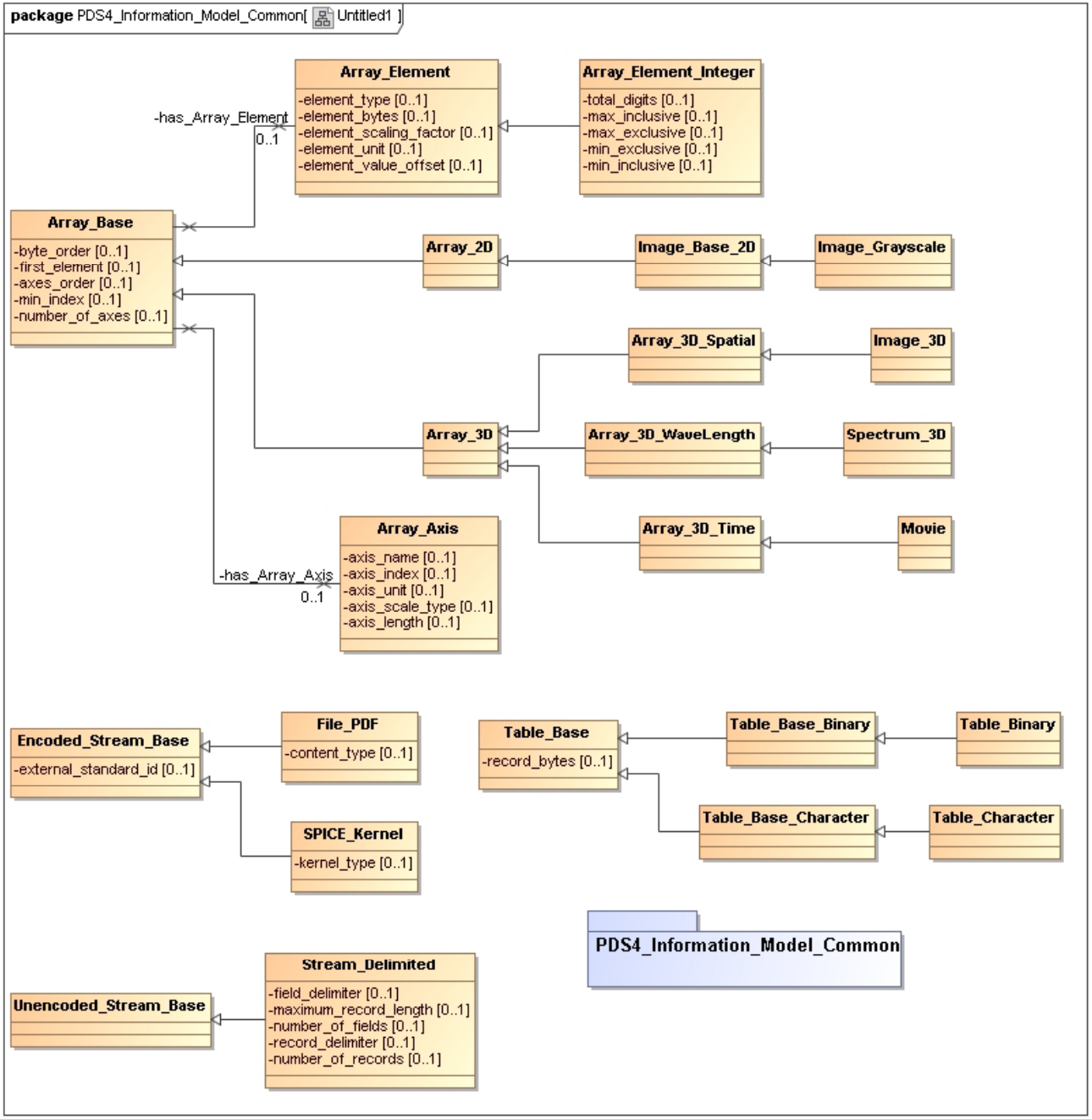
All the data we deal with can be broken down into one or more of the base formats.

- Arrays
- Tables
- Parseable byte streams
- Encoded files

Base Formats and Extensions



UML Class Diagram for Base Formats



Physical to Logical Mapping

This means no physical interleaving of logically disjoint sections of the data.

- Enhanced archive stability
- Efficiency in our own tool/utility programming

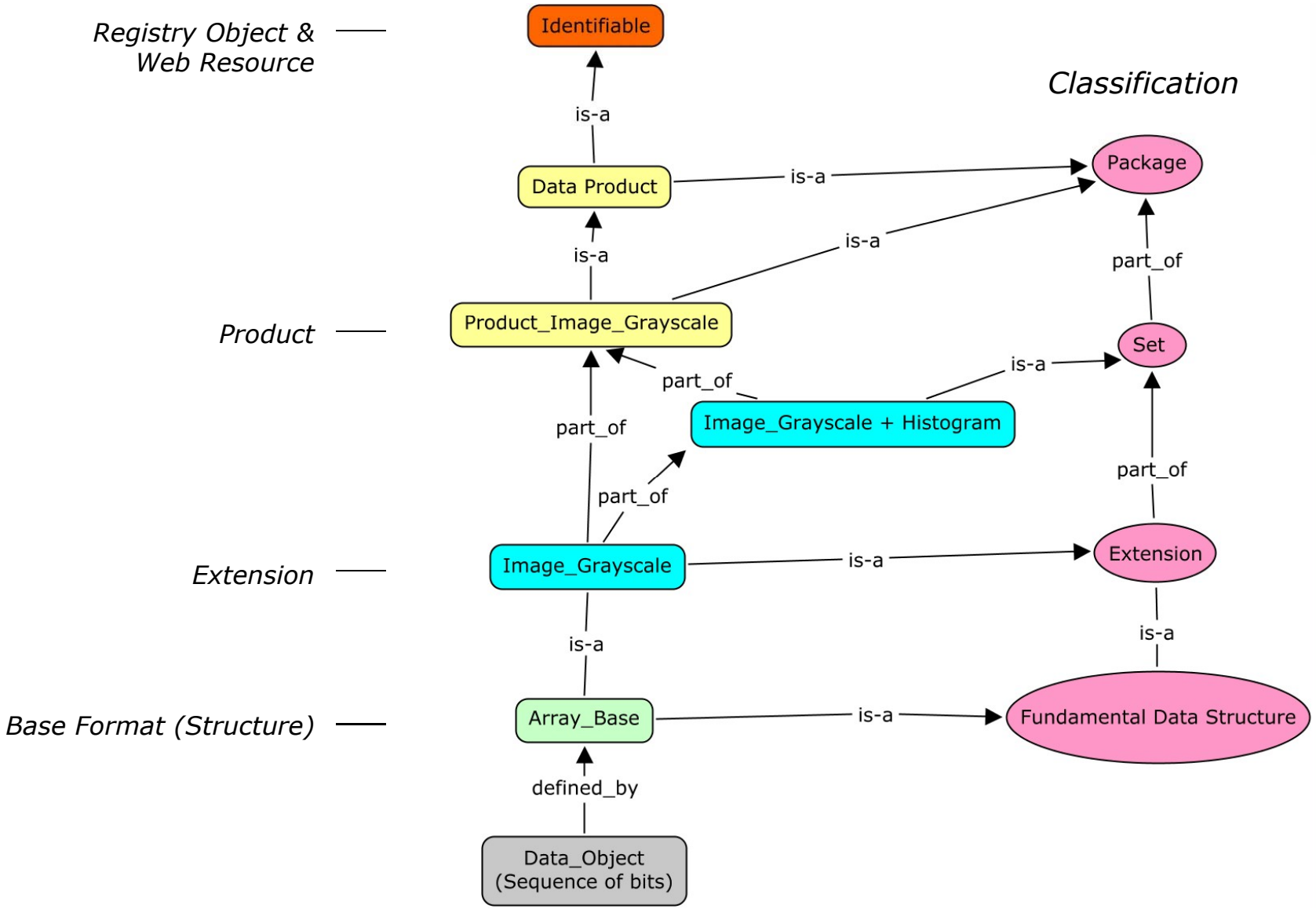
Note that this does not require bit manipulation.

All Products Are Equal

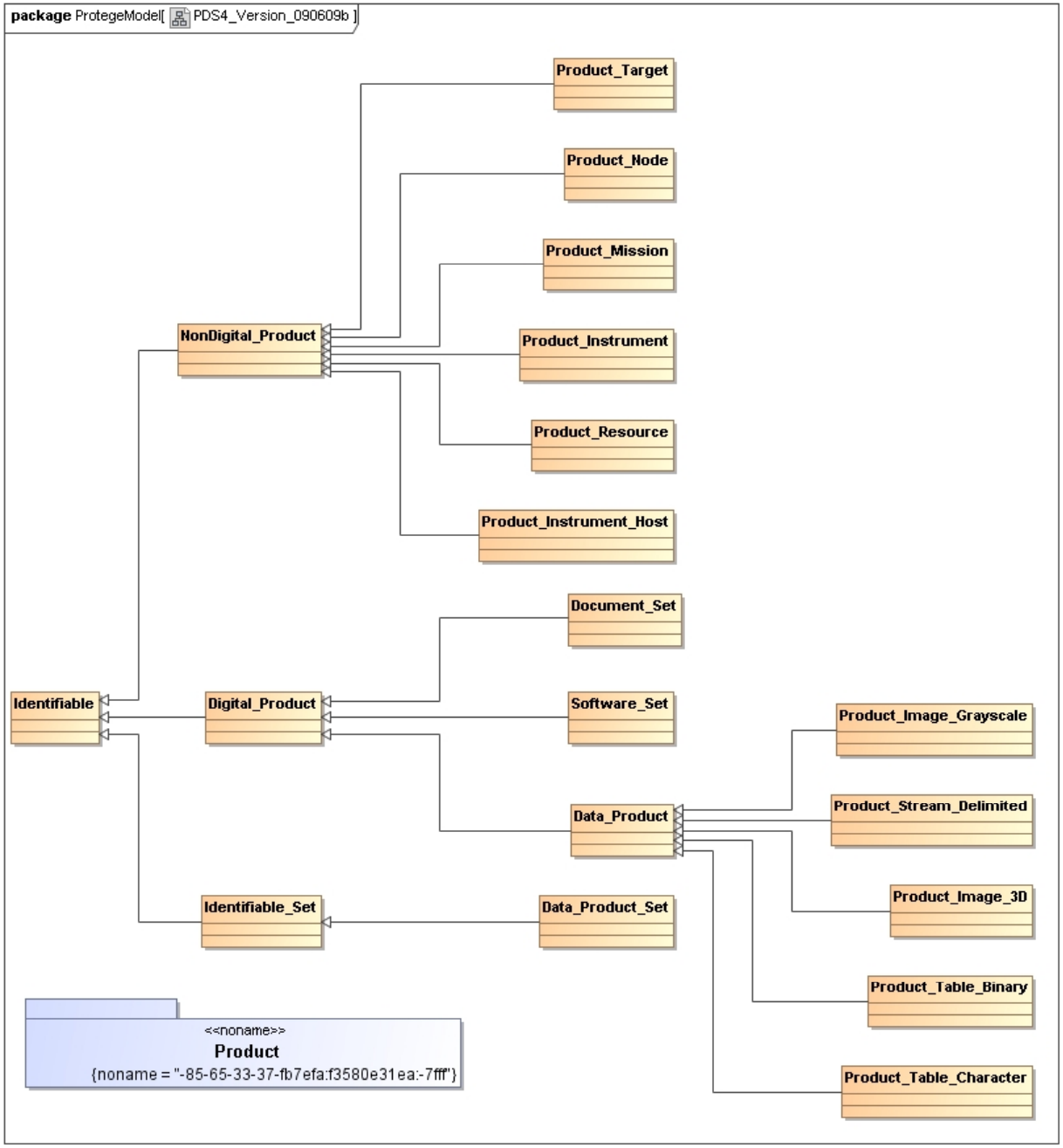
All products are treated with equal rigor in labelling and documenting.

- Ensures the ability to cross-reference throughout the archive holdings
- Supports interface selection and packaging options for users
- Necessary for tracking and processing formats that may require migration in future

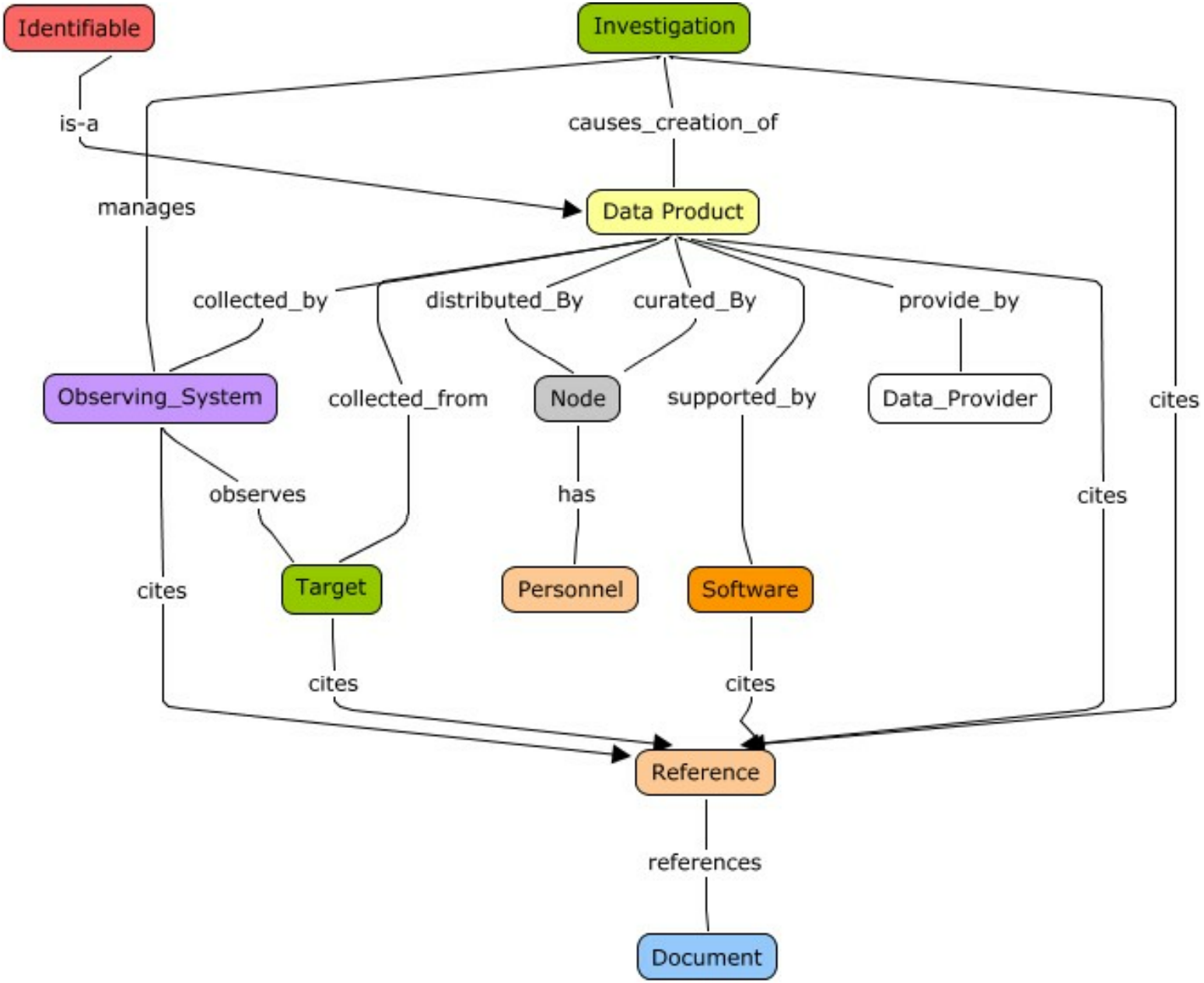
Data Product



Products


package ProtegeModel[PDS4_Version_090609b]
Product_Target
Product_Node
Product_Mission
NonDigital_Product
Product_Instrument
Product_Resource
Product_Instrument_Host
Document_Set
Identifiable
Digital_Product
Software_Set
Product_Image_Grayscale
Data_Product
Product_Stream_Delimited
Product_Image_3D
Identifiable_Set
Data_Product_Set
Product_Table_Binary
<<noname>>
Product
{noname = "-85-65-33-37-fb7efa:f3580e31ea:-7fff"}
Product_Table_Character

Data Product in Context



Backup

PDS4 Information Model Requirements

The PDS4 Information Model requirements have been derived from the PDS Level 3 Requirements and additionally compiled as an artifact of the PDS3 Information Model review.

1. The Information Model shall be developed and maintained independent from any specific
technology choices, implementations, or expressions.

2. The Information Model shall be defined using a formal data modeling notation.

3. The Information Model shall encompass all PDS stakeholder viewpoints, including the
contextual, conceptual, logical and physical.

4. The Information Model shall be rigorous and prescriptive.

5. The Information Model shall allow the definition and inclusion of data models from many domains.

6. The Information Model shall formally define “context” classes.

PDS4 Information Model Requirements

7. The Information Model classes shall formally define relationships between classes.

8. The Information Model shall define a standard set of “attributes” and “attribute values” for describing data maintained in the PDS archive. [PDS 1.4.2]

9. The Information Model shall define a set of data formats that are based on standard data structures. [PDS 1.4.2]

10. The Information Model shall be tightly coupled with a data dictionary to capture information that is not captured within the Information model. [PDS 1.4.2]

11. The Information Model shall support long term preservation (e.g., archiving) of the data in the PDS archive. [PDS 4.1]

12. The Information Model shall support long term use of the data in the PDS archive. [PDS 4.2]

13. The Information Model shall support distributed discovery and access to the data in the PDS archive. [PDS 2.8.1; PDS 2..8.2; PDS 2.8.3; PDS 3.1]

14. The Information Model shall support cataloging the holdings in the PDS archive. [PDS 2.6.2]

PDS4 Information Model Requirements

15. The Information Model shall support tracking of the data in the PDS archive. [PDS 2.2.2; PDS 2.4.5]

16. The Information Model shall be maintained in accordance with the PDS Data Standards process. [PDS 1.4.6]

17. The Information Model shall formally define the following classes: Data Set, Product, Mission, Instrument, Host, Target, Node, Person, Reference, Document, and Software.
Note: 1.4.4 PDS will establish minimum content requirements for a data set (primary and ancillary data)

18. The Information Model shall formally define the following classes: data object, data structure, data interpretation, data identification, and data metadata.
Note: 1.4.1 PDS will define a standard for organizing, formatting, and documenting planetary science data

19. The Information Model shall formally define a data dictionary model.
Note: 1.4.2 PDS will maintain a dictionary of terms, values, and relationships for standardized description of planetary science data

20. The Information Model shall have one or more grammars into which to express the information model classes.
Note: 1.4.3 PDS will define a standard grammar for describing planetary science data

PDS4 Information Model Requirements

21. The Information Model shall formally define the following classes: resource, release, and
housekeeping.
Note: 2.6.2 PDS will design and implement a catalog system for managing information about the holdings of the PDS – Assume that classes derived from 1.4.4 also support 2.6.2.

22. The Information Model shall formally define the following classes: repository, registry,
and identifiable.
Note: 3.2.1 PDS will develop and maintain online mechanisms allowing users to download portions of the archive

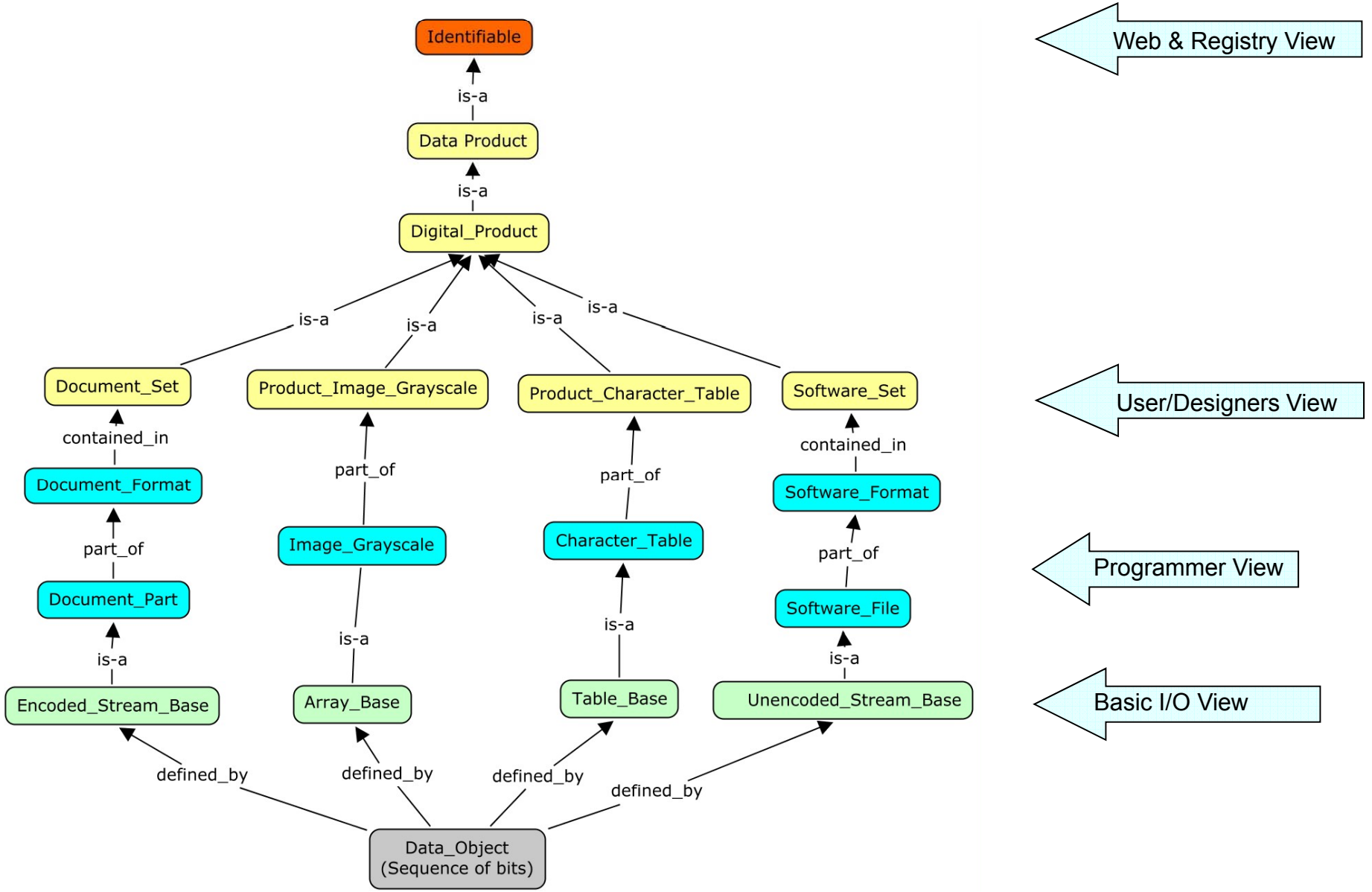
23. The Information Model shall formally define the following classes: coordinate system.
Note: 3.3.4 PDS will provide tools for translating archival products between selected coordinate systems

24. The Information Model shall formally define the following classes: manifest.
Note: 4.1.2 PDS will develop and implement procedures for periodically ensuring the integrity of the data.

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

Data Product Concept Map



Principles 2

- The information model is defined using a formal modeling tool
- The information model is independent of the implementation
- The information model is extensible enabling it to handle more complex data.