National Aeronautics and
Space Administration

PDS 4 Data Architecture Report

PDS 2010 Data Architecture WG

November 20, 2008

National Aeronautics and Space Administration

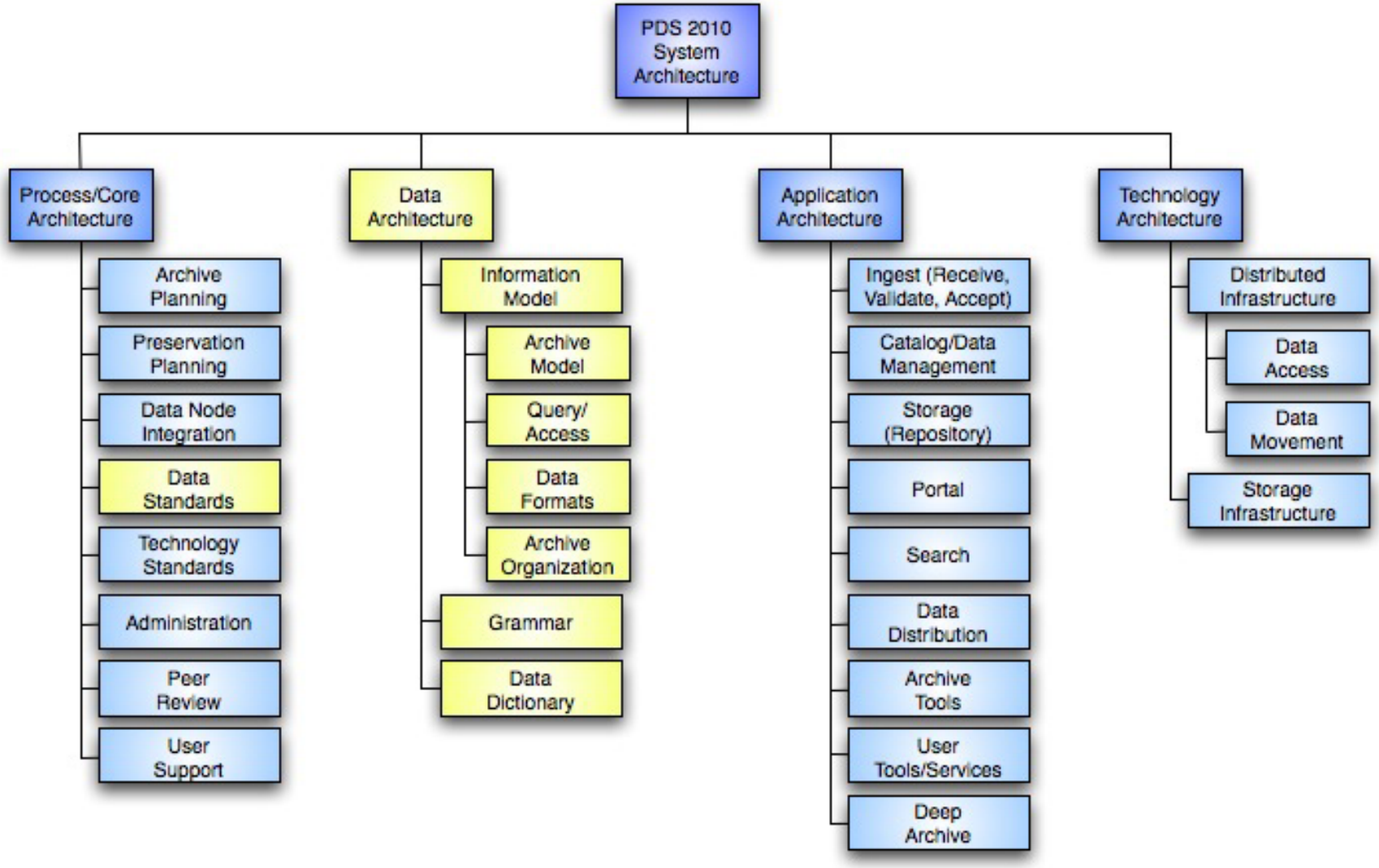
Background

- **At the July '08 Management Council (MC) meeting it was determined that the PDS3 Specification WG Report satisfies the action item that the PDS3 "data model" be explained by the Tech Group to MC**
  - The PDS3 Information Model Specification document was forwarded to the PDS 2010 project manager for disposition and the WG was disbanded

- **At the same meeting the charge was given to form a PDS4 Data Architecture Working Group (DAWG) to lay out the high level architecture and provide recommendations**

National Aeronautics and Space Administration

Data Architecture Scope



National Aeronautics and Space Administration

Response to Charge

- **Formed Data Architecture Working Group (DAWG)**
- **Identified architectural drivers and principles**
- **Used the PDS3 specification as input**
  - Identified the core elements of PDS3
  - Compiled PDS3 problems, issues, and anomalies from the PDS3 specification and PDS4 SCRs
  - Proposed fixes, estimated impact, and set priorities
- **Identified gaps in PDS3 requirements**
- **Proposed architectural recommendations**
- **Identified fundamental questions that require MC guidance**
  - Presented at tech session for review and comment
  - Have continued to clarify the issues with the help of the Tech Group

National Aeronautics and Space Administration

Architectural Principles - Synopsis

- **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
- **Model Driven**
  - Use the data architecture to guide system development
- **Common Vocabulary and Data Definitions**
  - Uses a standard data dictionary model
  - Allow the federation of the data dictionary

National Aeronautics and Space Administration

Key Architectural Drivers - Synopsis

- **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**
  - Support location independence

National Aeronautics and Space Administration

Findings – PDS3 Specification Work

- **The core concepts of the PDS3 standards and years of lessons-learn provide a good basis for designing PDS4.**
  - Examples of core concepts include:
    - Product is a package of “objects” that describe science data.
    - Data set is a collection of products and supporting information.
    - Data sets and products are given context through their association with Instruments, Missions, Targets, Nodes, ...
  - The standards are flexible.
  - The archived metadata is human and machine readable.
  - Science discipline experts are intricately involved in the development and evolution of the standards.
  - In general, the PDS was a pioneer in the development of science data archives.

National Aeronautics and Space Administration

Findings – PDS3 Specification Work

- **The WG has compiled 74+ anomalies, problems, issues, and items of note from the PDS3 specification work, gaps in the PDS requirements, PDS4 SCRs, and PSA input.**

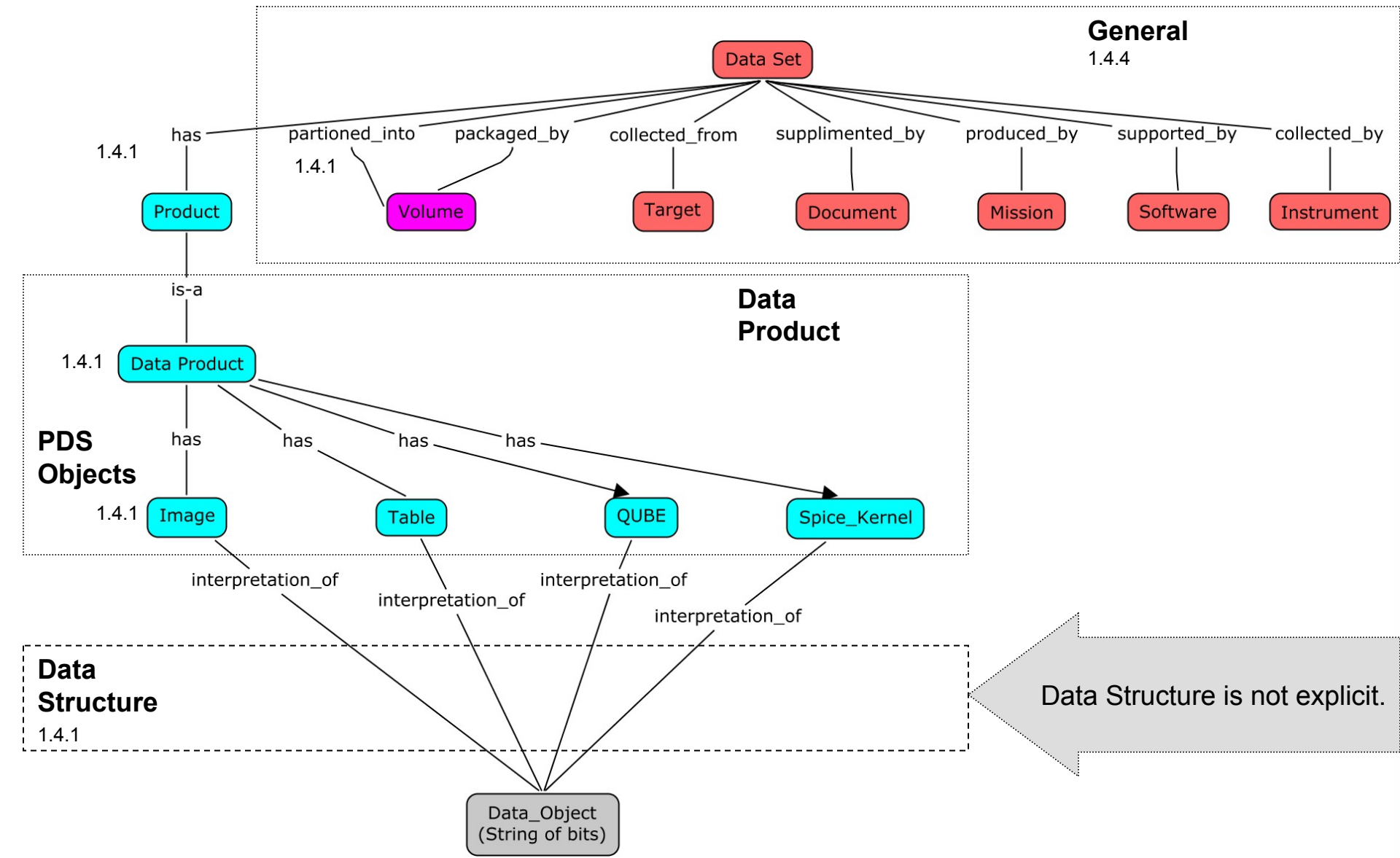
- **These items have been grouped into four categories.**
  - General (Methodology, Catalog, Operational, …)
  - Data Products
    - Data Structures
  - Data Dictionary
  - Object Description Language (ODL)

- **A key issue identified is the lack of explicitly defined data structures.**
  - PDS3 “objects” *describe* data structure but do not *define* data structure.
  - For example a PDS Image object provides the number of lines and line samples in a simple image but the 2 dimensional structure must be assumed. Typically it is the omission of the BANDS keyword.

National Aeronautics and Space Administration

PDS Data Model

Conceptual View



National Aeronautics and Space Administration

Findings – General (Methodology, Catalog, …)

- **Problem Summary**
  - The PDS data model in general is ambiguous and is riddled with assumptions.
  - Generally accepted data engineering principles are not being applied.
- **Proposed Fix**
  - Capture and manage the data model in a modern data engineering tool. *Done*
  - Fix or replace the PDS "objects" that have problems. E.g. Volume
  - Fix miscellaneous problems.
    - Use unique identifiers; Link "objects" that should be linked
  - Maintain the PDS Data Model independent from any implementation.
- **Source**
  - 36+14 Issues – PDS3 Spec and L3 Gaps, PDS4 SCRs
  - 6+ issues are associated with Volume alone
- **Impact**
  - Fixing any specific "object" (e.g. Volume) would affect labels, tools, and documentation and have impact on users, data providers and the system as a whole.

National Aeronautics and Space Administration

Findings – Data Product

- **Problem Summary**
  - The PDS3 data product(s) is not explicitly defined.
  - The PDS3 data structures are not explicitly defined.
  - The PDS3 “objects” in general are ambiguous, have too many assumptions, and are tightly bound to ODL.
- **Proposed Fix**
  - Capture and manage the data product model in a data engineering tool.
  - Fix the problematic “objects” and other outstanding issues
  - Capture the data structures that are implicit in the archive by factoring out the data structures used in the Object Access Library (OAL).
- **Source**
  - 28+14 Issues – PDS3 Spec and L3 Gaps, PDS4 SCRs
  - E.g. “either eliminate or replace the implicit File object”
- **Impact**
  - Changes to almost any “object” would affect product labels, tools, and documentation and have some impact on PDS users, data providers, and the system as a whole.
  - The definition of a set of data structures for all products in the archive would be almost impossible.
  - A set of “factored out” data structures would make some percentage of the PDS3 archive non-compliant.

National Aeronautics and Space Administration

Findings – Data Structure

- **Problem Summary**
  - The PDS standards do not explicitly define a set of data structures.
- **Proposed Fix**
  - Define a set of Data Structures
  - Derive a set of Data Formats (i.e. PDS “objects)
- **Source**
  - 1.4.1 PDS will define a standard for organizing, formatting, and documenting planetary science data
- **Impact**
  - A new set of “objects” would affect product labels, tools, and documentation and have impact on PDS users, data providers, and the system as a whole.
  - A significant percentage of the PDS3 archive will be made non-compliant.

National Aeronautics and Space Administration

Findings – Data Dictionary

- **Problem Summary**
  - The data dictionary has limited capability and many issues.
- **Proposed Fix**
  - Adopt a new data dictionary model
  - Migrate the data dictionary content
  - Clean up the data dictionary content
- **Source**
  - 9+8+10 Issues – PDS3 Spec and L3 Gaps, PDS4 SCRs, PSA input
  - E.g. “The PSDD does not capture relationships between data elements such as "has similar meaning" and "has similar valid values."
- **Impact**
  - A new data dictionary model would affect the data dictionary, related tools, services, and documentation.
  - Change to the data dictionary model should have minimal impact on users and data providers.

National Aeronautics and Space Administration

Findings – ODL

- **Problem Summary**
  - The Object Description Language (ODL) has issues and limited capability.
- **Proposed Fix**
  - Fix or replace ODL.
  - An additional grammar makes sense for operational purposes. (e.g. XML)
- **Source**
  - 2+3 Issues – PDS3 Spec and L3 Gaps, PDS4 SCRs
  - E.g. “When are ODL “Groups” appropriate?

    “To quote or not to quote”
- **Impact**
  - Fixing ODL would affect product labels, tools, and documentation and have impact on users, data providers and the system as a whole.
  - Depending on where it is used, a new language such as XML could have significant impact.

National Aeronautics and Space Administration

Recommendations for PDS4 Part 1

- **Fix the identified problems, issues, and anomalies in the *General, Data Dictionary*, and *ODL* categories.**

- **Examples**
  - Use a formal data modeling methodology to define the Data Model
  - 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, …
  - Fix the data dictionary model
  - Fix or replace ODL

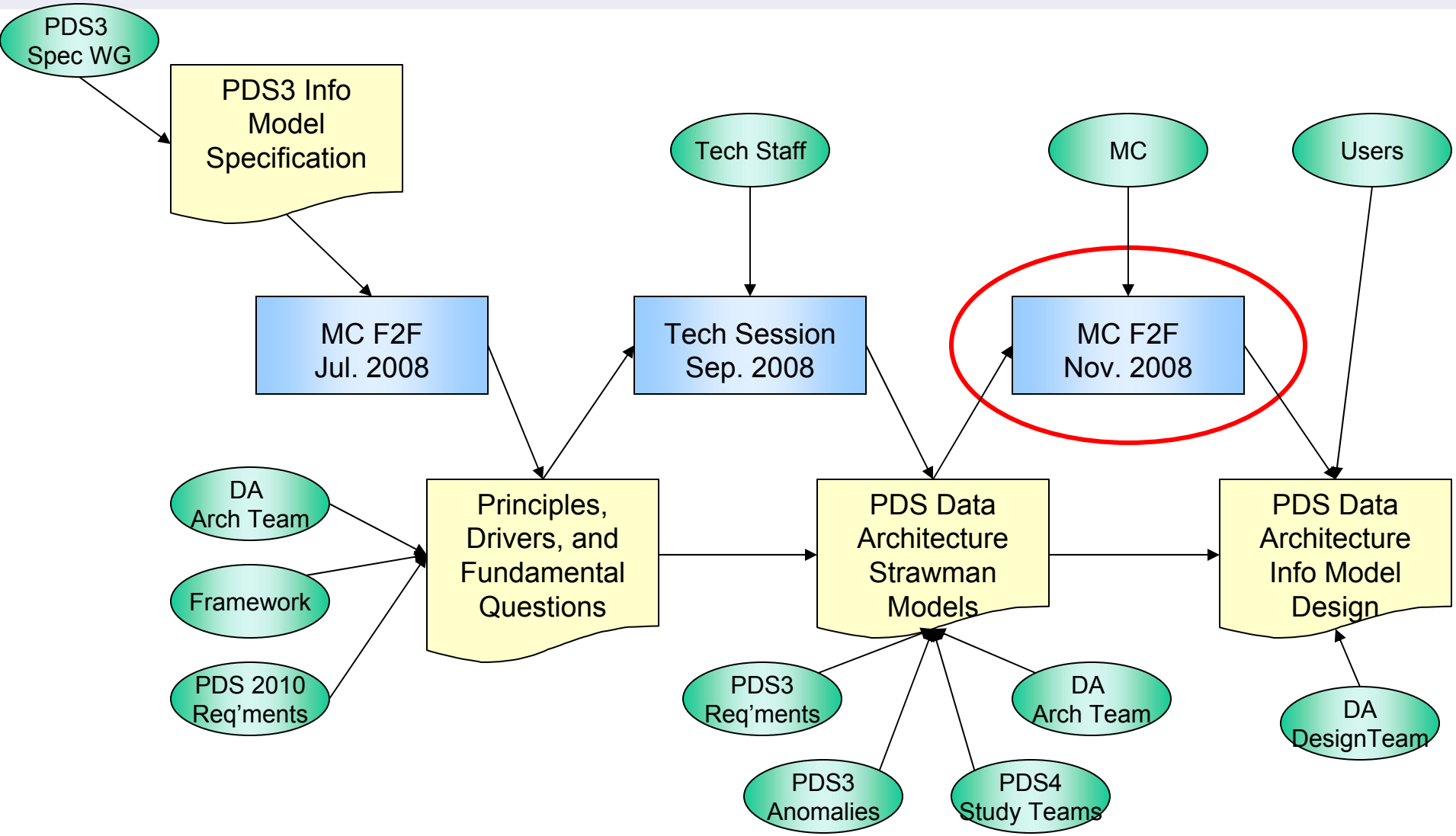
National Aeronautics and Space Administration

Recommendations for PDS4 Part 2

- Rigorously define the data product model
- A strawman data model has been started and will be elaborated on in the next talk

National Aeronautics and Space Administration

Road Map – Architecture to Design



National Aeronautics and Space Administration

Next Steps

- **Management Council**
  - Provide guidance on fundamental questions (Mitch to elaborate)
  - Provide guidance on design decisions

- **Form design team to implement architectural recommendation**
  - Follow PDS 2010 project structure.
  - Design and implement PDS4.
    - Standards Reference V4.x
    - Data Model Specification
    - New Data Dictionary
    - Consistent Grammar

National Aeronautics and Space Administration

Open Issues

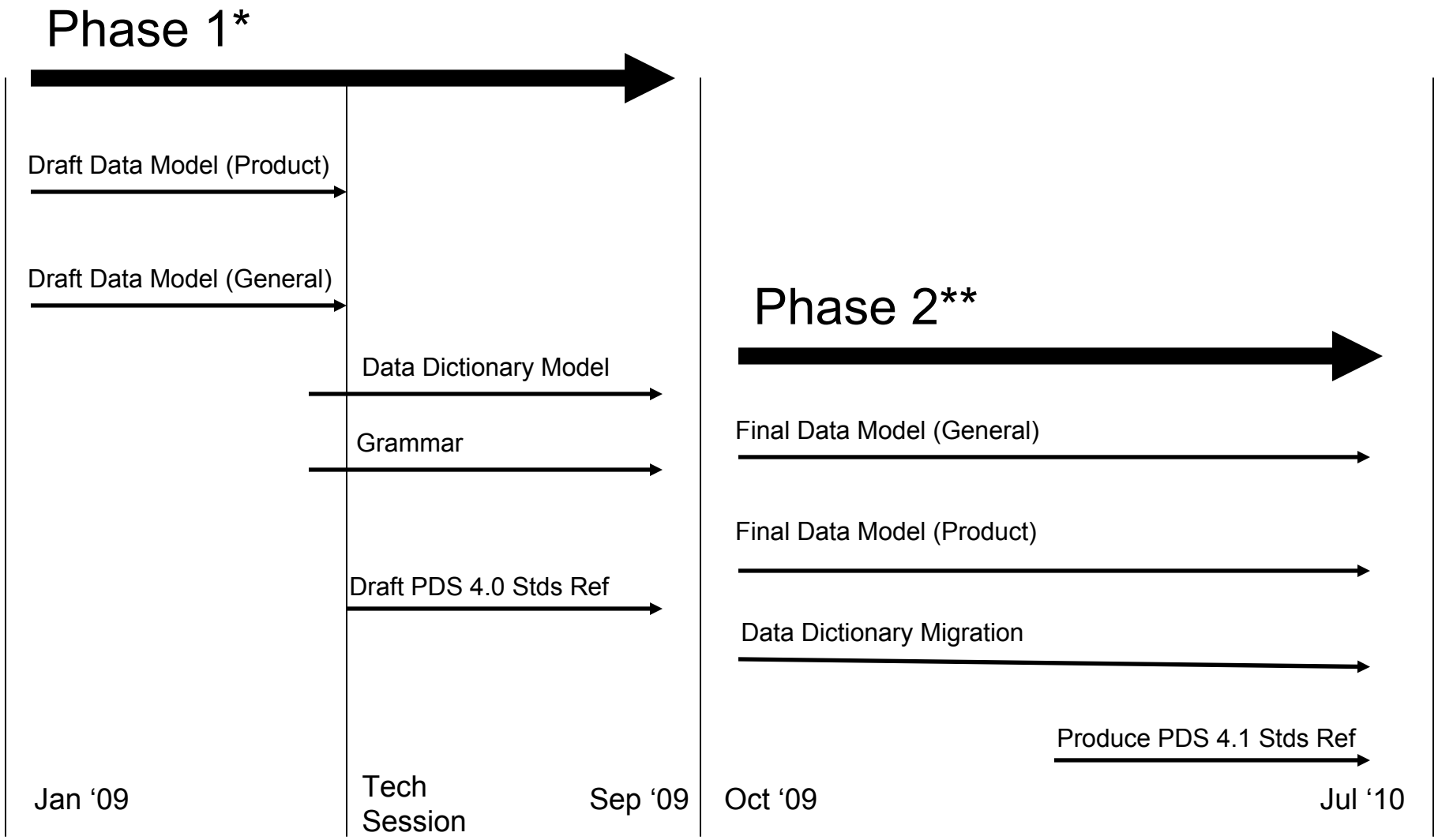
- **What about data migration?**
  - Some thoughts
    - Legacy data sets remains permanently in archive.
    - The appropriate migration approach will vary from node to node and data set to data set.
    - For example, repository gateways can be developed to transform data sets on request.

National Aeronautics and
Space Administration

Backup

National Aeronautics and Space Administration

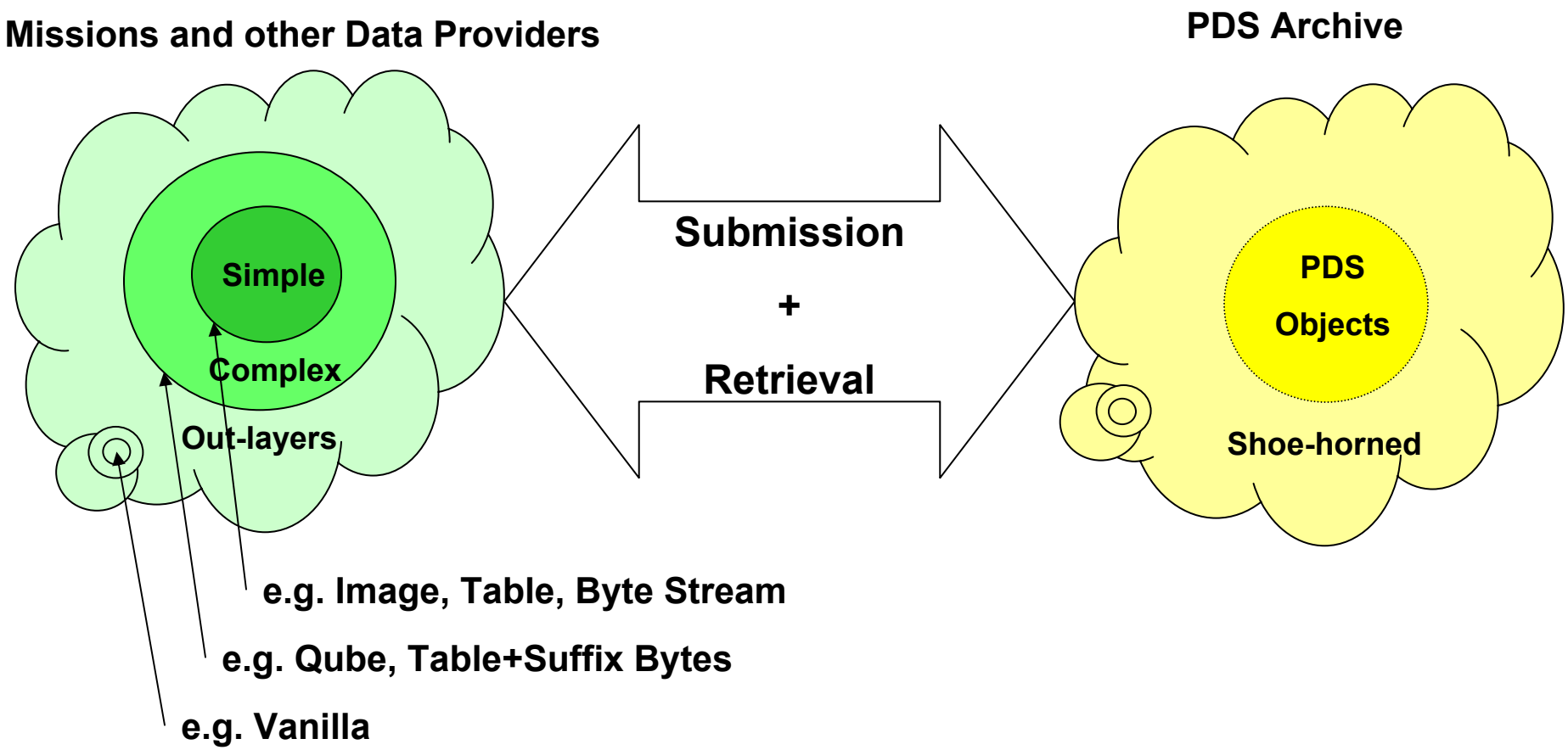
Data Standards Project Phases



*WG Design **PDS Node and Community Review and Implementation

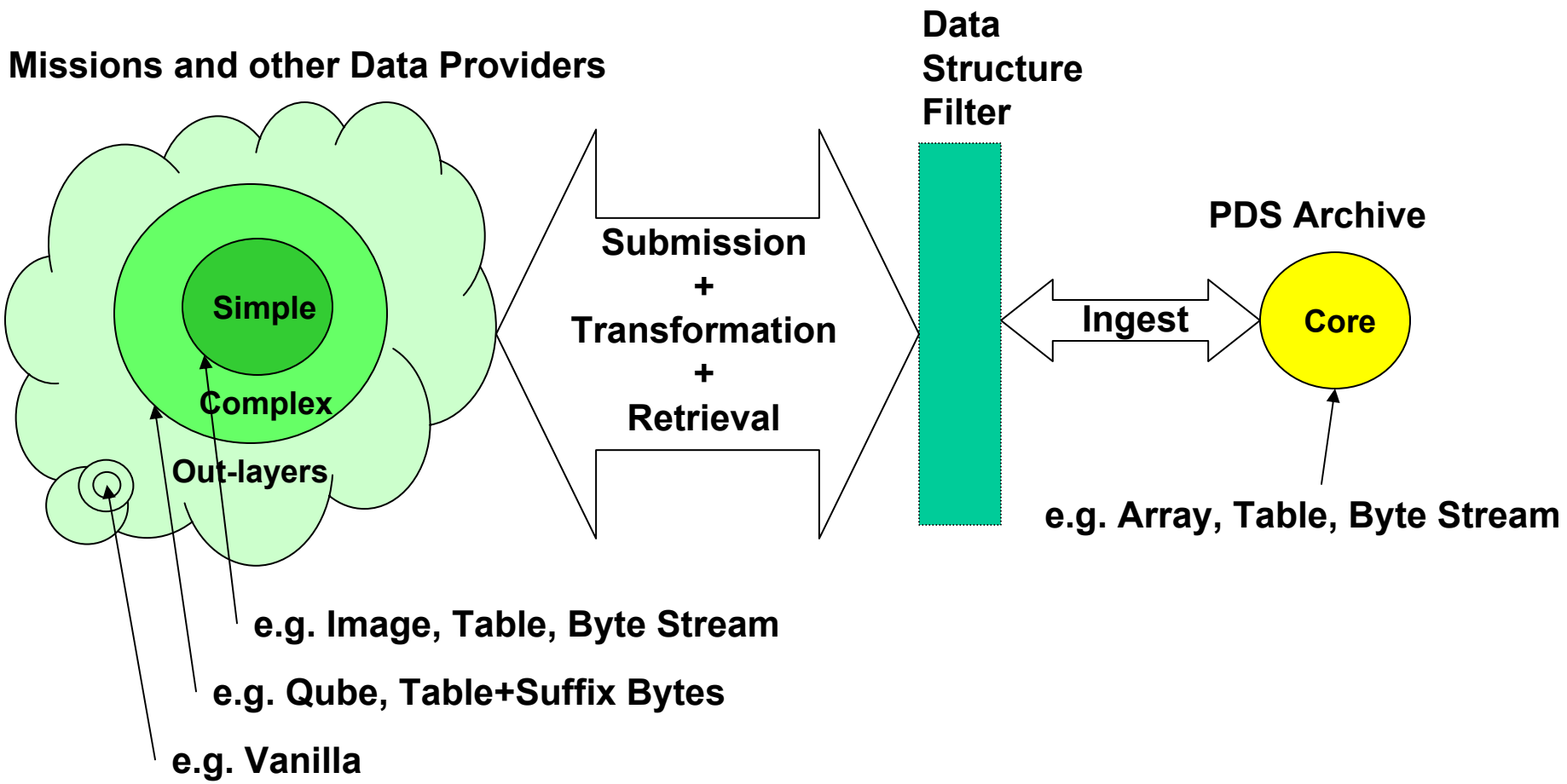
National Aeronautics and Space Administration

PDS3 Archive



National Aeronautics and Space Administration

PDS4 Archive



National Aeronautics and Space Administration

Image Object Labels

PDS3

```
OBJECT                  = IMAGE
 LINE_SAMPLES           = 256
 LINES                  = 256
 SAMPLE_BITS            = 32
 SAMPLE_TYPE            = "IEEE_REAL"
 AXIS_ORDER_TYPE        = "FIRST_INDEX_FASTEST"
 LINE_DISPLAY_DIRECTION       = "UP"
 SAMPLE_DISPLAY_DIRECTION  = "RIGHT"
 UNIT                   = "W/[M**2 SR UM]"
 MINIMUM                = -1.04848e-004
 MAXIMUM                = 1.82296e-002
 MEDIAN                 = 1.33682e-005
 STANDARD_DEVIATION = 1.30217e-004
END_OBJECT              = IMAGE
```

- **Image "object" is derived from Array**
- **Dimension is explicit**
- **Location information is local**
- **Optional attributes are distinct**

PDS4

```
class PDS4.2D_Image
    { data_location = [("MV0168757…_R.FIT",48960)];
      name          = "MRI Image";
      number_of_axes => 2;
      axis_length    = (256,256);
      axis_name      = ();
      element_type   = "IEEE_REAL";
      element_bytes  = 4;
      element_offset = 0.0;
      element_scaling_factor = 1.0;
      element_unit           = "W m**-2 sr**-1 micrometer**-1";
      property_map           = [("minimum", -1.04848e-004),
                                ("maximum",  1.82296e-002),
                                ("median",   1.33682e-005),
                                ("standard_deviation",
                                                1.30217e-004)];
    };
```

National Aeronautics and
Space Administration

Key Goals

- **Reduce the level of ambiguity in the PDS standards.**
  - Formally capture and document the PDS standards and generate a data model specification document.

- **Fix the identified problems in the data model, redesign the data dictionary, and either fix or replace ODL.**

National Aeronautics and Space Administration

Key impacts – Guidance needed

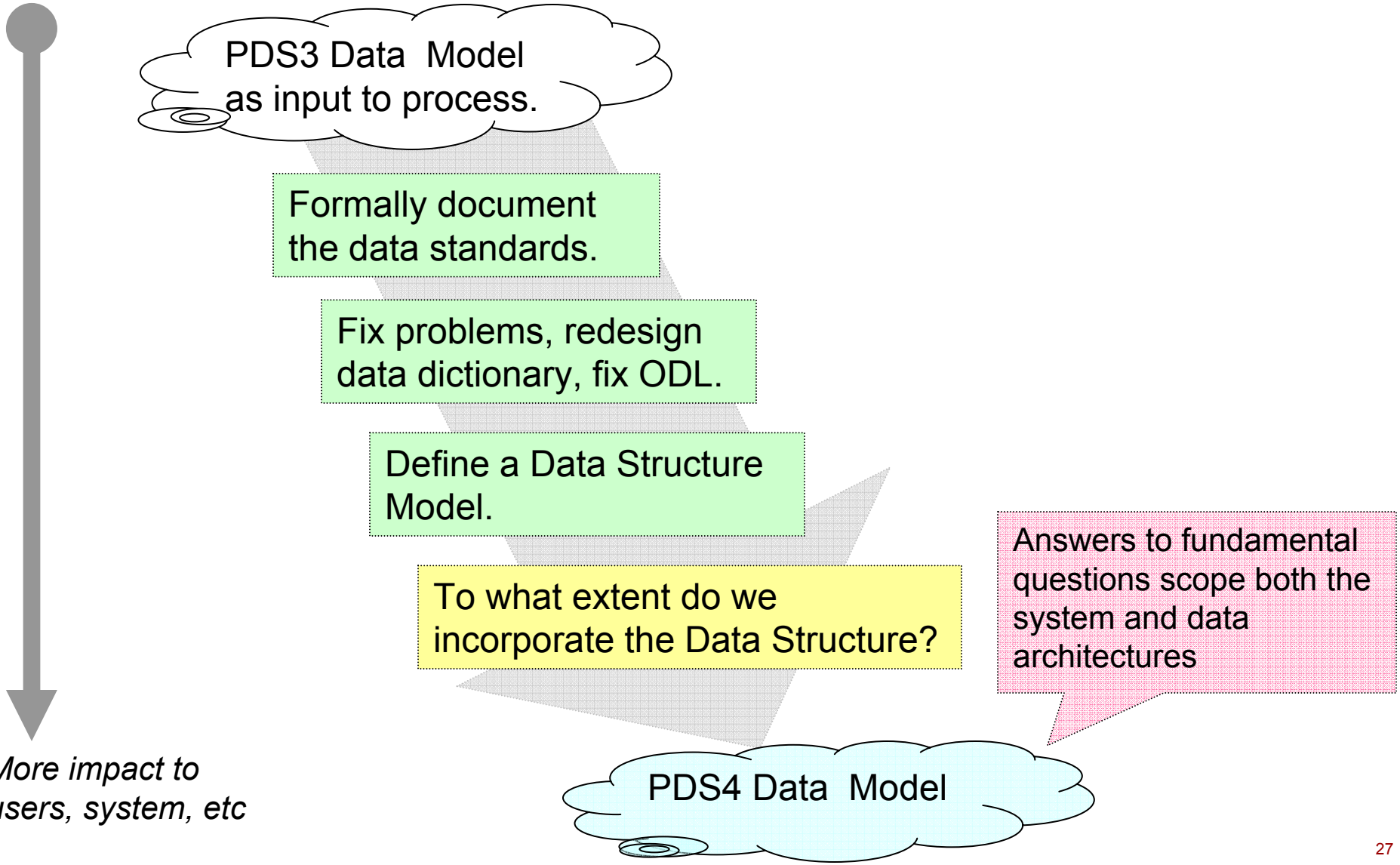
- **Data Transformation**
  - Restricting the archive to a few simple data formats implies that more complex data formats must be transformed before ingestion -- and then transformed before use.

- **Changes to labels, tools, system that affect users and data providers.**
  - Changes in the individual objects (e.g. Image) -- from name change (e.g. 2-D Image, 3-D Color Image) to new more explicit attributes (e.g. axis_length vs lines and line_samples) -- could have significant impact.

  - A new language means that the community must become familiar with a new look-and-feel for product labels.

National Aeronautics and Space Administration

Key Decision Waterfall



National Aeronautics and Space Administration

Data Dictionary

- **Data Dictionary**
  - Data Element Definitions
  - Data Element Relations (Data Element used as Object Attributes)

- **Scope**
  - Discrete groups of data elements
    - Common Data Elements
    - Discipline Data Elements
    - Mission/Instrument Data Elements
    - Ground Data System/Instrument Team Operational Data Elements
    - PDS Operational Data Elements

- **Control Authority**
  - The PDS is the registration authority.
  - Each group has its own submitter/steward.
  - The data dictionary is federated.

- **Data Dictionary Model**
  - Current PDS DD model is very limited
  - Use Standard Data Dictionary Model
    - E.g. ISO/IEC 11179 Metadata Registry Specification
  - Data dictionary is expressed into target languages for specific functions. (e.g. ODL Tool Data Dictionary)

National Aeronautics and Space Administration

Grammar (Language)

- **The language (s) into which we express the PDS Information Model.**
  - E.g. PDS3-ODL, PDS4-ODL, PVL, XML, Other

- **Information Model concepts are expressed into a target language.** (e.g. Image Class -> ODL Object)
  - If two or more languages used, then the mapping is done from the model to each language individually. I.e. We do not derive XML elements from ODL objects

National Aeronautics and Space Administration

Migration

- **Legacy data sets remains permanently in archive.**
- **The appropriate migration approach will vary from node to node and data set to data set.**
- **One approach**
  - Gateways are developed to transform data sets on request
  - New versions of legacy data sets can be ingested.
    - For example frequently used data sets.
  - Non-compatible versions of data sets have to transformed via manual intervention.

National Aeronautics and Space Administration

Principles - Data Stewardship

- **PDS will collaborate with data providers as early as possible in the data creation process to ensure that PDS Standards and tools are adopted and utilized effectively.**
  - Data architecture that is unambiguous and well-documented.
- **Data will be preserved for the long-term. Although the definition of long-term is up for debate, it can be measured in terms of at least 10s of years.**
  - Data architecture that promotes well-described, self-contained data sets.
- **Data will be managed in a way that preserves its meaning and promotes it understanding. This implies that software is available to read and transform the data for use in current day environments.**
  - Data architecture that provides data formats that are easy to understand and transform.

National Aeronautics and Space Administration

Principles - Model Driven

- **The information model will be defined using a formal data modeling notation.**
  - Data architecture is defined clearly and precisely.

- **The data architecture includes a data model that encompasses the PDS.**
  - All things are defined
  - Overlapping Sub models exists and are extracted
    - Data Formats, Archive Model, Query Model, Operations

- **The model will be flexible and extensible**
  - Common things will be defined by consensus for use across the PDS.
    - Minimal extensions are allowed for discipline specific needs.
    - Examples - Data Set, Mission
  - Other things will be partially defined.
    - Maximal extension are allowed for discipline specific and mission needs.
    - Examples - Image, Table

National Aeronautics and Space Administration

Principles - Common Vocabulary and Data Definitions

- **A data dictionary must be established and utilized uniformly throughout the PDS.**
  - Use a common data dictionary model.
  - Allow the federation of the data dictionary.

National Aeronautics and Space Administration

Drivers (1 of 2)

- **More Data - PDS storage requirements are projected to increase from 40 TB to over 500 TB in just three years**
  - Capability to package and partition large volumes of data.

- **More Complexity - Missions, instruments, and data are all becoming more complex.**
  - Capability to efficiently handle more complex data and associated information.

- **More Producer Interfaces - PDS is facing an increasing number of missions, a greater number and diversity of data providers, and smaller, focused missions.**
  - Flexibility to meet the demands of a broad range of data providers.

National Aeronautics and Space Administration

Drivers (2 of 2)

- **Greater User Expectations - The World Wide Web has led users to expect well-documented data to be readily available via text-based or graphical search systems with data delivery in a variety of formats compatible with their data processing systems.**
  - Capability to provide users well-documented data in a variety of contemporary formats.

- **Limited Funding – The emphasis on smaller, faster, cheaper missions which often include international partners may limit the ability to provide products suitable for analysis by the broader science community.**
  - Capability to efficiently meet the demands of a broad range of data users.

- **Creating a "System" from the Federation – The current PDS nodes operate autonomously and independently with limited distributed access via PDS-D to node repositories.**
  - Support location independence in a federated system.

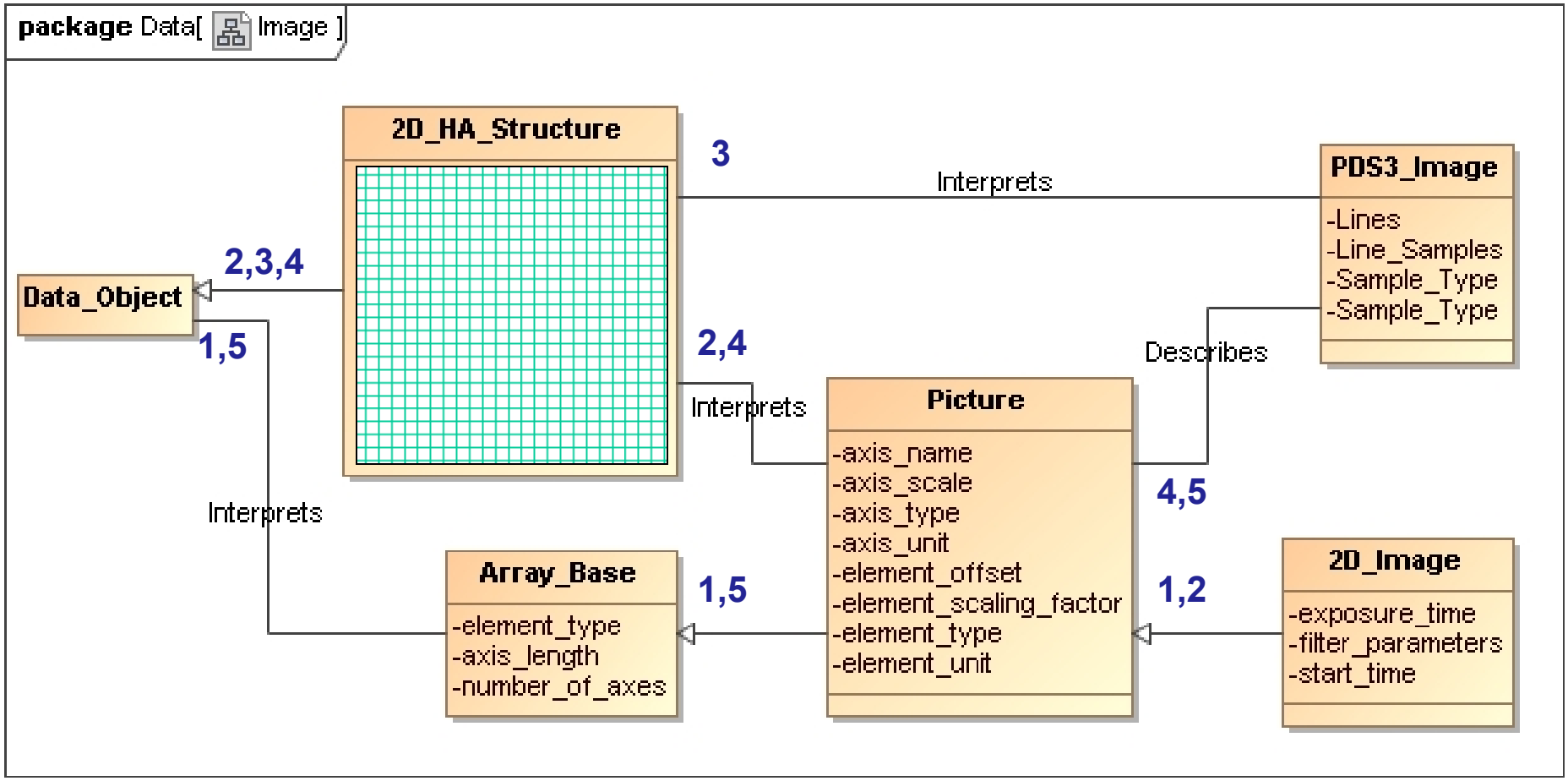
National Aeronautics and Space Administration

Model Components 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

National Aeronautics and Space Administration

Simple Image Models



National Aeronautics and Space Administration

What is a Data Model?

- Everything that isn't the real thing is a *model.*
- A *data* model is an abstract model that describes how data is represented and accessed.
- *Data modeling* is the process of creating a data model instance by applying a data model theory, typically to meet a set of requirements.
- A major component of data modeling is to visually represent the rules that the community wishes to enforce on data.
- An *information model* is a set of related data models.
  - E.g. Product vs Operational; Conceptual versus Logical

National Aeronautics and Space Administration

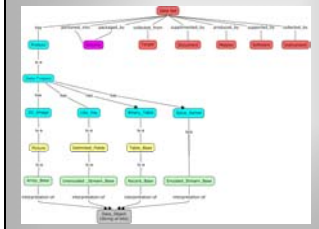
Data Architecture Design

Refine the context. Define the boundaries of the system.

Context Model

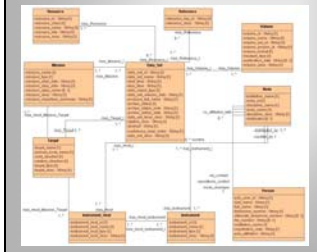
Define the community model of data from a manager's point of view. Includes concepts and key relationships.

Concept Model



Define the system model from a designer's point of view. Includes entity classes, attributes, and relationships in rigorous terms

Logical Model



Identify implementation technology. Determine detailed representation for the particular technology. E.g. ODL, XML

Physical Model

Target specific artifacts

- Primary/Foreign Keys
- ODL Objects
- XML Element/Attributes

Independence

Specificity

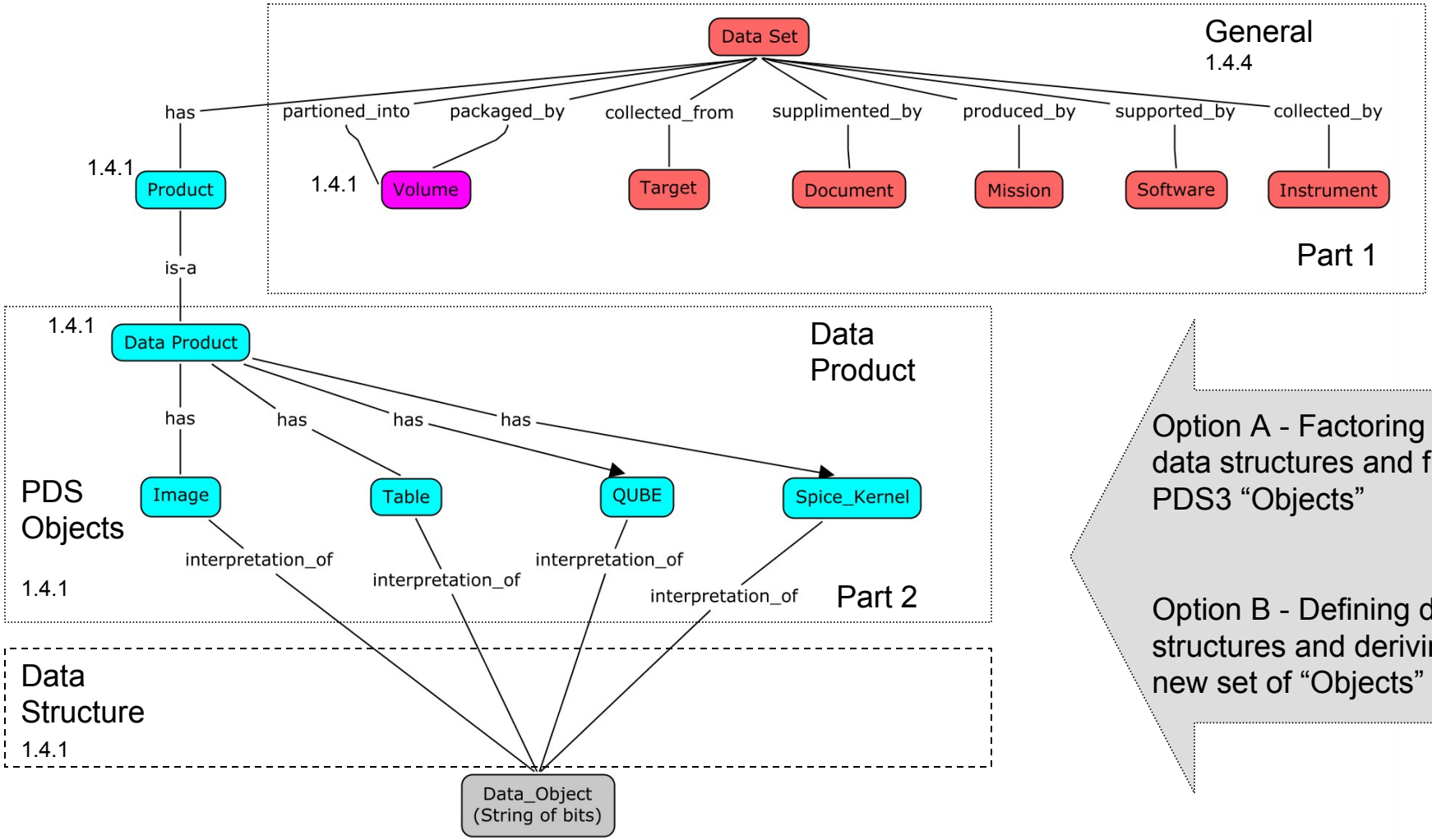
National Aeronautics and
Space Administration

Topics

- Background
- Data Architecture Scope
- Response to MC Charge
- Findings
- Roadmap
- Node Consensus
- Recommendations
- Identifying Tradeoffs
- Next Steps

National Aeronautics and Space Administration

PDS Conceptual Data Model
Architectural View



Option A - Factoring out data structures and fixing PDS3 “Objects”

Option B - Defining data structures and deriving a new set of “Objects”

National Aeronautics and Space Administration

PDS4 = Part 1 + Part 2

- **The following combinations illustrate how PDS4 can be built**

1. Part 1 + Part 2 (Option A)
2. Part 1 + Part 2 (Option B)

Part 1 *– "Path Independent" fixes to things in general.*
Includes fixes to data dictionary model and ODL.

Part 2 *– Two options for addressing PDS products and PDS "Objects".*

National Aeronautics and Space Administration

Examples of Differences Exercise 1

- While most nodes listed Archiving and current users support as the top priorities, two nodes considered helping data submitters to be a top priority, and two (different) nodes considered facilitating PDS operations to be a top priority.
- Half the nodes responding considered archiving data in simple formats to be the top priority - the other half rated this at zero.
- One node considers "one-stop-shopping" to be a high priority. Only one other node even placed this in the top 10.
- On the subject of archiving in "easy-to-use" formats, three nodes rated this as very high priority; the rest rated it at the lowest priority.
- One node rated "easy data submission" at a very high priority; most rated it around the middle, but three nodes rated it 2 or lower.
- One node rated archiving in contemporary formats as its highest priority. Only one other node included this in its top ten, and rated it only 1.

National Aeronautics and Space Administration

Examples of Differences Exercise 2

- While most nodes indicated a preference for strictly centralized data structures, two nodes indicated a preference for the possibility of node-specific structures.
- While most nodes indicated a desire for the possibility of node-specific metadata, three nodes leaned more heavily towards most metadata being centrally defined.
- Five nodes indicated strong preference for the standards to be globally defined, but three nodes indicated a desire to accommodate node-specific supplements (or extensions, if you prefer) to the common standards.