PDS4 Data Standards

PDS System Design Review II
Greenbelt, Maryland
June 21-22, 2010

Steve Hughes
PDS4 Data Design Working Group

Topics

- Overview
- Status and Next Steps
- Information Model
- Data Dictionary
- Grammar
- Support for Data Ingest and Distribution
- Standards Management

What is PDS4?

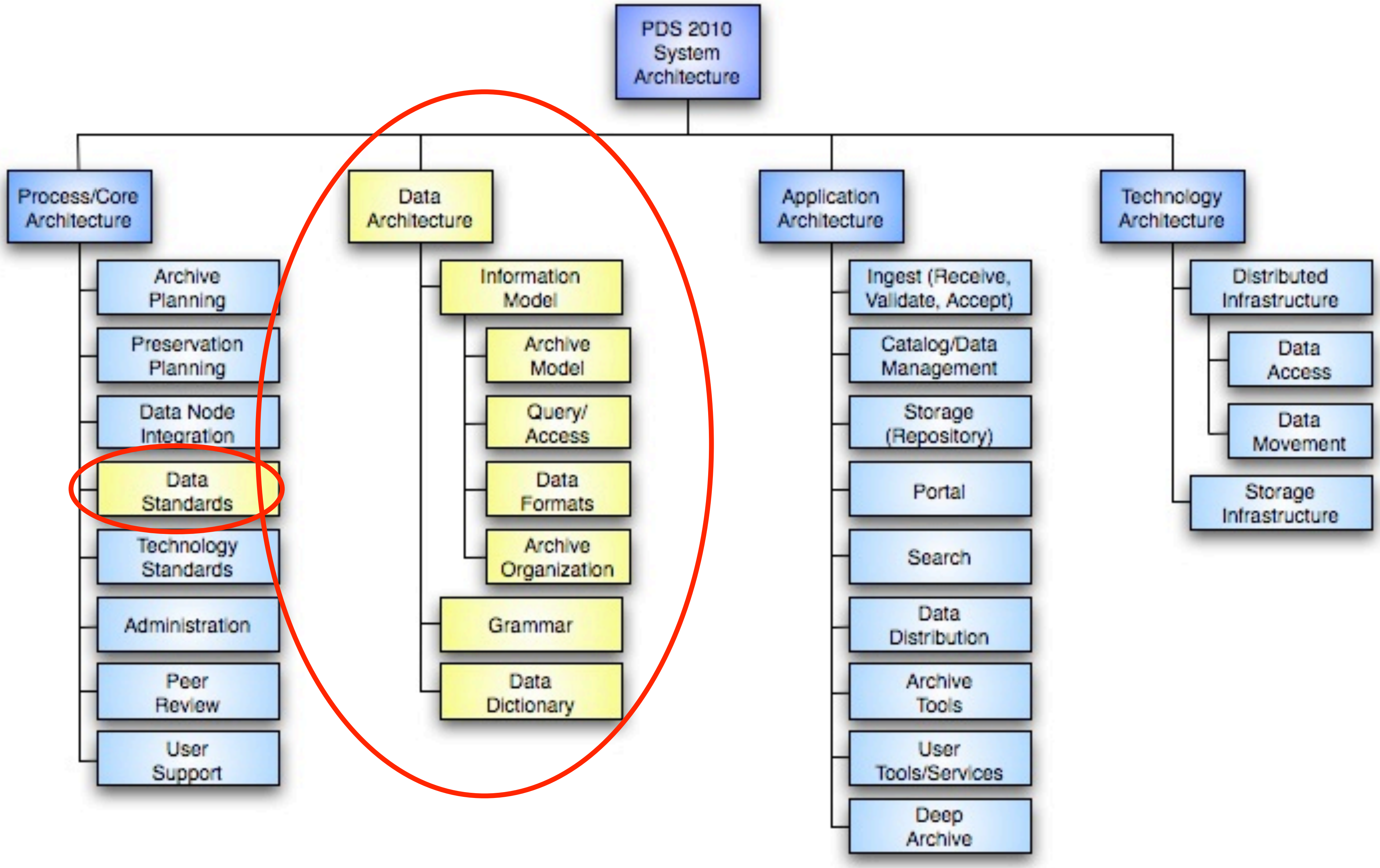
- A transition from a 20-year-old collection of data standards to a modern set of data standards constructed using best practices for standards development.
- Fewer, simpler, and more rigorously defined formats for science data products.
- Use of XML, a well-supported international standard, for data 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.

Data Architecture in Context

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

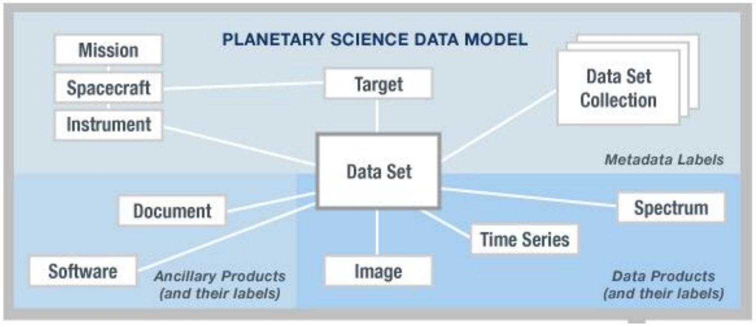
- The Data Architecture is a set of data standards for a planetary science archive data system
  - It guides system design, implementation and operations

PDS 2010 Architecture



Data Architecture Concepts

Information Model



Expressed As

Data Element

has

Class

Planetary Science Data Dictionary

Extracted/Specialized

Label Schema

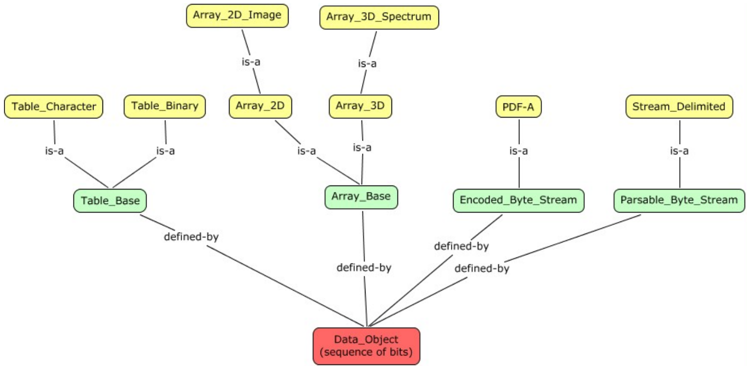
Validates

Used to Create

Product

Tagged Data Object

(Information Object)



Describes

Data Object

Level 2 and 3 Requirements Applicable to Data Architecture

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

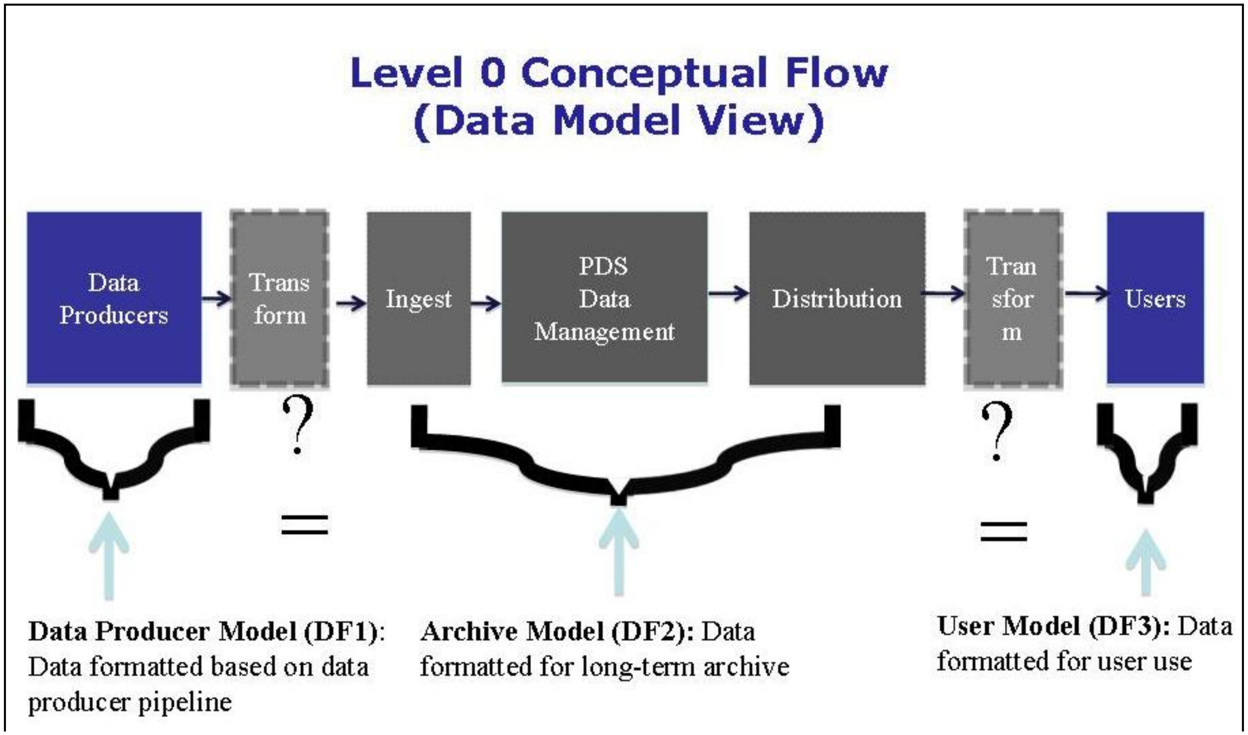
3.1.2 PDS will develop and maintain online interfaces for discipline-specific searching

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

Data Architecture Objectives

- 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.



Data Producer Model (DF1): Data formatted based on data producer pipeline

Archive Model (DF2): Data formatted for long-term archive

User Model (DF3): Data formatted for user use

Data Architecture Documents

The following eight ‘documents’ on the next two slides describe the Planetary Data System version 4 (PDS4).

1. Introduction – A guide to get you started.*

2. Concepts Document – Introduction to PDS4 key concepts — the view from 10000 feet, avoiding gory details.

3. Glossary – A concise set of definitions for key PDS4 terms. Although primarily intended as a quick reference, the Glossary is organized functionally, presenting terms in the approximate order in which you are likely to encounter them.

4. Jumpstart Guide – A brief introduction to PDS4 in terms of analogous PDS3 vocabulary. Experienced PDS3 users should read it once, noting both the parallels and the differences; then set it aside. People not familiar with PDS3 should skip it; concentrate on the Concepts Document.

5. Data Provider’s Handbook – A cookbook to guide data providers step-by-step through the process of developing an archive.

Data Architecture Documents (cont’d)

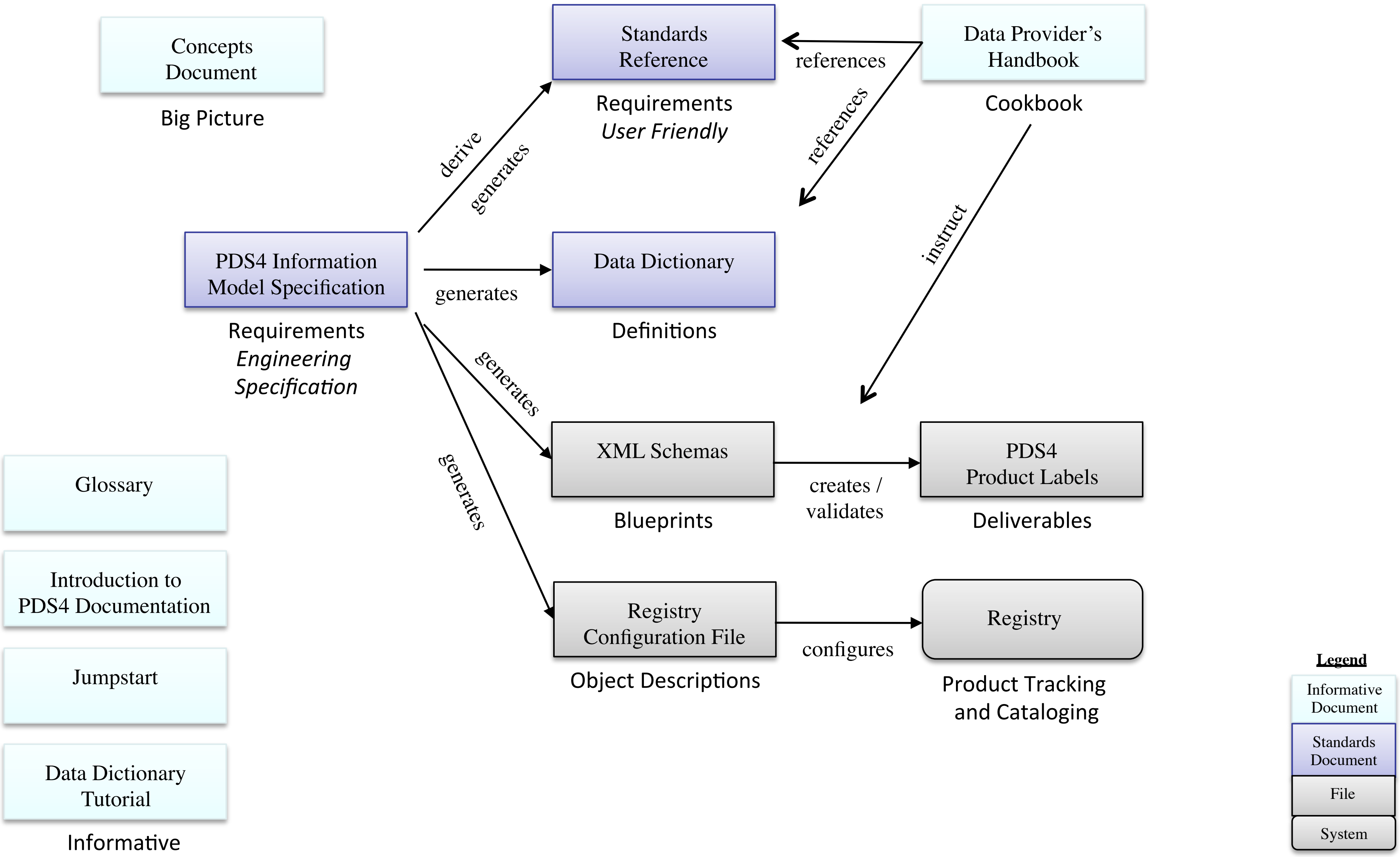
Reference documents:

6. Standards Reference – One of the two fundamental reference documents for PDS4. You will need this as you work your way through the Data Provider’s Handbook and as you prepare an archive.

7. Data Dictionary – The other fundamental reference for PDS4. It comes in two versions, abridged and unabridged. Use the abridged version unless you encounter a specific instance in which the information in the more detailed unabridged version is required. The abridged version has been abstracted from the unabridged version with the needs of data providers and data end users in mind. It contains full definitions but not all the fine detail or repetition necessary to support the underlying Information Model.

8. Examples – A set of products, collections, bundles, and packages that illustrates design concepts and goals. Frequently referenced by [5] and to be used in conjunction with [5-7] when constructing an archive.

Data Architecture Documents in Context



Concepts Document

Big Picture

Standards Reference

Requirements

User Friendly

Data Provider's Handbook

Cookbook

references

references

instruct

derive

generates

PDS4 Information Model Specification

Requirements

Engineering Specification

generates

Data Dictionary

Definitions

generates

generates

XML Schemas

Blueprints

creates / validates

PDS4 Product Labels

Deliverables

Registry Configuration File

Object Descriptions

configures

Registry

Product Tracking and Cataloging

Glossary

Introduction to PDS4 Documentation

Jumpstart

Data Dictionary Tutorial

Informative

Legend

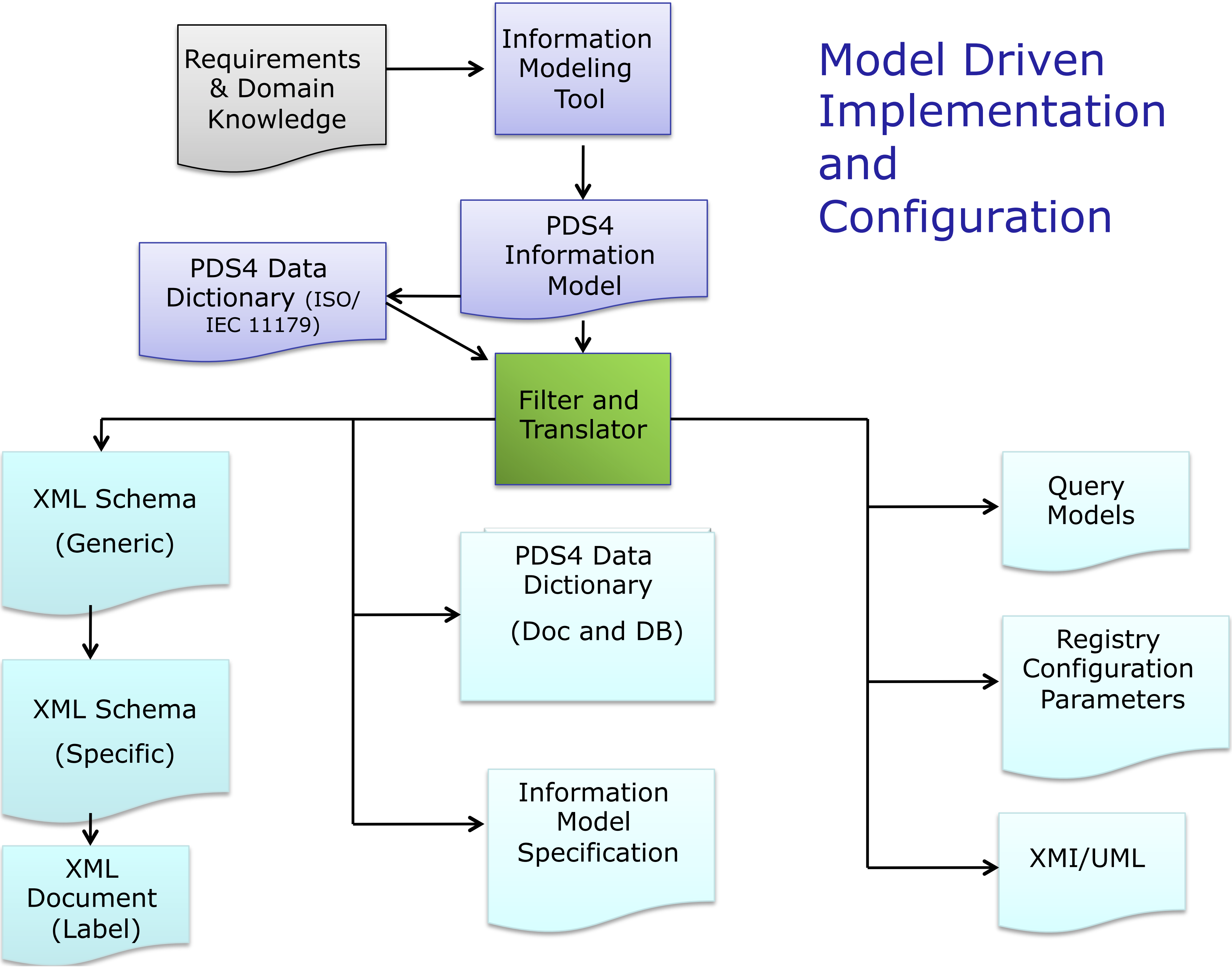
Informative Document

Standards Document

File

System

Model Driven Implementation and Configuration



Requirements & Domain Knowledge

Information Modeling Tool

PDS4 Information Model

PDS4 Data Dictionary (ISO/ IEC 11179)

Filter and Translator

XML Schema (Generic)

XML Schema (Specific)

XML Document (Label)

PDS4 Data Dictionary (Doc and DB)

Information Model Specification

Query Models

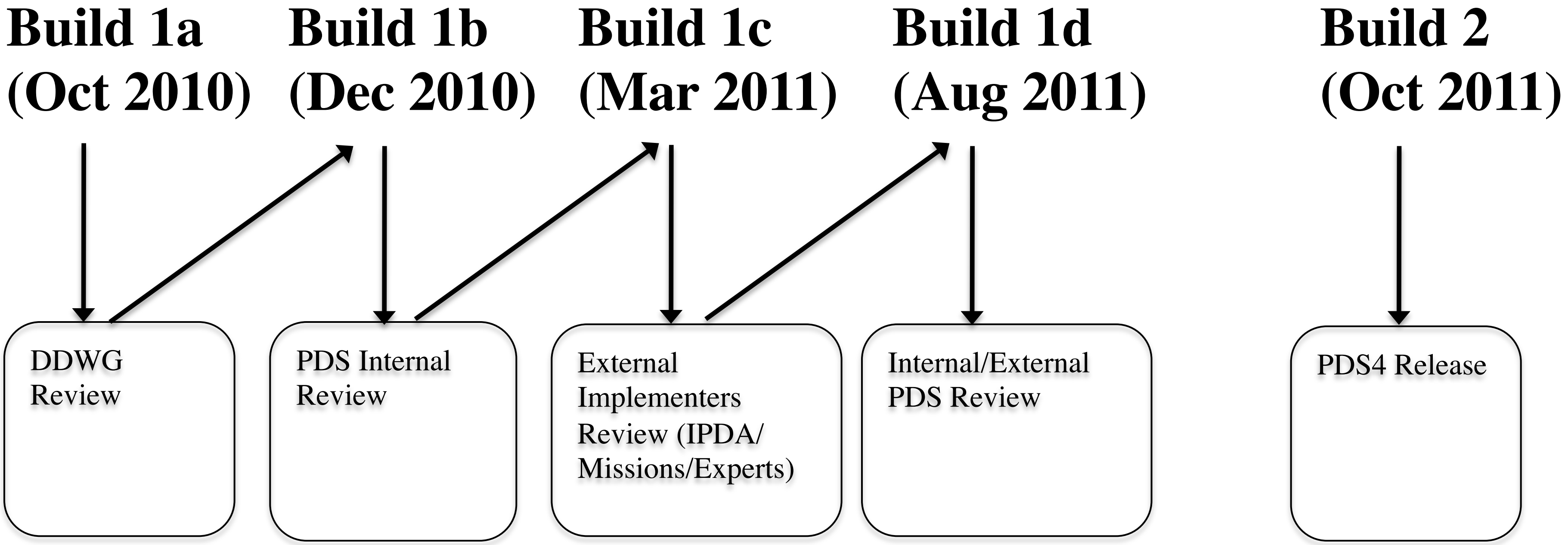
Registry Configuration Parameters

XMI/UML

Topics

- Overview
- Status and Next Steps
- Information Model
- Data Dictionary
- Grammar
- Support for Data Ingest and Distribution
- Standards Management

Build/Assessment Alignment



Build 1a (Oct 2010)

Build 1b (Dec 2010)

Build 1c (Mar 2011)

Build 1d (Aug 2011)

Build 2 (Oct 2011)

DDWG Review

PDS Internal Review

External Implementers Review (IPDA/ Missions/Experts)

Internal/External PDS Review

PDS4 Release

Finding from Internal Review*

- Clarification/Ambiguity (62)
- Completeness/Incomplete (49)
- Complexity (33)
- Kudos (31)
- Consistency/Conflict (20)
- Omission/Missing Items (16)
- Duplication (9)
- Bugs/Errors (7)
- Examples (4)
- Focus (2)
- Format (7)
- Organization (4)

* From Hughes and Simpson rollup

Status (Build 2)

	Document/Artifact	*Reviews*	*Status*	*Next Steps*
1	Introduction	3	Mature	Ready for Build
2	Concepts Document	2, 3	Mature	Ready for Build
3	Glossary	2, 3	Mature	Ready for Build
4	Jumpstart Guide	2, 3	Mature	Ready for Build
5	Data Provider's Handbook	1, 2, 3	Cleanup in Progress	External Review
6	Standards Reference	2, 3	Cleanup in Progress	External Review
7	Data Dictionary	1, 2, 3	Cleanup in Progress	External Review
8	Examples	2, 3	Make consistent \w model	External Review
10	Schemas	1, 2, 3	Consistent with model	External Review
11	Information Model	1, 2, 3	Core – Almost complete	Release at Build
			Discipline Level – Phase 1	Release at Build

	Reviews
1	IPDA -1
2	Internal PDS
3	IPDA -2
4	External

Build 2 – Oct ‘11

- Complete the core information model for Build 2
  - Provenance, Targets
- Complete the design of the initial set of discipline node classes necessary for early PDS3 data product migration and the first missions using PDS4 data standards (LADEE, MAVEN).
- Finalize the processes and interfaces for the mission data dictionary.
- Deliver Version 1.0 of the PDS4 Data Standards Documents.
- Prepare for the Operational Readiness Review.

Build 3 – Summer ‘12

- Design the next increment of discipline node and mission classes.
- Design the next increment of core components (e.g. Qube).
- Mature the standards management processes and operational procedures.
- Continue to used the nodes and the IPDA to test and exercise new elements of the model.
- Deliver Version 1.1 of the PDS4 Data Standards Documents.

Post Summer ‘12

- Evolve the PDS4 Data Standards as required for the changing planetary science community.
  - Design the model elements necessary for new missions, instruments, and data products.
  - Release new versions of the documents as necessary.
- Continue to support the PDS3 data product migration effort.
- Continue to mature the standards management processes and operational procedures.

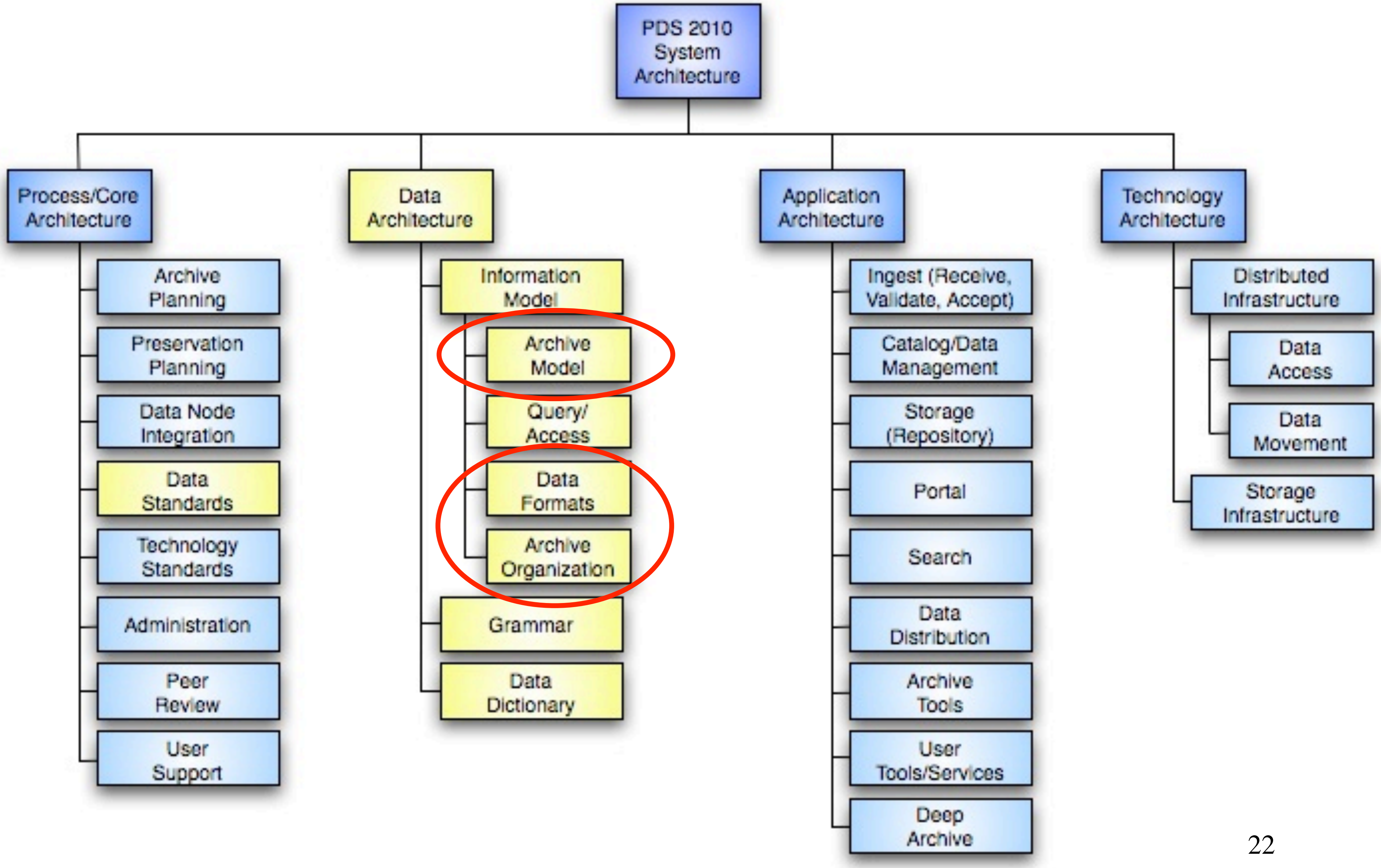
Provenance

- Attribution - The sources or entities that contributed to create the artifact in question.
- Process - The activities (or steps) that were carried out to generate or access the artifact at hand.
- Versioning - Records of changes to an artifact over time and what entities and processes were associated with those changes.
- Justification - Documentation recording why and how a particular decision is made.
- Entailment - Explanations showing how facts were derived from other facts.

Topics

- Overview
- Status and Next Steps
- Information Model
- Data Dictionary
- Grammar
- Support for Data Ingest and Distribution
- Standards Management

PDS 2010 Architecture



Design Approach

- 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 schemas, and other development artifacts.
- Leverage existing standards.

Key Features of the Information Model

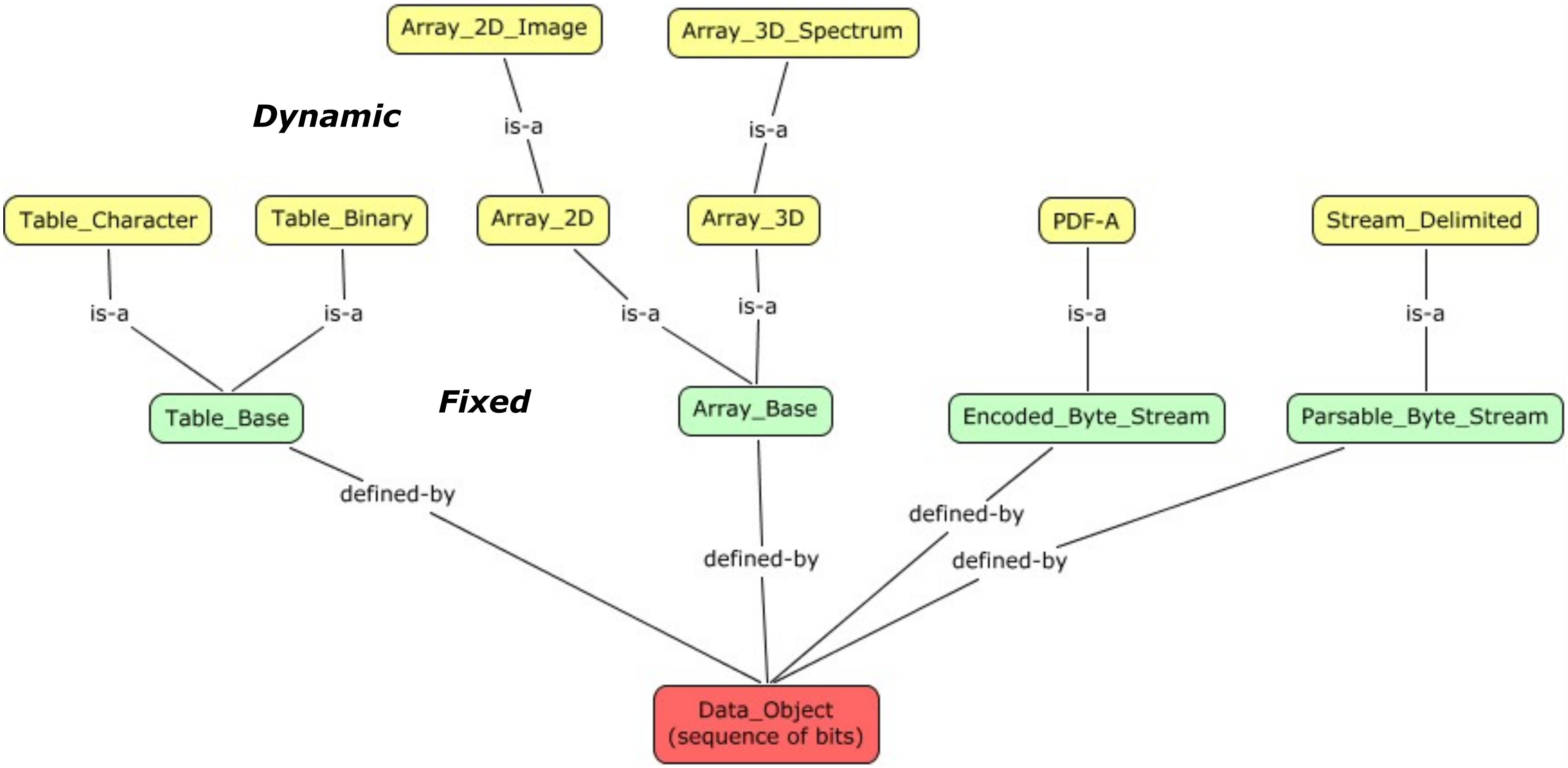
- Four base formats for all archived information
- Physical data segments map directly to logical segments
- Documents, software and ancillary data treated as rigorously as observational data
- Keyword content sorted into independent classes
- Product Centric
  - Products are registry objects

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



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

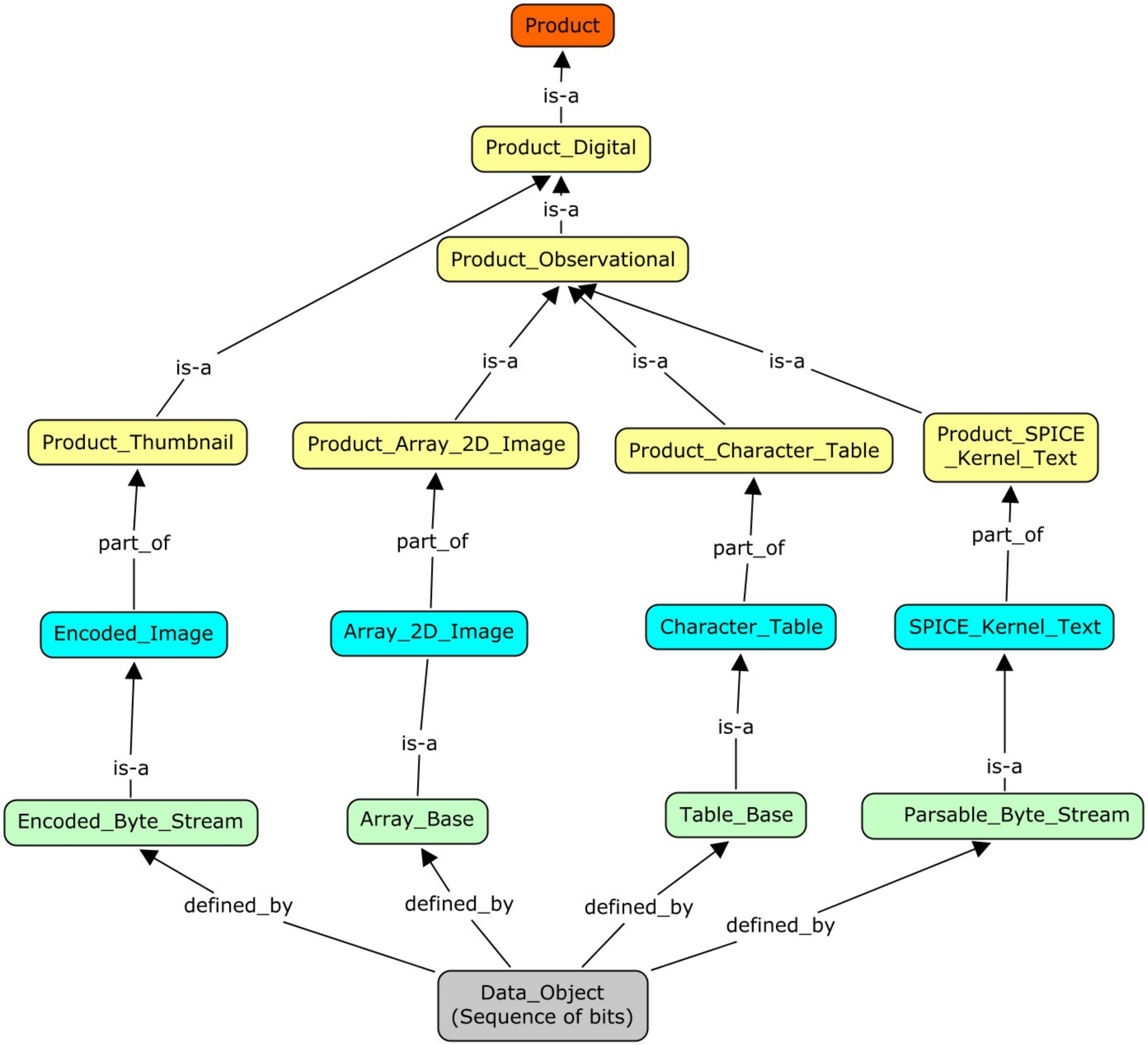
Note that this does not require bit manipulation.

All Products Are Equal

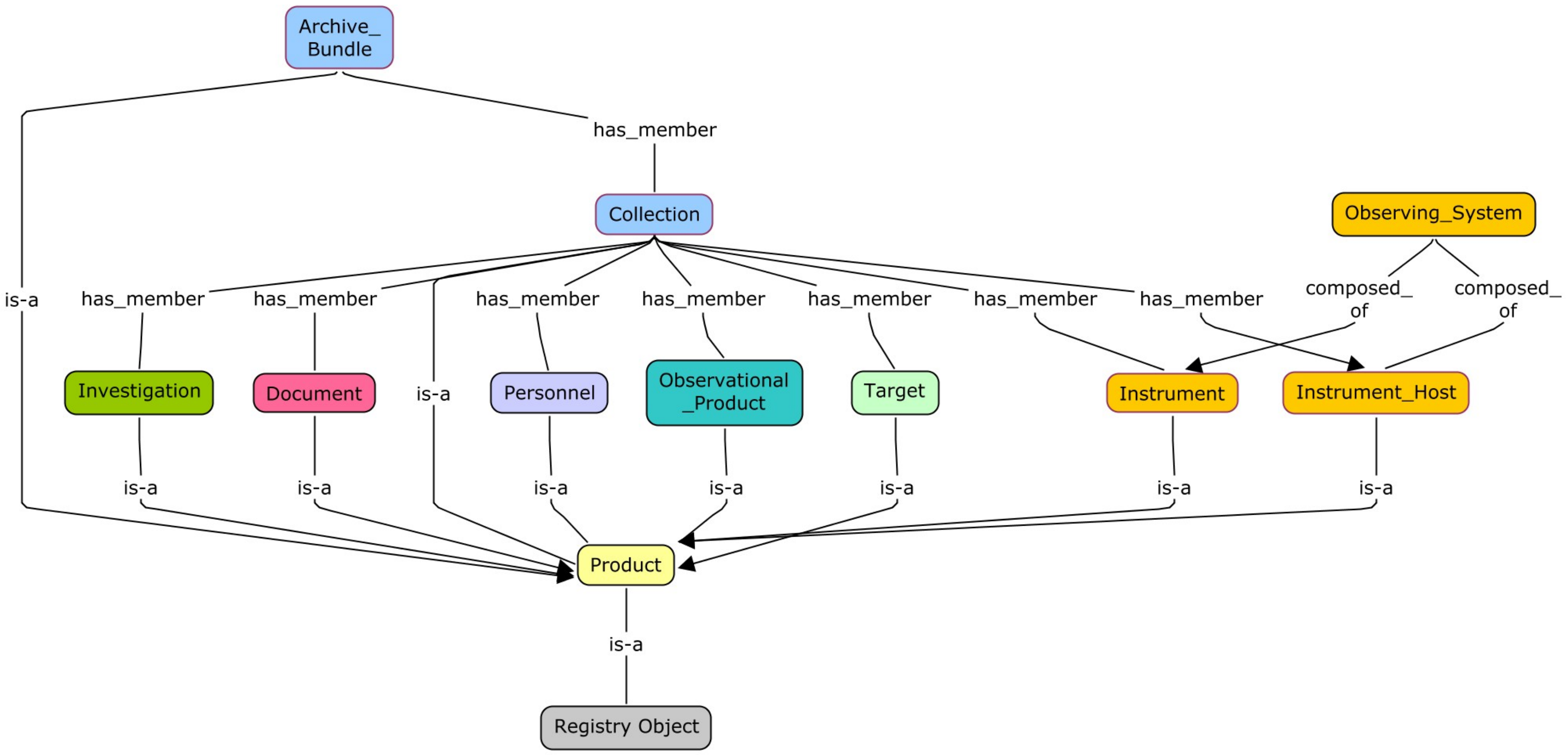
All products are treated with equal rigor in labelling and documenting.

- 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

Observational Product – Concept Map



Observational Product in Context



Industry Standards* Referenced and Controlling

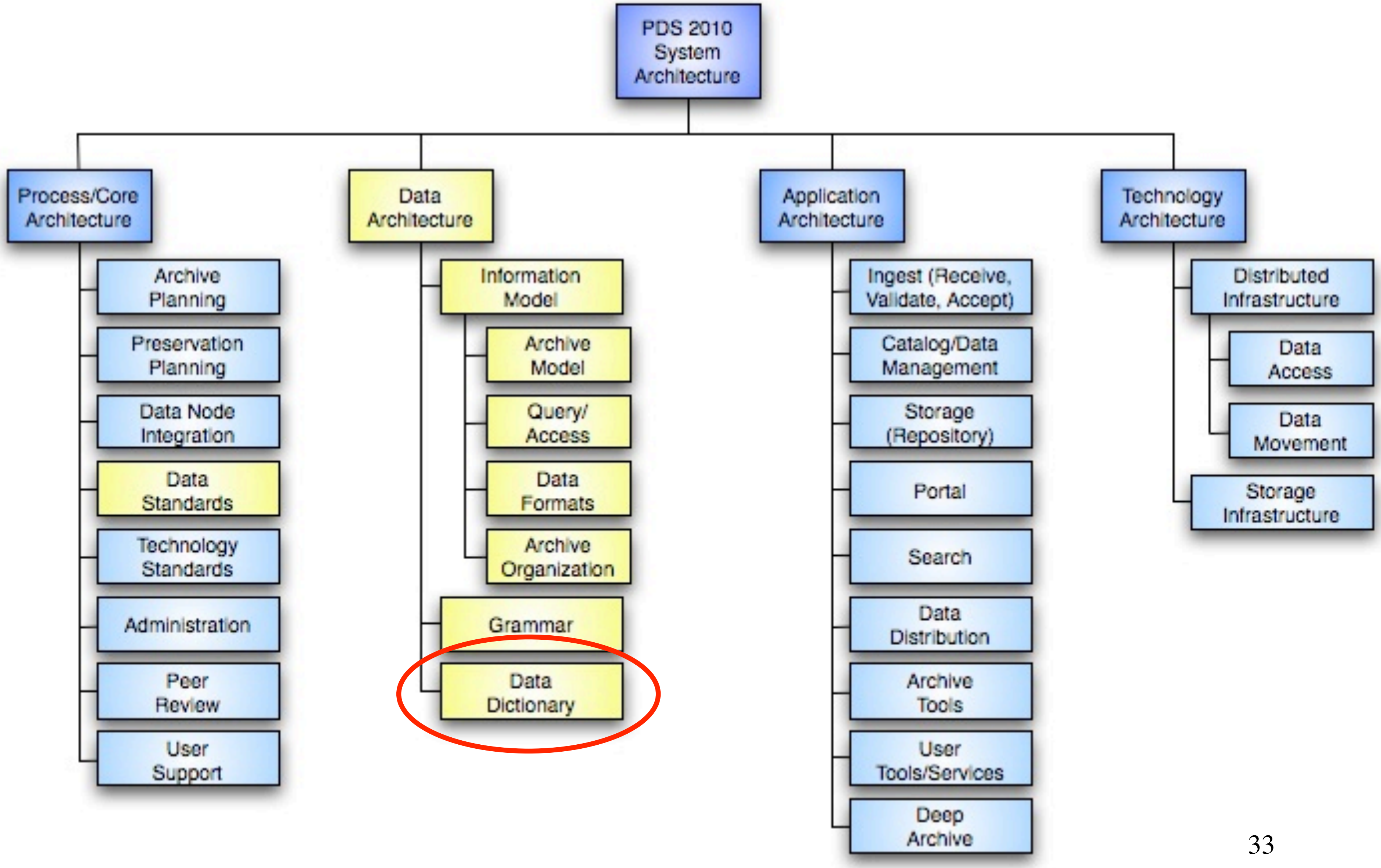
- ISO/IEC 11179:3 Registry Metamodel and Basic Attributes specification - Adopted for the data dictionary schema.
- ISO/IEC 11404:2007(E) - Provides the specification for language-independent data types.
- Reference Architecture for Space Information Management (RASIM) - CCSDS 312-0.G-1 – Provides the 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

Topics

- Overview
- Status and Next Steps
- Information Model
- Data Dictionary
- Grammar
- Support for Data Ingest and Distribution
- Standards Management

PDS 2010 Architecture



Design Approach

- 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 keywords

- Support international requirements

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

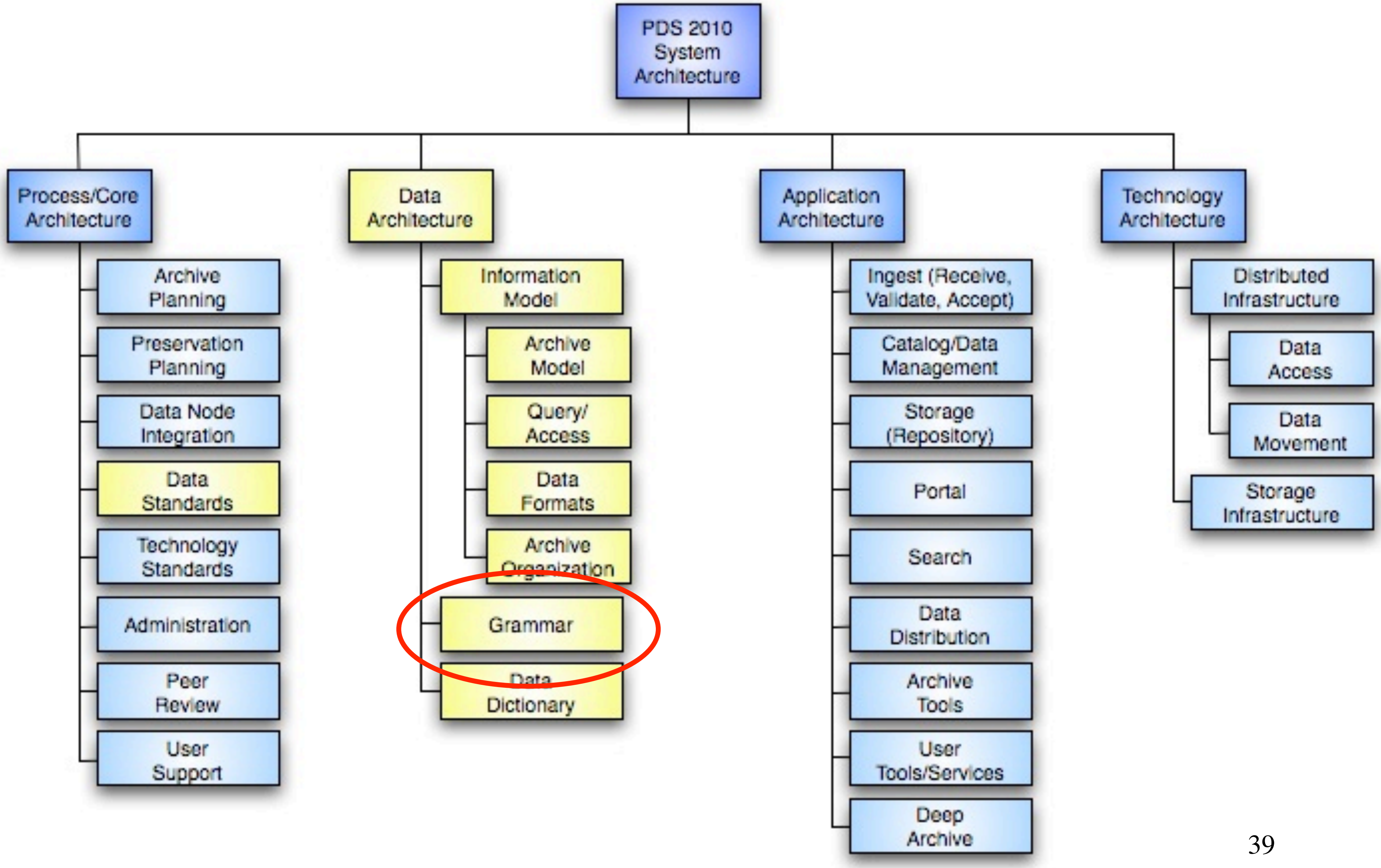
Data Dictionary Governance

- 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.

Topics

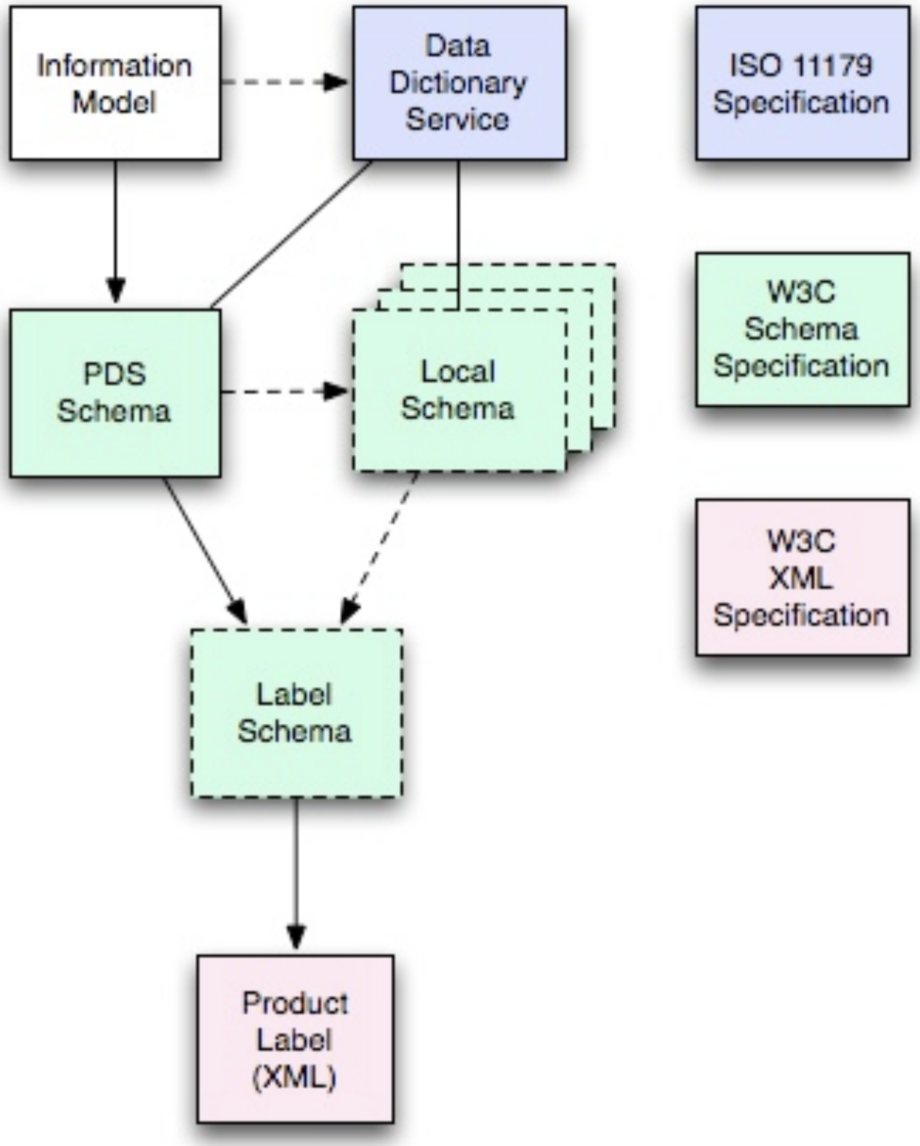
- Overview
- Status and Next Steps
- Information Model
- Data Dictionary
- Grammar
- Support for Data Ingest and Distribution
- Standards Management

PDS 2010 Architecture



XML Integration

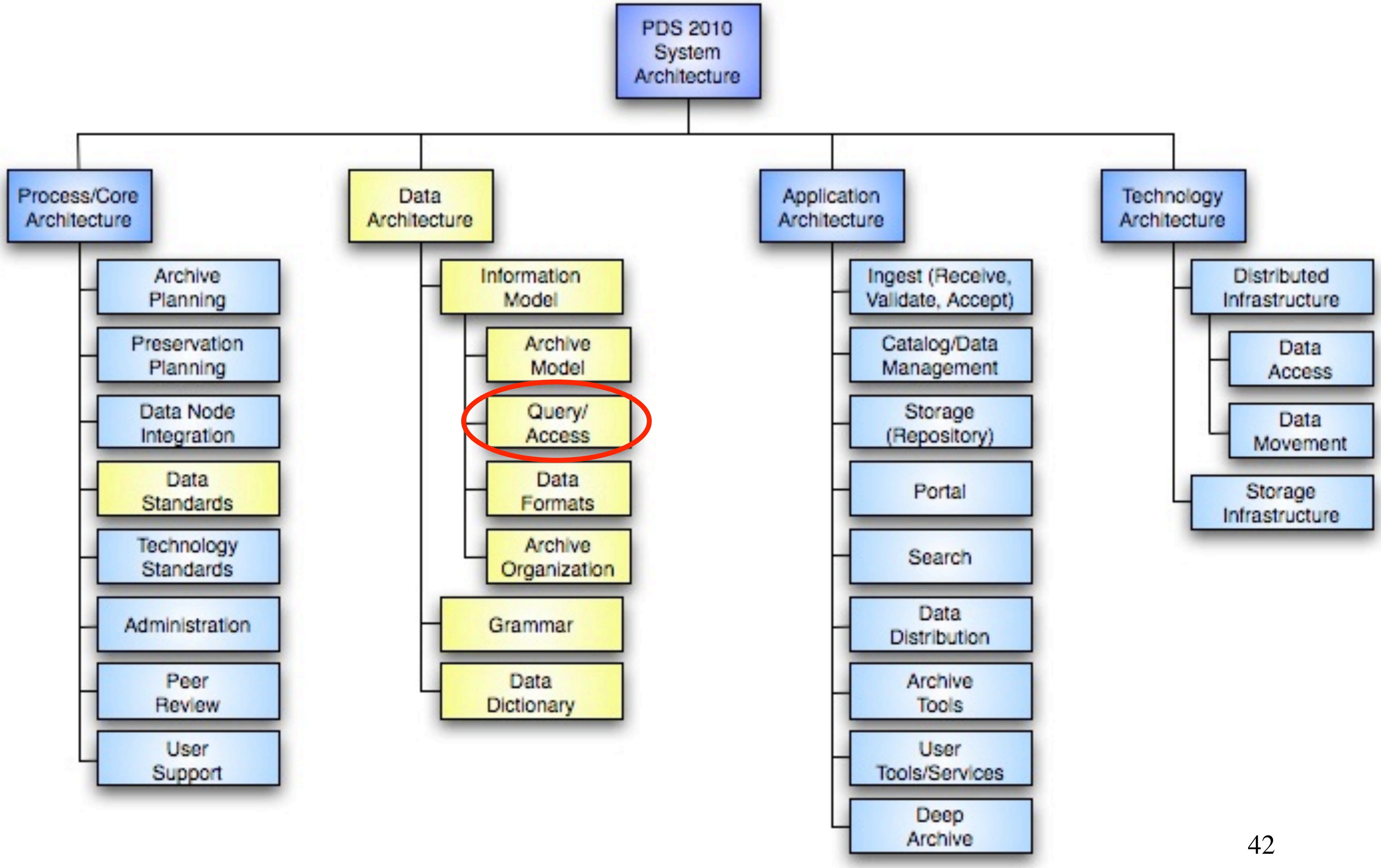
- Data Dictionary Service – manages dictionaries using ISO 11179 Model. Exports dictionaries to a XML Schema
- PDS Schema – Captures the types, elements, and structures within the PDS namespace
- Local Schema – Captures the types, elements, and structures for some mission, node, etc. Builds on and inherits 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



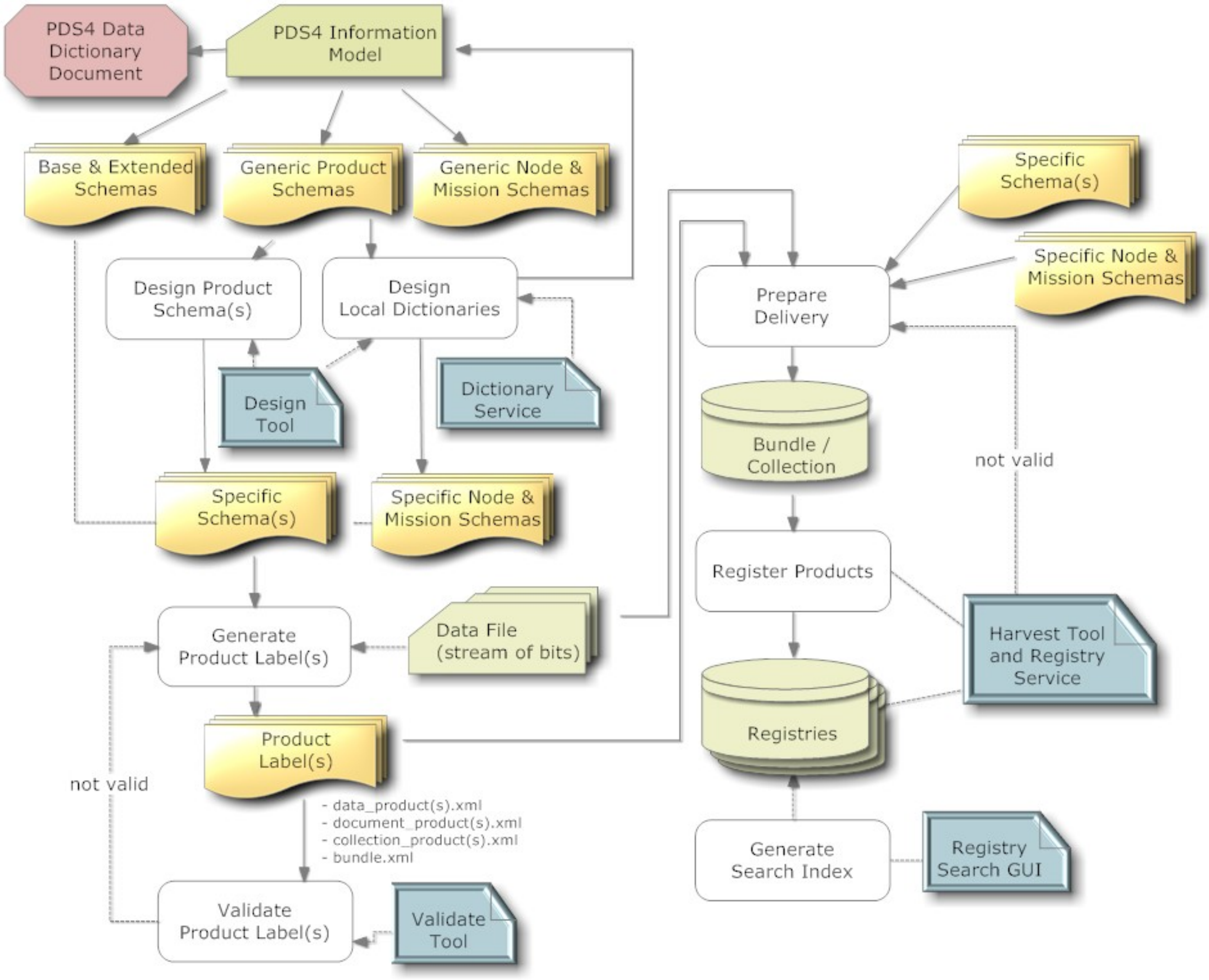
Topics

- Overview
- Status and Next Steps
- Information Model
- Data Dictionary
- Grammar
- Support for Data Ingest and Distribution
- Standards Management

PDS 2010 Architecture



Process for Data Product Creation



Product Identification

```
<Identification_Area_Product>
      <logical_identifier> urn:nasa:pds:VG2-J-PLS</logical_identifier>
      <version_id>1.0</version_id>
      <product_class> Product_Table_Character </product_class>
      <title> Voyager Electron density and moment …</title>
      <alternate_title> … </alternate_title>
      <alternate_id> … </alternate_id>
      <last_modification_date_time>2011-04-15T00:36:08.000Z </last
      <product_subclass> … </product_subclass>
      <type>Observational_Product</type>

         …

   </Identification_Area_Product>
```

Product Versioning

```
<Identification_Area_Product>
      <logical_identifier> urn:nasa:pds:VG2-J-PLS</logical_identifier>
      <version_id>1.0</version_id>
      <product_class> Product_Table_Character </product_class>
      <title> Voyager Electron density and moment …</title>
      <alternate_title> … </alternate_title>
      <alternate_id> … </alternate_id>
      <last_modification_date_time>2011-04-15T00:36:08.000Z </last
      <product_subclass> … </product_subclass>
      <type>Observational_Product</type>

         …

   </Identification_Area_Product>
```

Product Typing

```
<Identification_Area_Product>
        <logical_identifier> urn:nasa:pds:VG2-J-PLS</logical_identifier>
        <version_id>1.0</version_id>
        <product_class> Product_Table_Character </product_class>
        <title> Voyager Electron density and moment …</title>
        <alternate_title> … </alternate_title>
        <alternate_id> … </alternate_id>
        <last_modification_date_time>2011-04-15T00:36:08.000Z </last
        <product_subclass> … </product_subclass>
        <type>Observational_Product</type>

            …

    </Identification_Area_Product>
```

Product Cross Referencing

```
<Cross_Reference_Area_Product>
  …
  <Reference_Entry_Product>
    <lid_reference>urn:nasa:pds:instrument.PLS_VG2</lid_reference
    <reference_association_type>has_instrument</reference_associat
  </Reference_Entry_Product>
 …
</Cross_Reference_Area_Product>
```

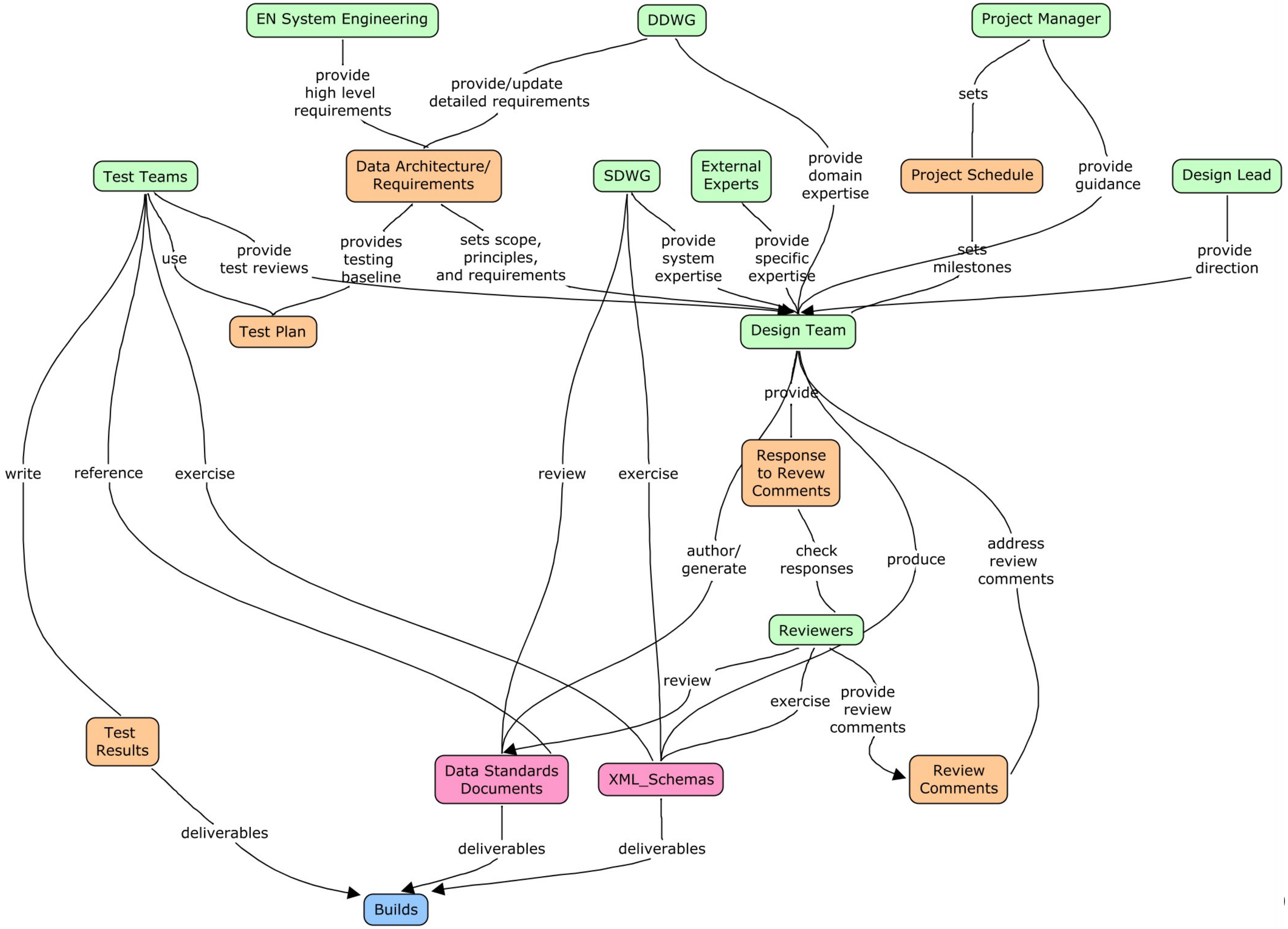
Product Search Parameters

```
<Identification_Area_Product>
  …
  <Subject_Area>
    <target_name>JUPITER</target_name>
    <instrument_name>PLASMA SCIENCE EXPERIMENT</instru
    <instrument_host_name>VOYAGER 2</instrument_host_name>
    <investigation_name>VOYAGER</investigation_name>
    <keywords>…</keywords>
  </Subject_Area>
</Identification_Area_Product>
```

Topics

- Overview
- Status and Next Steps
- Information Model
- Data Dictionary
- Grammar
- Support for Data Ingest and Distribution
- Standards Management

Data Design Process

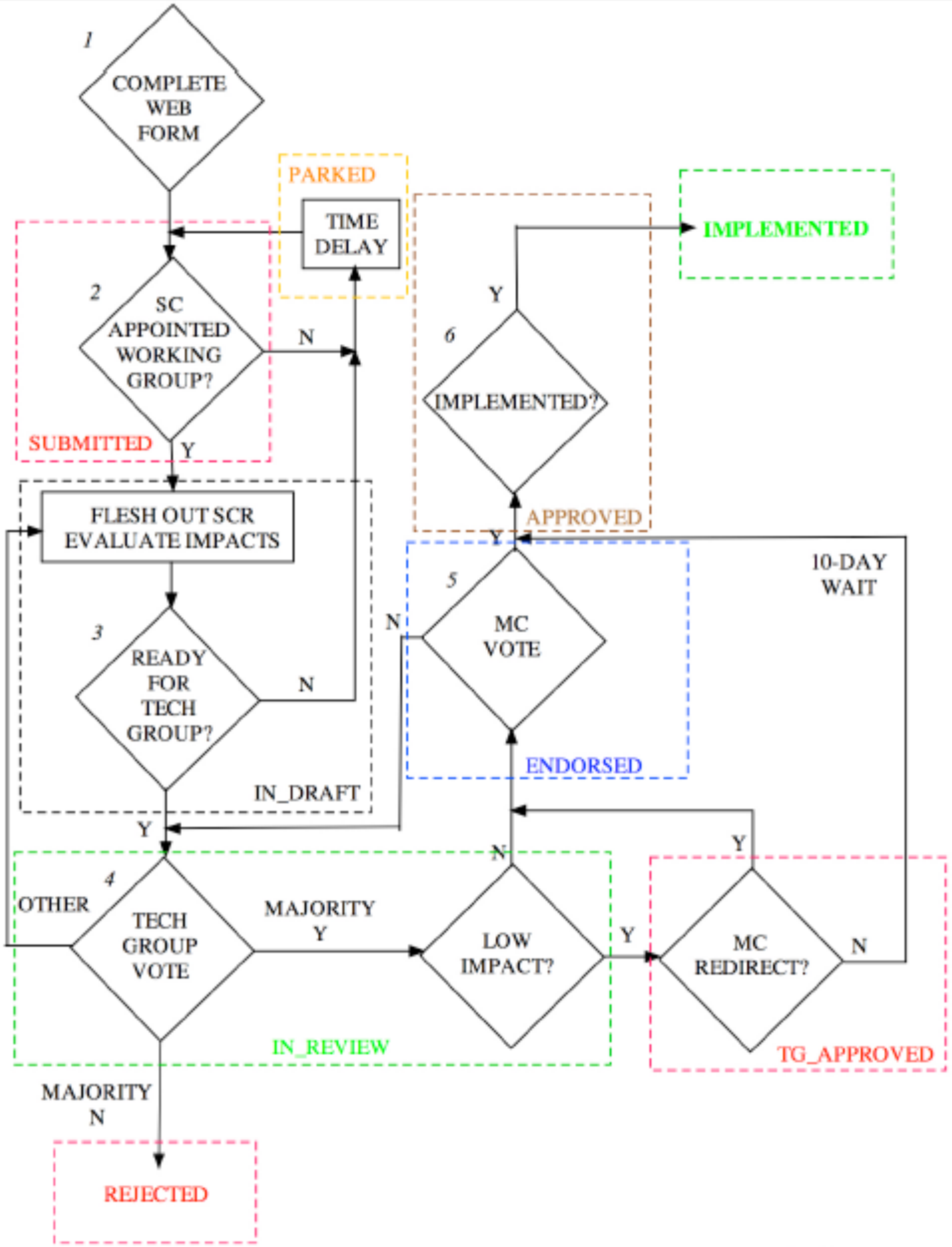


PDS4 Standards Management

- 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

Acknowledgements*

Ed Bell
Richard Chen
Dan Crichton
Amy Culver
Patty Garcia
Ed Grayzeck
Ed Guinness
Mitch Gordon
Sean Hardman
Lyle Huber
Steve Hughes
Chris Isbell
Steve Joy
Ronald Joyner
Debra Kazden
Todd King
Joe Mafi
Mike Martin
Thomas Morgan
Lynn Neakrase
Paul Ramirez
Anne Raugh
Elizabeth Rye
Boris Semenov
Dick Simpson
Susie Slavney

* Anyone who sat through a DDWG 2-hour telecon or provided useful input.

Online Resources

- PDS4 Deliverables
  - http://pds-engineering.jpl.nasa.gov/index.cfm?pid=145&cid=167
- PDS4 Project Wiki
  - https://oodt.jpl.nasa.gov/wiki/display/pdscollaboration/Data+Design+Working+Group

Thank You!

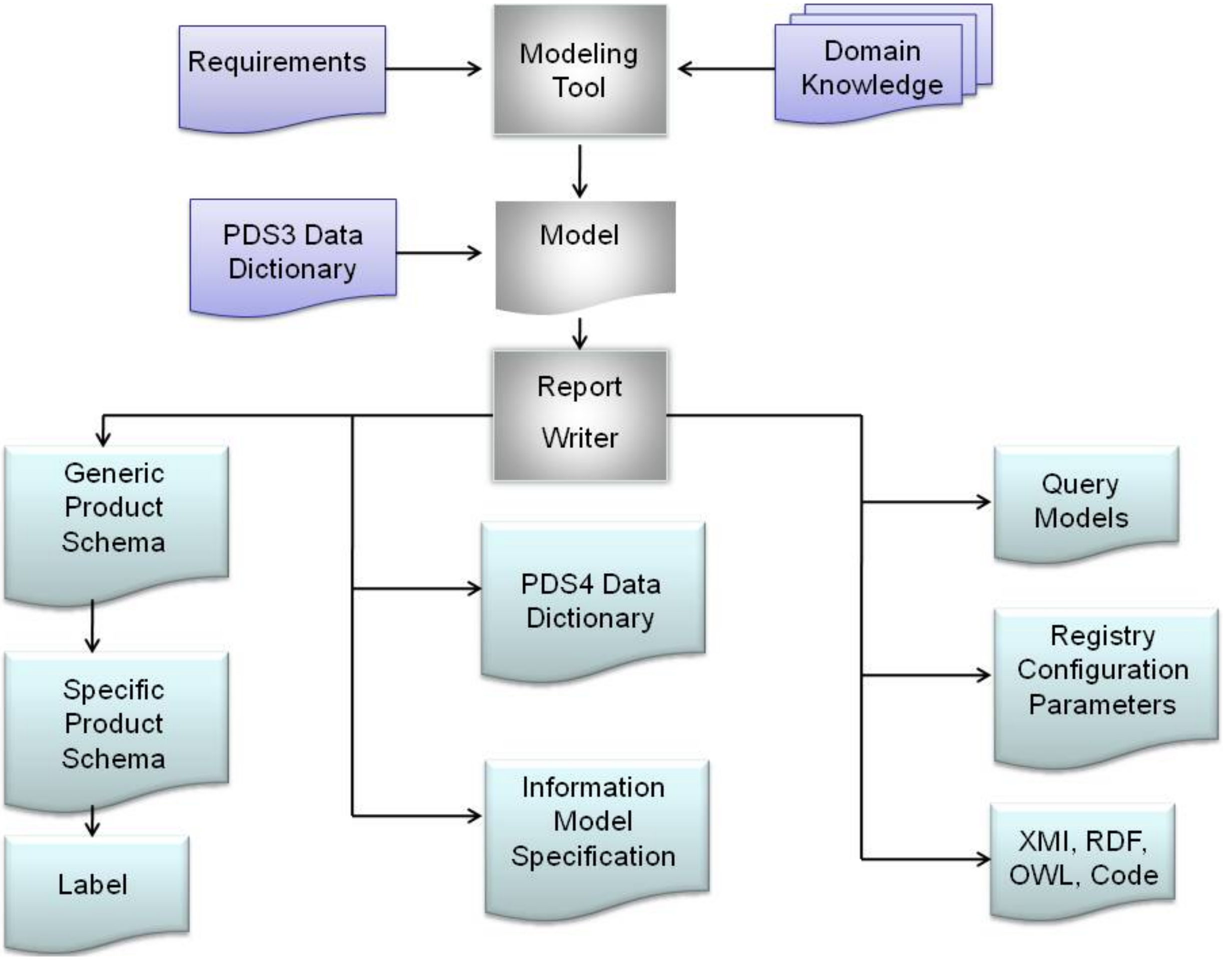
Backup

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.



Provenance

Provenance of a resource is a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource.

Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility.

Provenance assertions are a form of contextual metadata and can themselves become important records with their own provenance.