An Ontology Driven Information Architecture for Interoperable Disparate Data Sources

EGU General Assembly 2011
EGU2011-9509 ESSI16

J. Steven Hughes,
Dan Crichton, Sean Hardman, Ronald Joyner, Chris Mattmann, Paul Ramirez, Sean Kelly

Presented by Rebecca Castano

Jet Propulsion Laboratory
California Institute of Technology

This work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Topics

- Introduction and Problem Statement
- Information Object
- Domain Ontology
- Federated Registry
- Conclusion

The Planetary Data System (PDS)

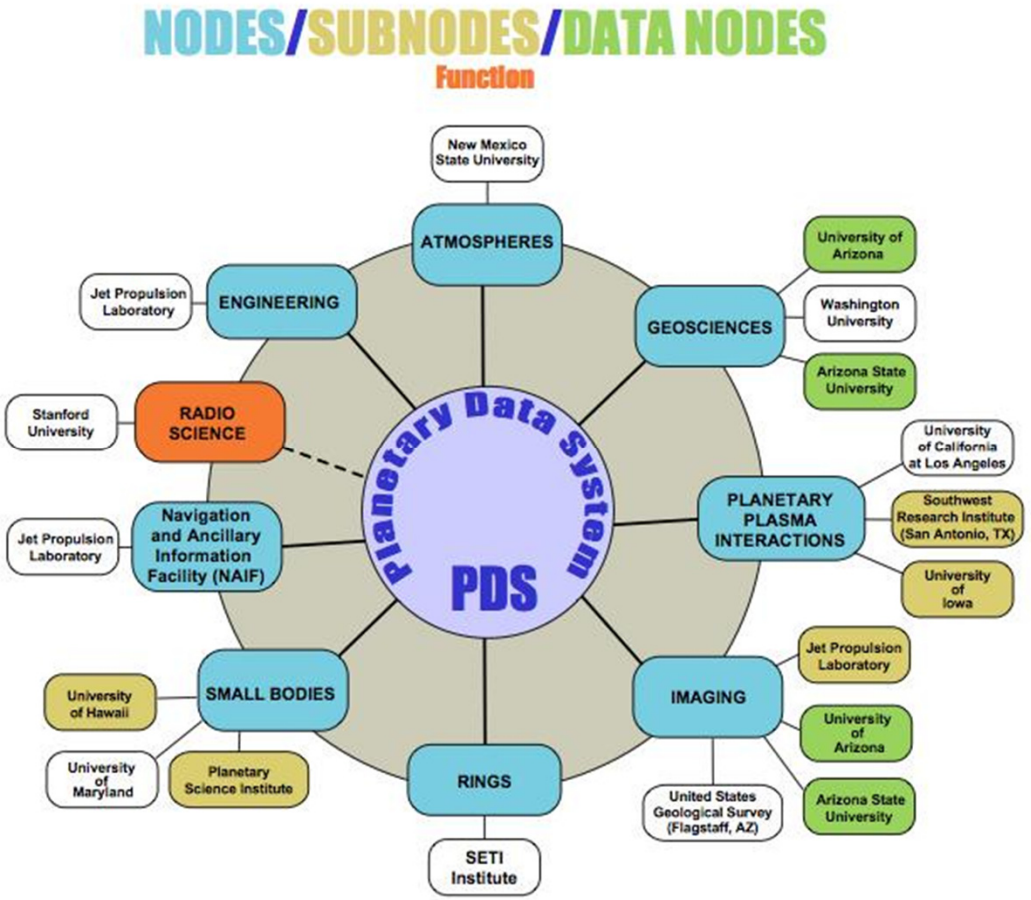
- **Mission:** The mission of the Planetary Data System is to facilitate achievement of NASA's planetary science goals by efficiently collecting, archiving, and making accessible digital data produced by or relevant to NASA's planetary missions, research programs, and data analysis programs

- **Vision:**
  - To gather and preserve the data obtained from exploration of the Solar System by the U.S. and other nations
  - To facilitate new and exciting discoveries by providing access to and ensuring usability of those data to the worldwide community
  - To inspire the public through availability and distribution of the body of knowledge reflected in the PDS data collection

- PDS is a federation of heterogeneous nodes including science and support nodes

A Federation of Nodes

- **NASA's official archive for Planetary Science Data**
  - Missions are required to archive data with PDS
  - Data is peer reviewed as part of the archiving of data
  - PDS works to support planetary science R&A
- **Federation of nodes**
  - Science discipline nodes provide scientific and data management expertise
  - Central engineering (JPL) addresses PDS-wide software and standards
  - Management node (GSFC) provides management support to PDS
  - Governed by the PDS Management Council which is formed from the node managers



Problem Statement

- Objective

  Efficiently provide product level identification, versioning, tracking, association, classification, search, retrieval and subscription/notification

- Challenge

  Support federate distributed nodes covering eight science disciplines and with thousands of product types

Solution

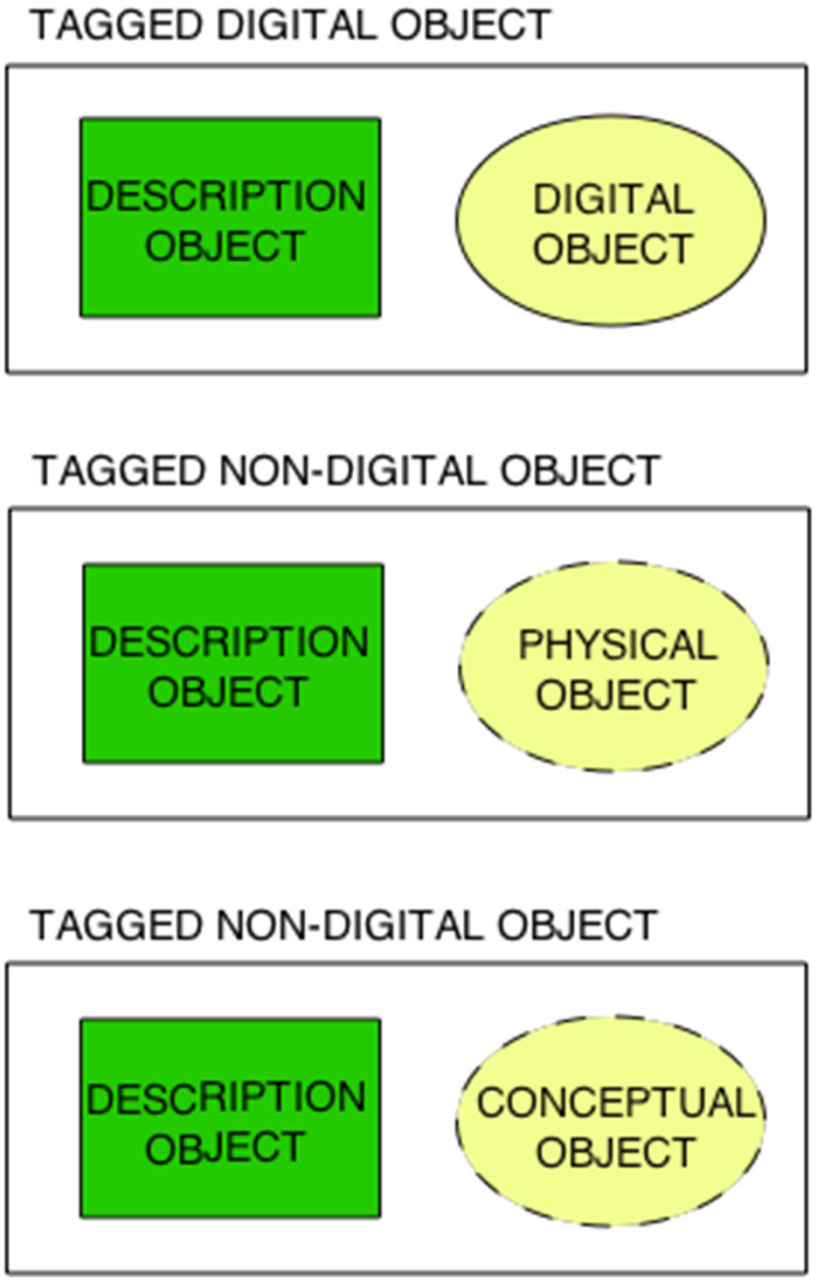
An information architecture that enables interoperability between the disparate data sources that comprise the archive

Information architecture elements

- Information objects
- Domain ontologies
- Federated registry

Information Object

- Data object paired with its description (metadata object)
- Information objects are combined to create archive products.
- From the Open Archival Information System (OAIS) Reference Model
  - CCSDS 650.0-B-1 of the Consultative Committee for Space Data Systems
  - ISO 14721:2003
- Provides a standard for the unification of digital, conceptual, and physical data objects



- **digital object**: An object which is real data — for example, a binary image of a redwood tree or an ASCII table of atmospheric composition versus altitude.

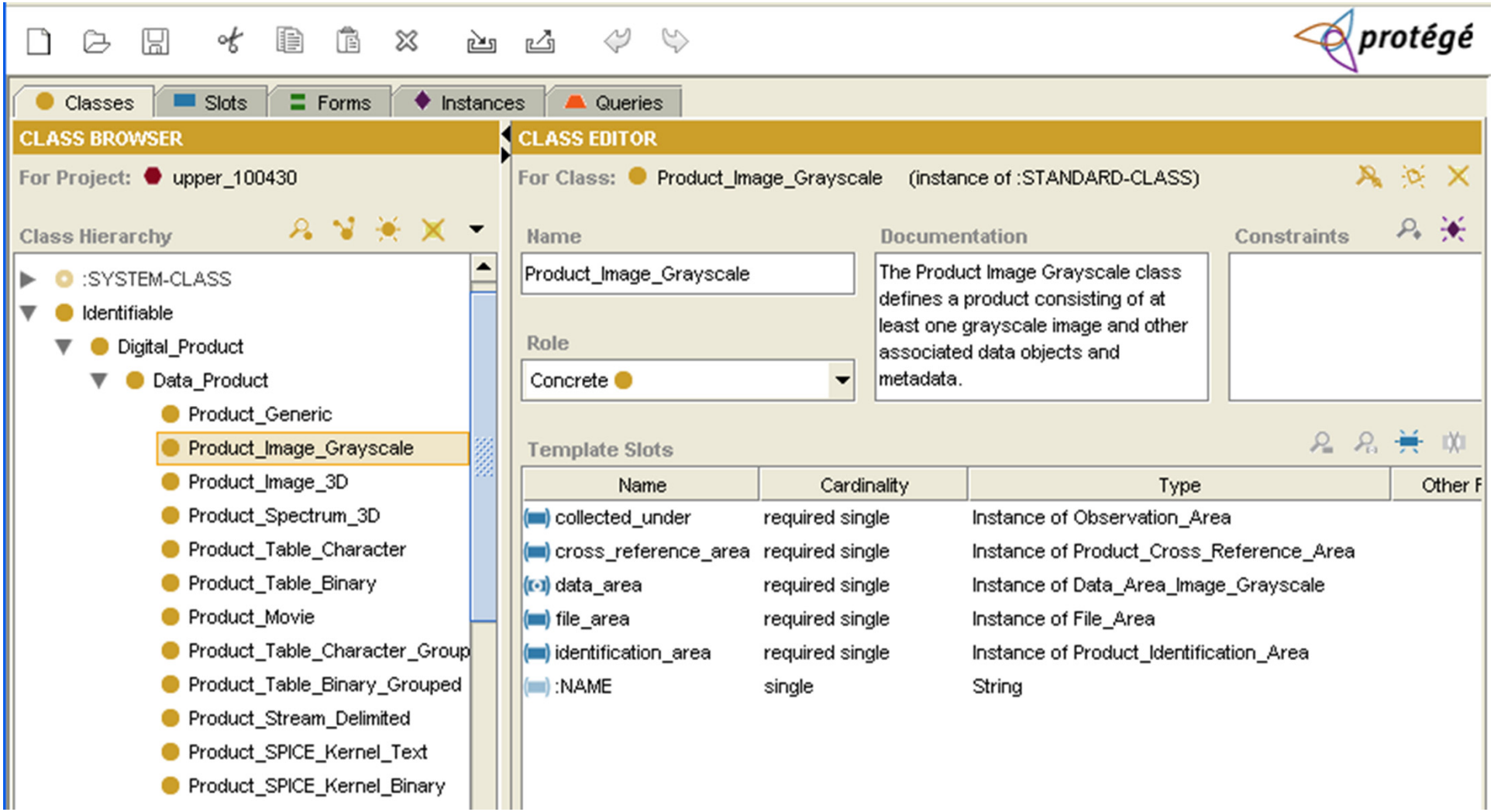
- **physical object**: An object which is physical or tangible (and, therefore, does not itself fit into a digital archive). Examples of 'physical objects' include the planet Saturn and the Venus Express magnetometer.

- **conceptual object**: An object which is intangible (and, because it is intangible, does not fit into a digital archive). Examples of 'conceptual objects' include the Cassini mission and NASA's strategic plan for solar system exploration.

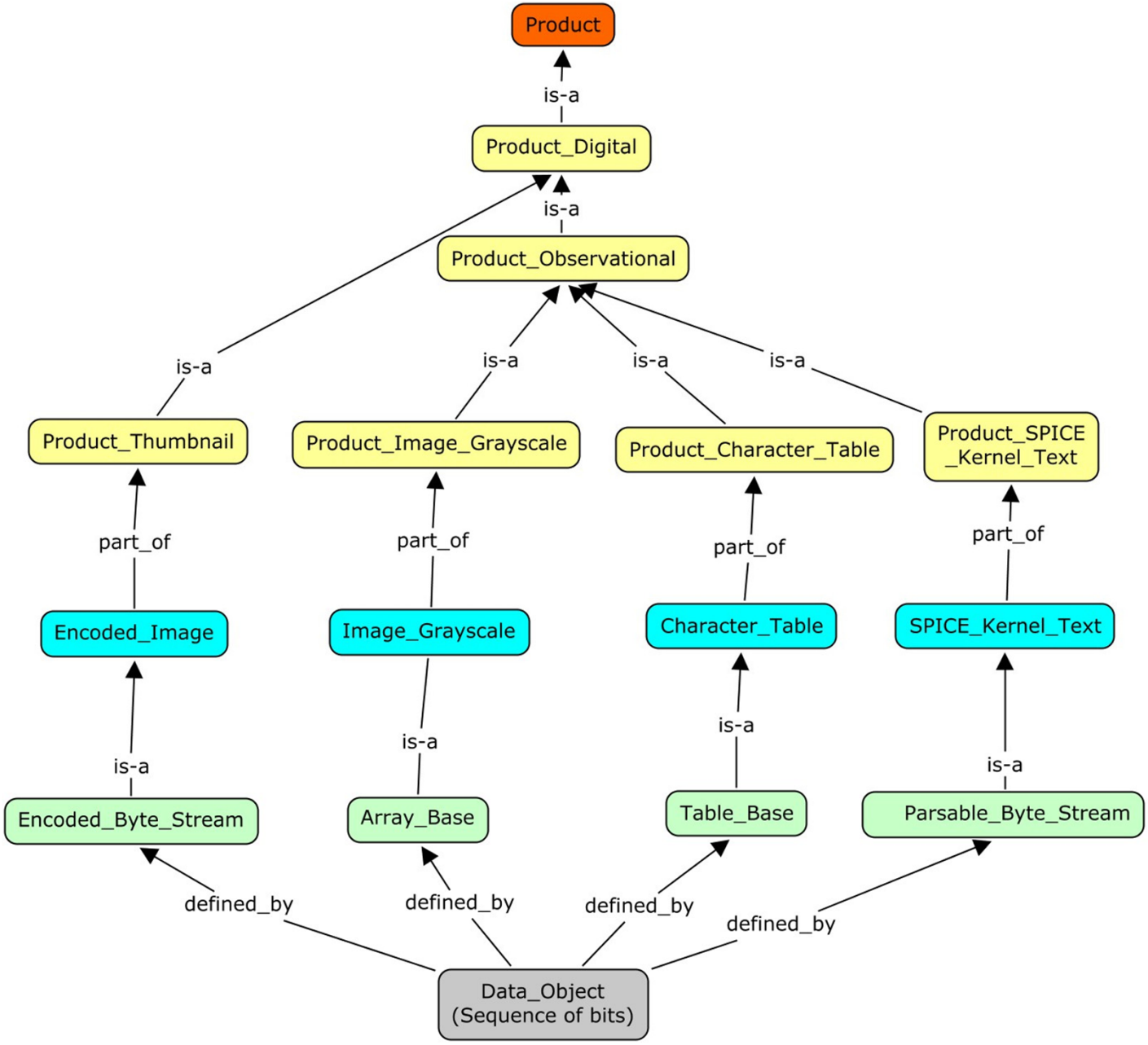
Domain Ontology

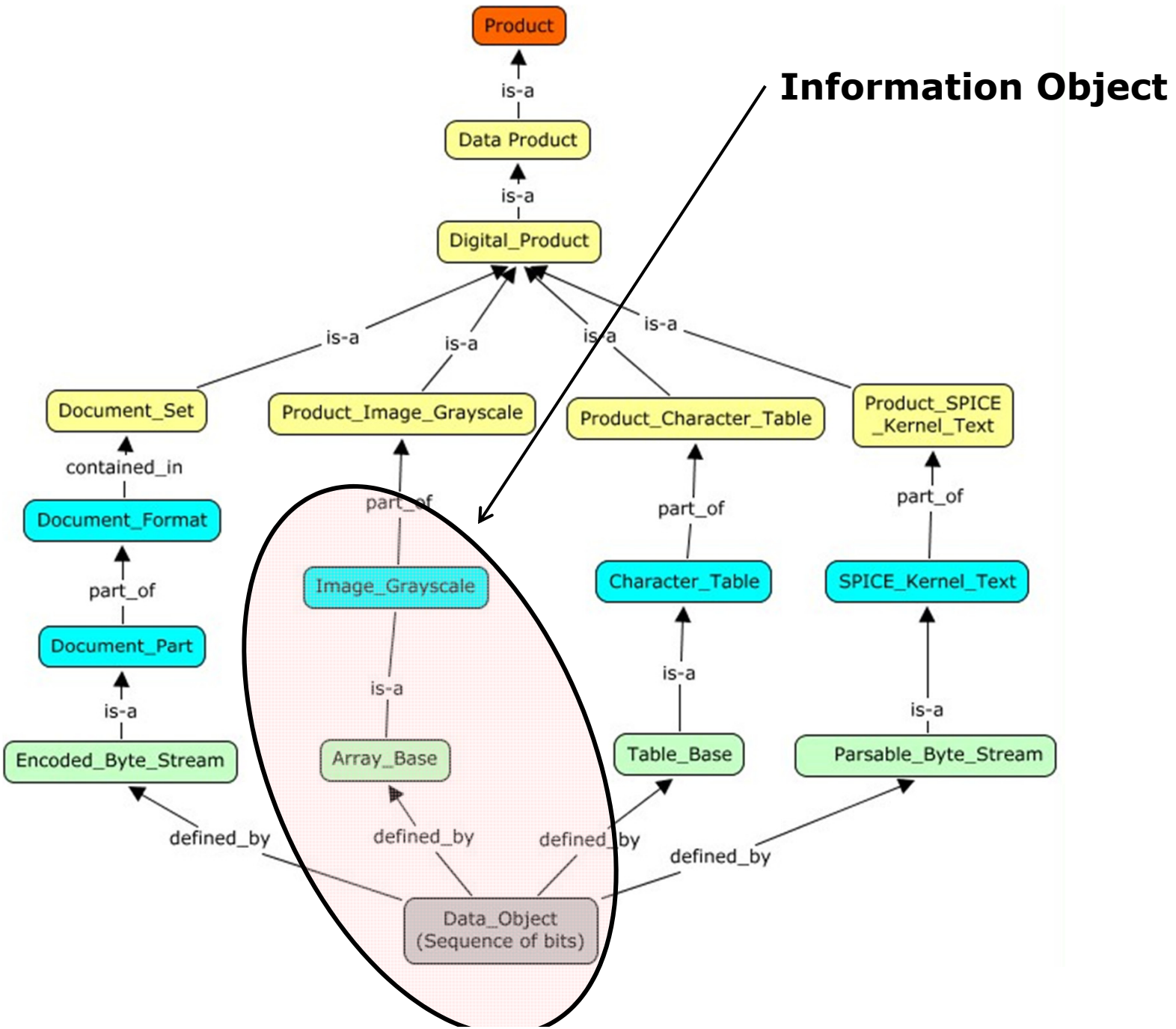
- Used to define things in the science data domain
  - Data structures
  - Metadata for the science interpretation of the data
  - Context within which the data was captured, processed, and archived
  - Relationships between data
  - Provenance of the data
- Once defined the ontology is used to
  - Generate schemas for data labeling and validation
  - Capture and validate all metadata
  - Configure the product registry
  - Generate user documentation
- *Defines the metadata part of an information object*

Ontology - Data Product

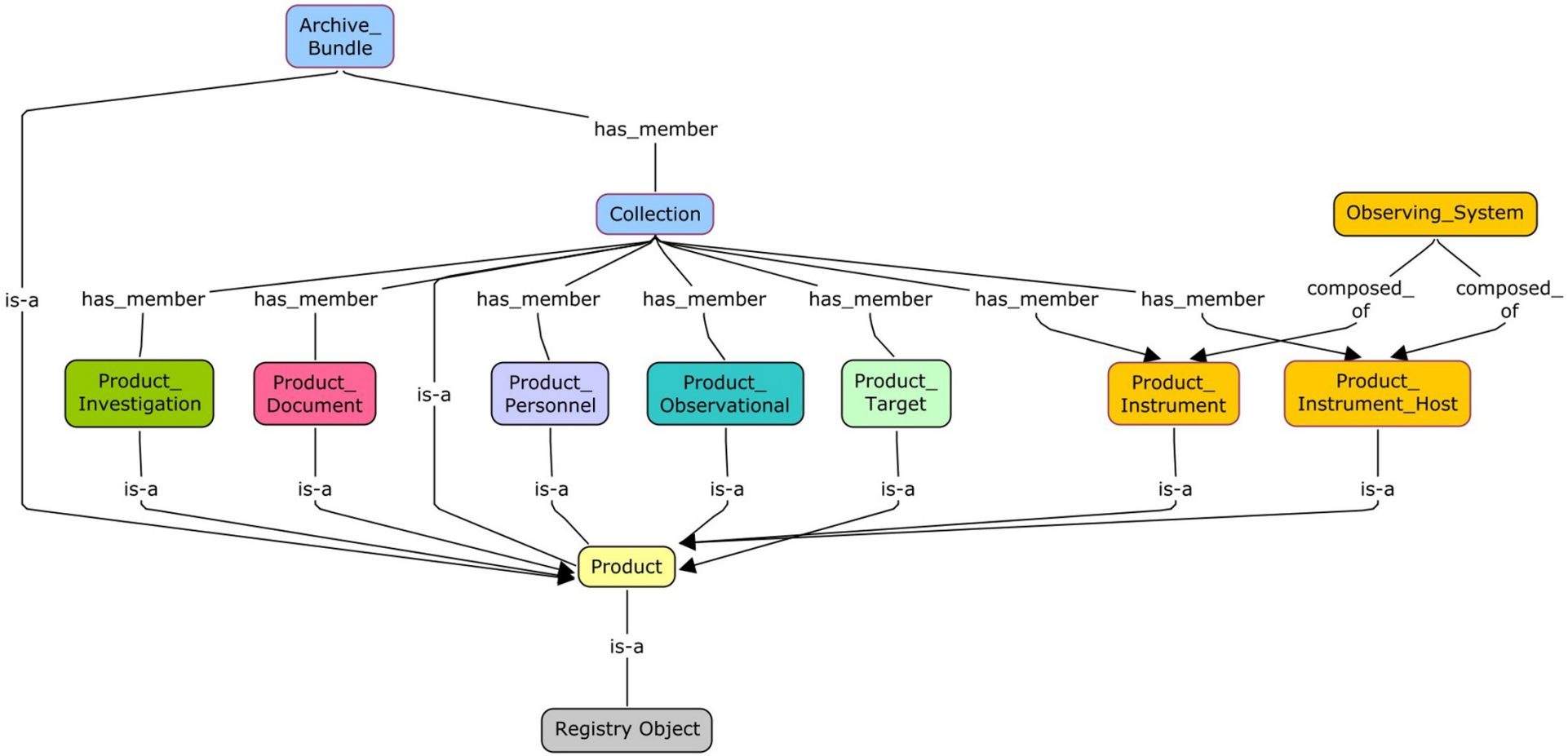


Observational Product – Concept Map

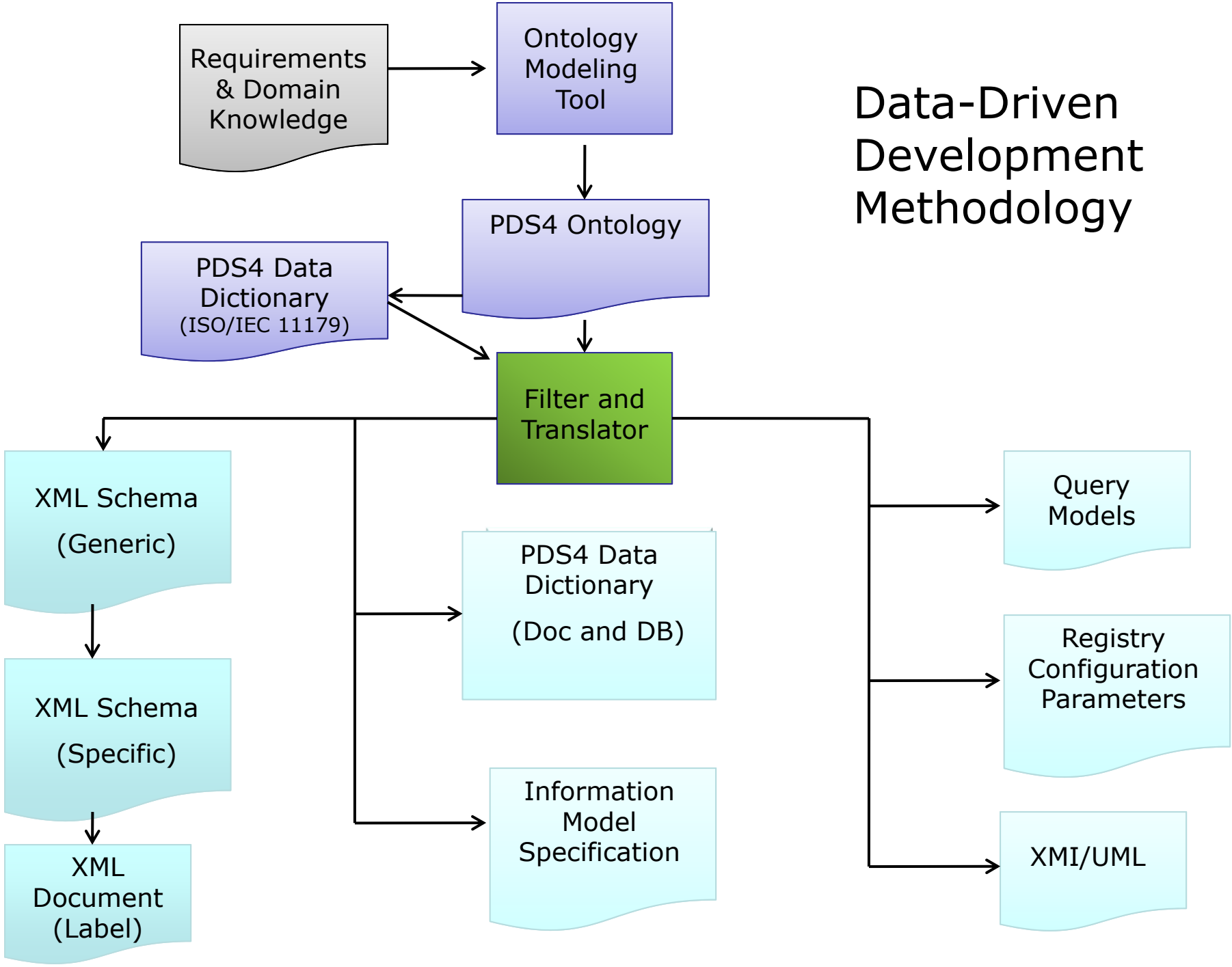



Information Object
Product
is-a
Data Product
is-a
Digital_Product
is-a
is-a
is-a
is-a
Document_Set
Product_Image_Grayscale
Product_Character_Table
Product_SPICE
_Kernel_Text
contained_in
part_of
part_of
part_of
Document_Format
Image_Grayscale
Character_Table
SPICE_Kernel_Text
part_of
is-a
is-a
is-a
Document_Part
Array_Base
Table_Base
Parsable_Byte_Stream
is-a
Encoded_Byte_Stream
defined_by
defined_by
defined_by
defined_by
Data_Object
(Sequence of bits)

High Level Concept Map



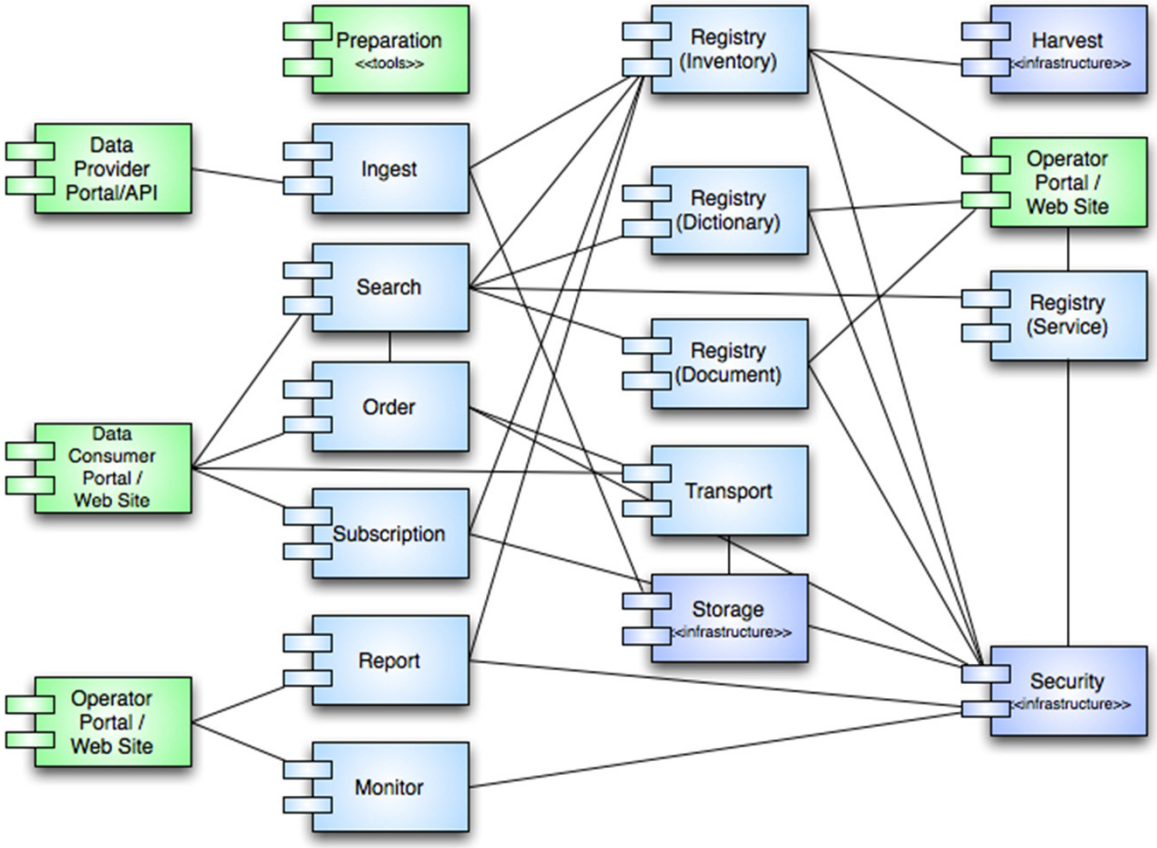
Data-Driven Development Methodology



Federated Registry

- Electronic Business XML (ebXML) federated registry reference model
  - Provides a standard to support federated registry functions
- Extrinsic objects
  - Opaque to registry (content is not known)
  - Configuration file is generated from the domain ontology
  - Notifies registry of product types to expect
    - Provides the associated associations, classification schemes and search attributes descriptions
    - Actual associations, classification schemes, and search attributes for each product are provided at ingestion
- Intrinsic objects
  - Content known to registry (defined in registry schema)
    - Generic Registry Object
    - Associations, Classification Schemes, Slots
- Federated
  - Local governance with federated query, linking, and replication

Registry in Context



Conclusion

- The Planetary Data System (PDS) has adopted this architecture for it next generation information system, PDS 2010.
- The resulting information system will help meet the expectations of modern scientists by providing more data interconnectedness, correlative science, and system interoperability across diverse data sources.

THANK YOU!

Backup

International Collaboration

- PDS Standards are currently used as the de facto standard for archiving planetary science data
  - ESA has adopted PDS Standards

- In 2006, ESA and NASA proposed and started the *International Planetary Data Alliance* to improve efficiency in coordinating data archiving for international missions and improving access to international archives

- In 2008, COSPAR passed a resolution recognizing IPDA and supporting its efforts to establish standards for archiving and sharing planetary science data

- Representatives include: ESA, NASA, ISRO, JAXA, DLR, BNSC, CNES, ASI, CNSA, RSA/IKI with Japan as the current chair

- Information model and registry being assessed for use in the IPDA.

Controlling Standards

- 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.
- Open Archival Information System (OAIS) Reference Model - Provides a standard for the unification of digital, conceptual, and physical data objects.
- Electronic Business XML (ebXML) federated registry/repository information model – Provides a standard to support federated registry/repository functions.
- The Dublin Core Metadata Initiative – Provides a vocabulary of fifteen properties for use in resource description.

Important ISO/IEC 11179 Features

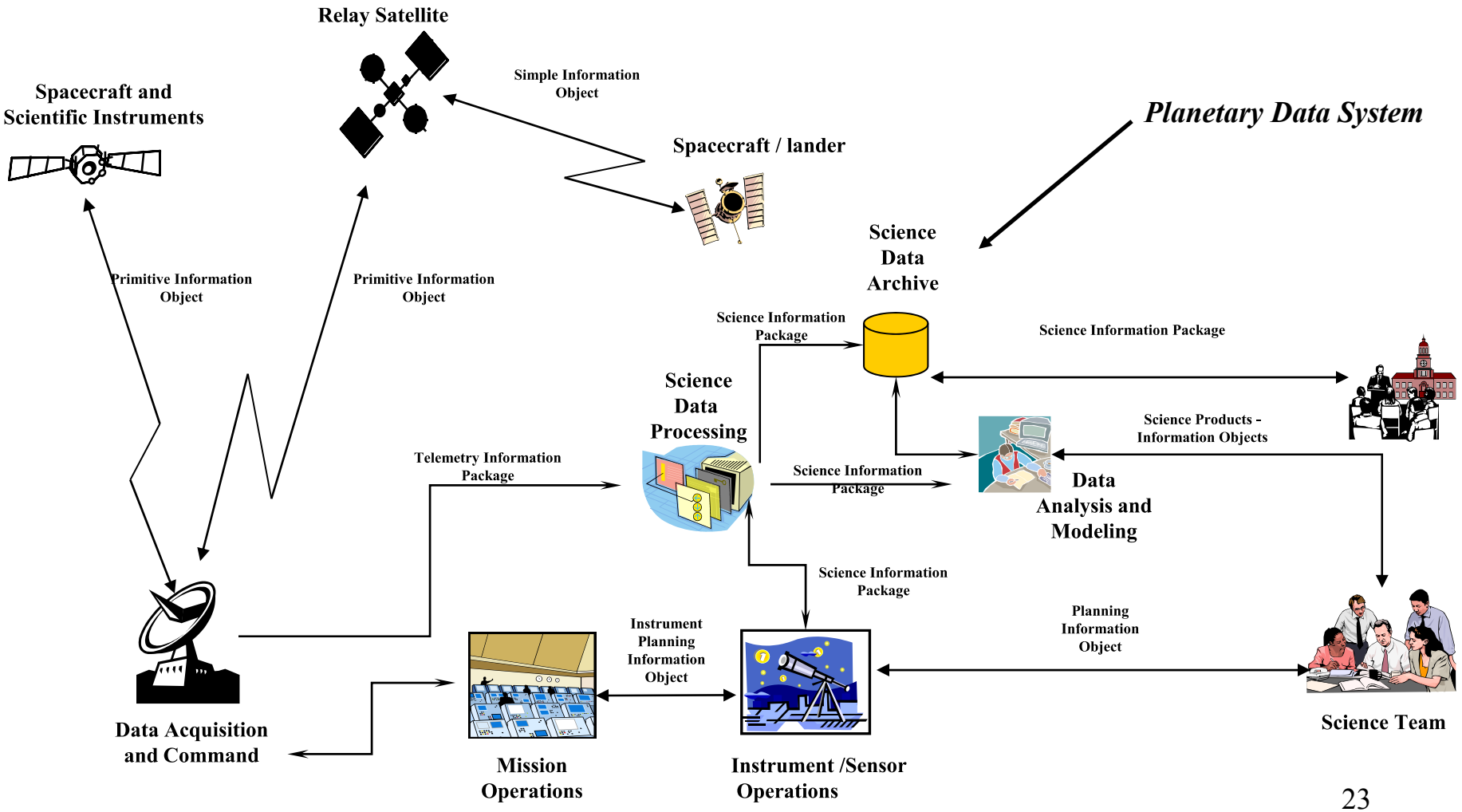
Data Element

- 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

Value Domain

- Permissible Value
- Value Meaning
- 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

PDS's Role in a Distributed Space Systems Architecture



Goals of the Data Architecture

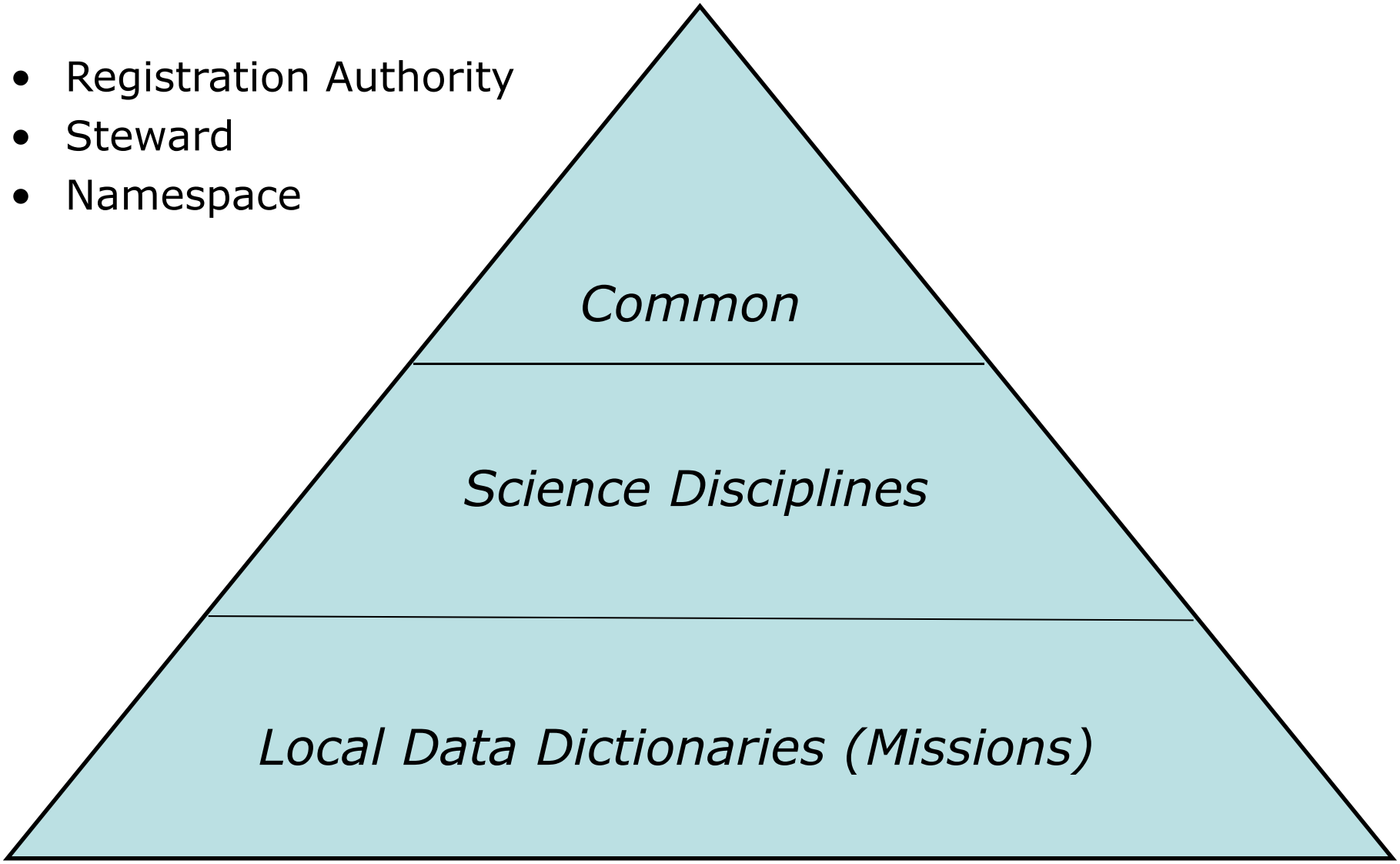
- A transition from a 20-year-old collection of data standards to a modern set of data standards constructed using best practices for standards development.
- Fewer, simpler, and more rigorously defined formats for science data products.
- Use of XML, a well-supported international standard, for data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

- The data dictionary content is tightly coupled with the information model.
  - A Registration Authority is responsible for all attributes in one 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

Data Dictionary Governance

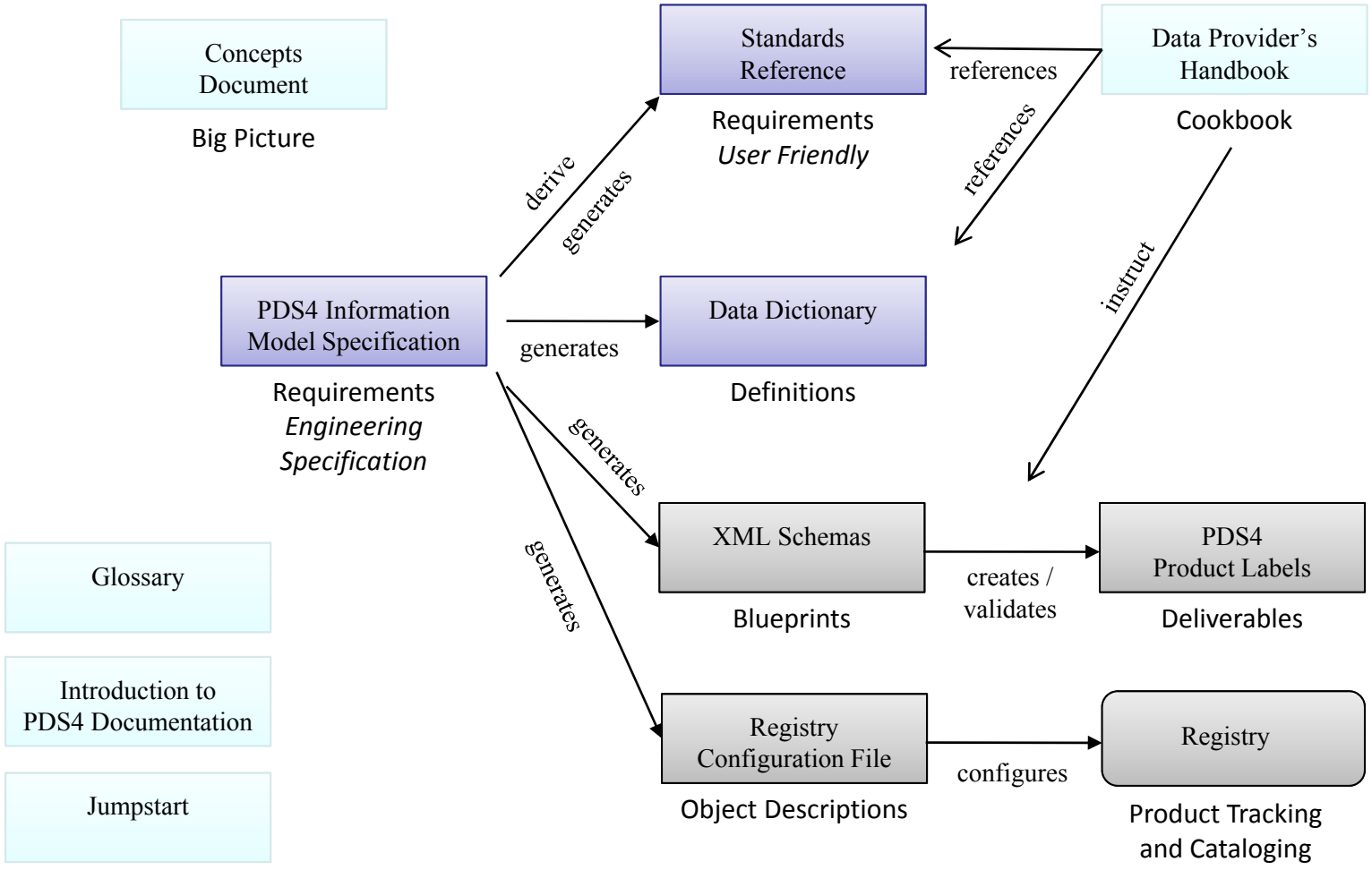
- Registration Authority
- Steward
- Namespace



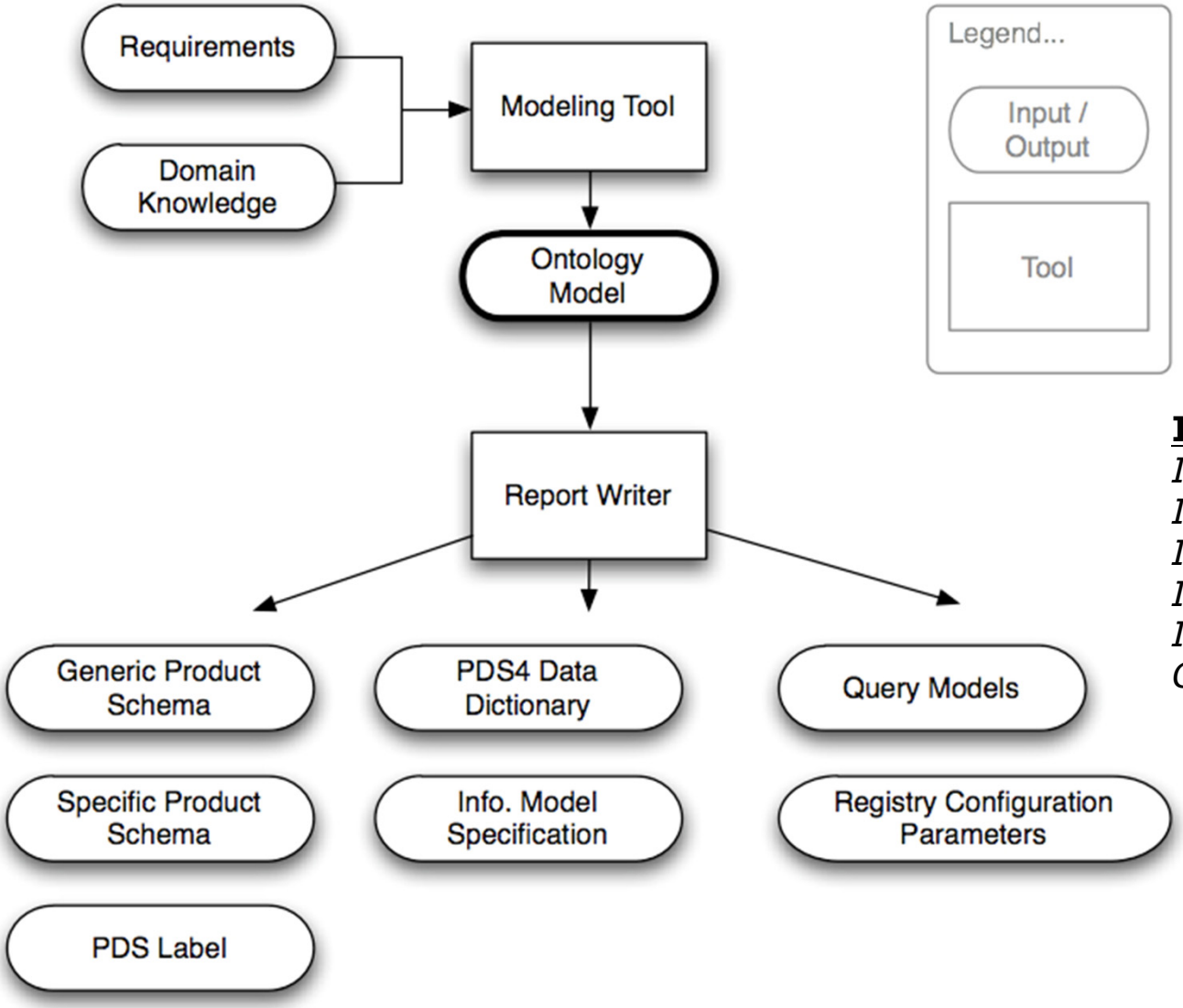
PDS 2010

- Simplified, but rigorous, archiving standards that are consistent, easy to learn, and easy to use
- Adaptable tools for designing archives, preparing data, and delivering the results efficiently to PDS
- On-line services allowing users to access and transform data quickly from anywhere in the system
- A highly reliable, scalable computing infrastructure that protects the integrity of data, links the nodes into an integrated data system, and provides the best service to both data providers and users

PDS4 Documents and their Relationships



Data-Driven Development Methodology



Information Standards

ISO/IEC 19502-MOF
ISO/IEC 11179
ISO/IEC 19503 - XMI
ISO 639-RDF
ISO/IEC 19501 – UML
OWL-DL

Time Line

- Project Start – December 2008
- Build 1 – Test System – October 2010
- Build 2 – Operational System – October 2011

Design Concepts

- Redesign the PDS data standards using current best practices for data system development.
- Design fewer, simpler, and more rigorously defined data formats for science data products.
- Use XML, a well-supported international standard, for data product labeling, validation, and searching.
- Implement a hierarchy of data dictionaries built to the ISO 11179 standard, designed to increase flexibility, enable complex searches, and make it easier to share data internationally.