National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

PDS: A Model-Driven Planetary Science Data Architecture for Long-Term Preservation

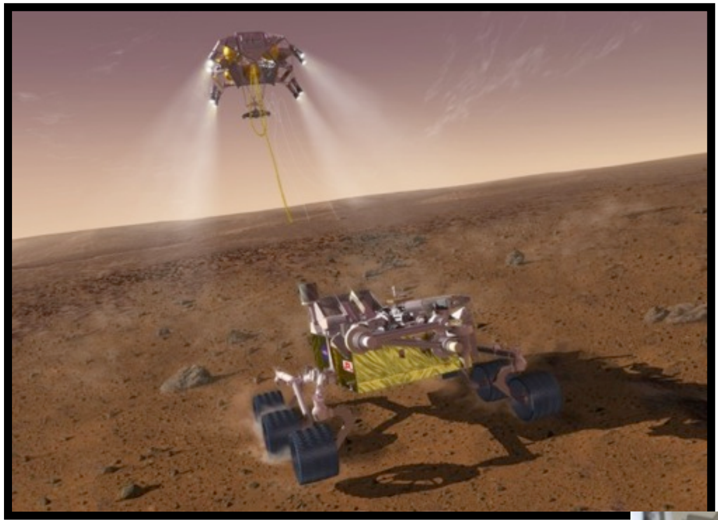
LOPS@ICDE 2014

**J. S. Hughes, D. Crichton, S. Hardman,
E. Law, R. Joyner, P. Ramirez**

Topics

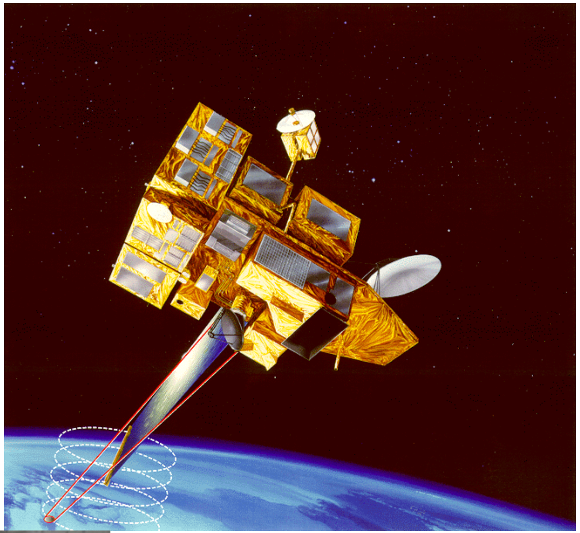
- Big Data in the Space Sciences
- The Planetary Data System
- PDS4: The Next Generation PDS
- The PDS4 Information Model
- The PDS4 System Architecture
- Recommendations

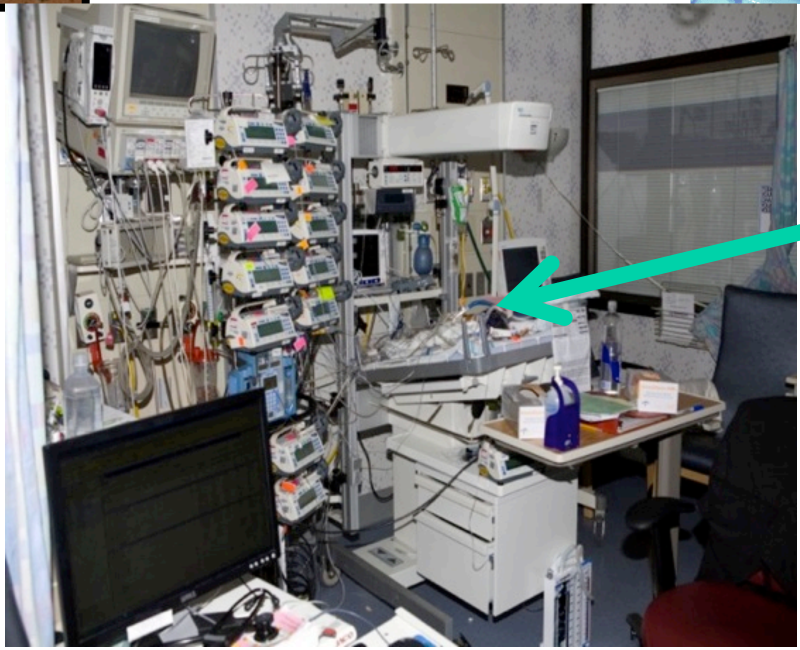
Observational Science Platforms



?

What do these have
In common?

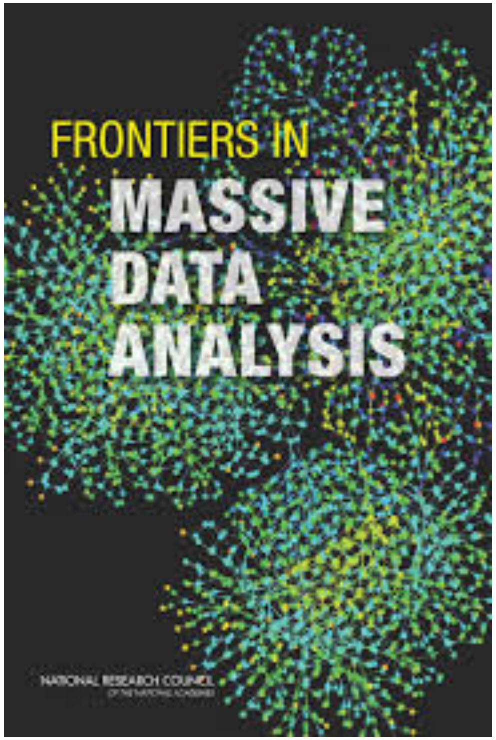




What's being
observed

NRC Report: Frontiers in the Analysis of Massive Data

- Chartered in 2010 by the National Research Council
- Chaired by Michael Jordan, Berkeley, AMP Lab (Algorithms, Machines, People)
- Consideration of the architecture for big data management and analysis
- Importance of systematizing the analysis of data
- Need for end-to-end lifecycle: from point of capture to analysis
- Integration of multiple discipline experts
- Application of novel statistical and machine learning approaches for data discovery



2013

- A Major Shift from Compute-Intensive to Data-Intensive -

Technology Trends*

- Distributed systems (access, federation, linking, etc)
- Scalable infrastructures and technologies for optimizing compute and data-intensive applications
- Service-oriented architectures
- Ontologies, models and information representation
- Scalable database systems with different underlying models
- Federated data security mechanisms
- Technologies for moving large data sets

* Frontiers in Massive Data Analysis (2013)

A Disciplined, Architectural Approach

- **Consider the architectural (information, software) viewpoint**
  - *Address the data definition and lifecycle from point of collection to data integration and analysis*
  - ***Separate the technical infrastructure from the data to drive an overall data architecture on top of a scalable, big data infrastructure***
  - *Apply advanced computer science techniques to address data access, discovery, integration, and extraction across highly distributed environments to support data analytics*

- **Adapt, adopt, develop and research techniques and technologies for increasing the efficiency of data analysis for distributed environments**
  - *Reduce time to perform analytics by distributing the computation*
  - *Develop mechanisms for comparing measurements against predictive models*
  - *Manage the balance between sampling strategies and uncertainty in inferences*

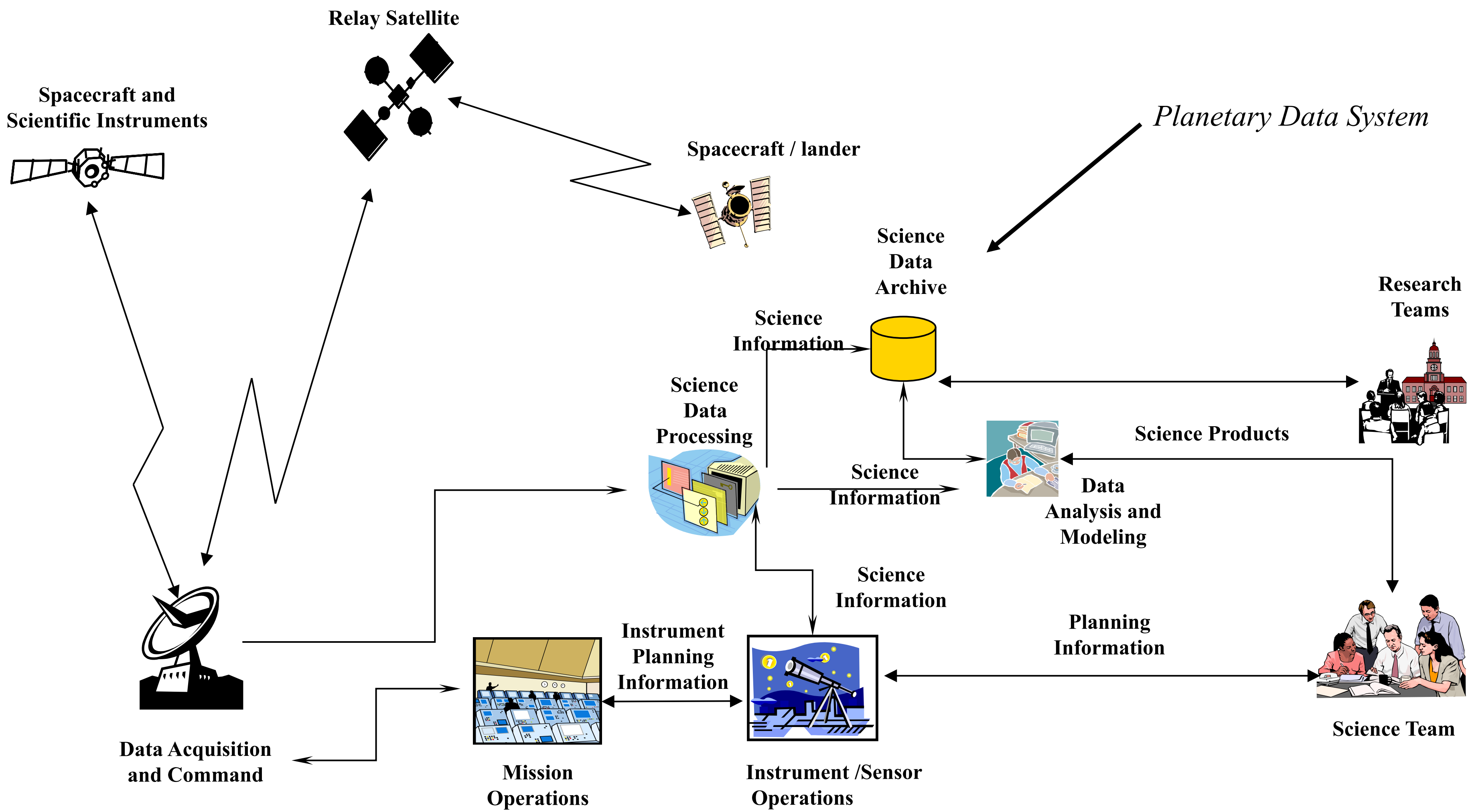
Challenges in Space Data Systems

- **Space systems and instruments are deployed world-wide; data is generated across complex, multi-organizational systems**
  - *Many producers of data*
  - *Data is managed in highly distributed environments*
  - *Limited data sharing occurring between organizations*

- **Systems are very heterogeneous**
  - *Data systems are often developed around the point of collection*
  - *Access to data has traditionally been difficult*
  - *Data is represented in different formats and structures*

- **Massive data sets are being generated challenging traditional analysis approaches**

The Planetary Data System



Credit: CCSDS Reference Architecture for Space Information Management

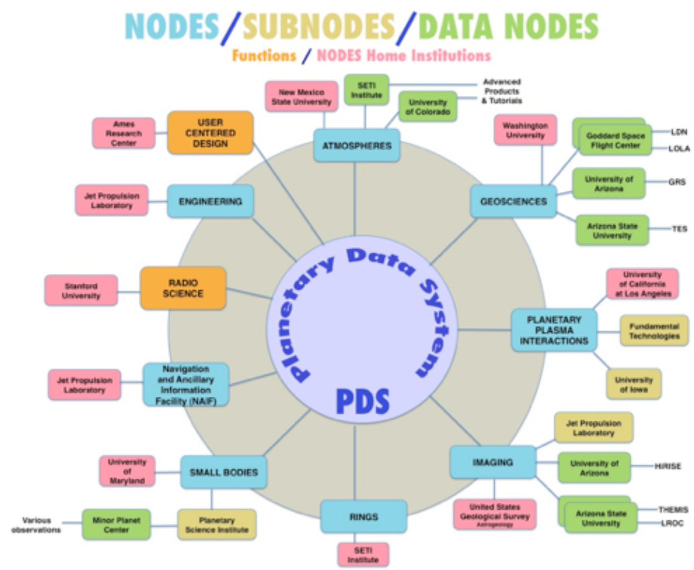
National Aeronautics and
Space Administration

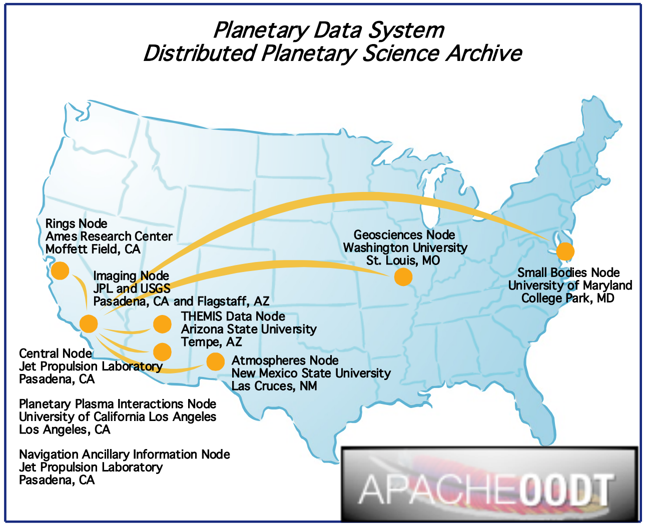
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

NASA Planetary Data System: The Planetary Science Archive

- <u>Purpose:</u> to collect, archive and make accessible digital data and documentation produced from NASA's exploration of the solar system

- <u>Infrastructure:</u> a highly distributed infrastructure with planetary science data repositories implemented at major government labs and academic institutions
    - **All data is captured based on a common set of data standards (models, structures, etc)**
    - Approximately 600 TBs of data
    - Movement towards international interoperability
    - Implemented an open source cyerinfrastructure developed at JPL (Apache OODT)
    - Movement to an information-model driven architecture





Timeline of PDS Technical Implementations and Upgrades

- **PDS 1 – < 1990**
  - *High-Level Catalog for finding data sets by mission, instrument, spacecraft and target.*
  - *Archive volumes stored and distributed on tape.*
  - *The Object Description Language (ODL) is invented for product labeling and capturing catalog information.*
- **PDS 2 - 1990**
  - *CD-ROM becomes the archive and distribution volume of choice.*
  - *High-Level Catalog simplified by using more text instead of keywords to capture descriptive information.*
- **PDS 3 - 1992**
  - *PDS sets up and maintains a web presence.*
  - *Movement to online distribution of products (PDS-D). (~2002)*
  - *On-line mass storage and data bricks replace CD/DVD as archive and distribution media.*
- **PDS4 - 2010**
  - *Movement to a distributed, service architecture*
  - *Integrated federation*
  - *New data standards, data formats and structures*
  - *International Collaboration*

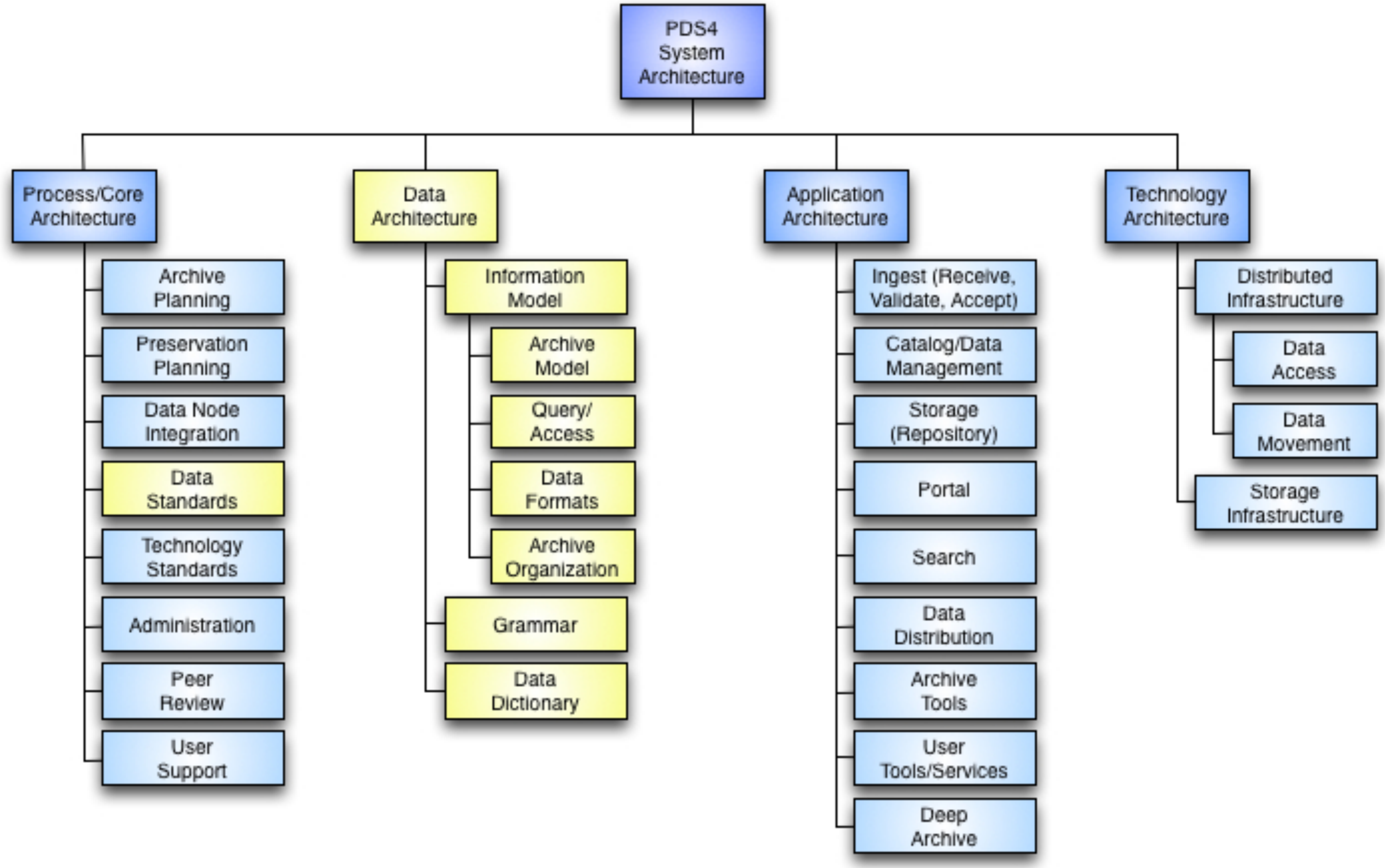
PDS4: The Next Generation PDS

- The NASA Planetary Data System (PDS) after about 20 years of operations is developing PDS4, a major revision and transition to a modern system based on best practices for data system development.
  - *A single information model*
- PDS4 will have fewer, simpler, and more rigorously defined formats for science data products.
- PDS4 will use XML, a well-supported international standard, for data product labeling, validation, and searching.
- PDS4 incorporates a hierarchy of data dictionaries built to the ISO/IEC 11179 standard.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

PDS4 Architecture



The PDS4 Information Model

- Defines the data structure (format)
- Defines the science interpretation of the data
- Defines the context within which the data was captured, processed, and archived
- Defines the relationships between the data

The Design Principles of the Information Model

- The information model should remain independent of its implementation.
  - Disentangles the model from the implementation
  - Information model evolves independent of information technology

- A changing domain suggests that the information model should drive both the development and management of the information system.

- The modeling language should be semantically richer than the other languages in the framework.

Knowledge Acquisition for the Information Model

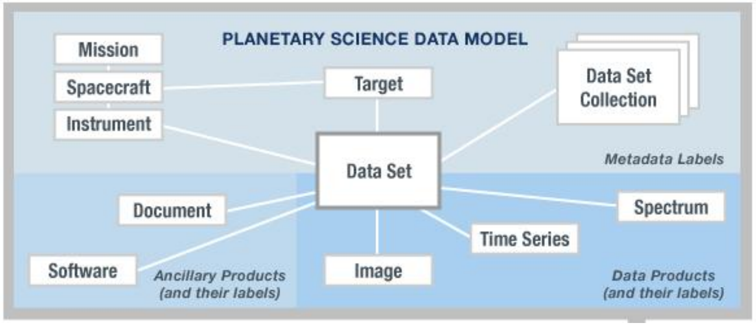
- Domain expertise is captured in an ontology.
  - A working group was formed with at least one domain expert from each of the science disciplines.
  - Each thing-of-interest in the domain was defined and then related to other things-of-interest.
  - The resulting model represents the consensus of domain experts across the PDS science and engineering disciplines.
- The model is subsequently used as the single authoritative source for the PDS4 Data Standards.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

The Primary Role of the Information Model

Information Model



Expressed As

Planetary Science Data Dictionary

Data Element

has

Class

Extracted/Specialized

Label Schema

Used to Create

Validates

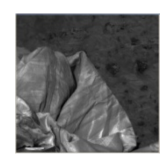
Product

Tagged Data Object

(Information Object)

```
<local_identifier>MPFL_M_IMP_IMAGE</local_identifier>
   <offset unit="byte">0</offset>
   <axes>2</axes>
   <axis_index_order>Last_Index_Fastest</axis_index_order>
   <encoding_type>Binary</encoding_type>
   <Element_Array>
      <data_type>SignedMSB4</data_type>
      <unit>pixel</unit>
   </Element_Array>
   <Axis_Array>
      <axis_name>Line</axis_name>
      <elements>248</elements>
      <sequence_number>1</sequence_number>
   </Axis_Array>
   <Axis_Array>
      <axis_name>Sample</axis_name>
      <elements>256</elements>
      <sequence_number>2</sequence_number>
   </Axis_Array>
</Array_2D_Image>
```

Describes

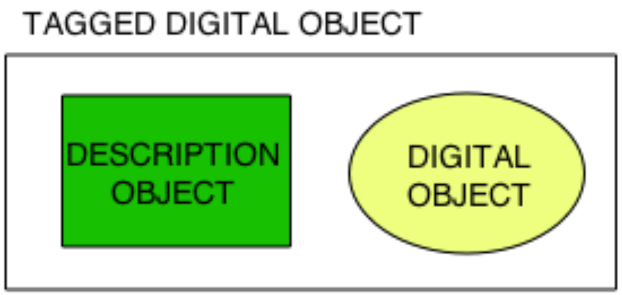


Data Object

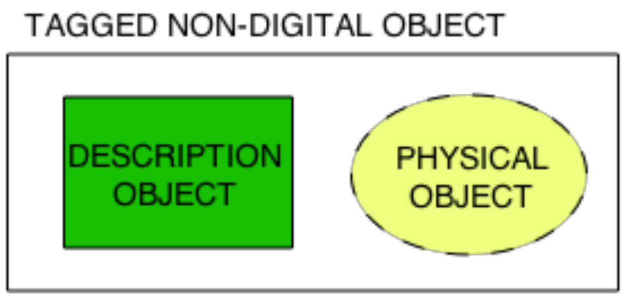
National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

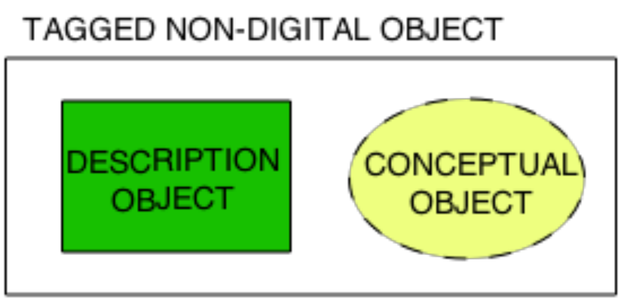
Information Object Model [1]



- digital object: An object which is real data — for example, a binary image of a redwood tree.



- physical object: An object which is physical or tangible – for example the planet Saturn and the Venus Express magnetometer.



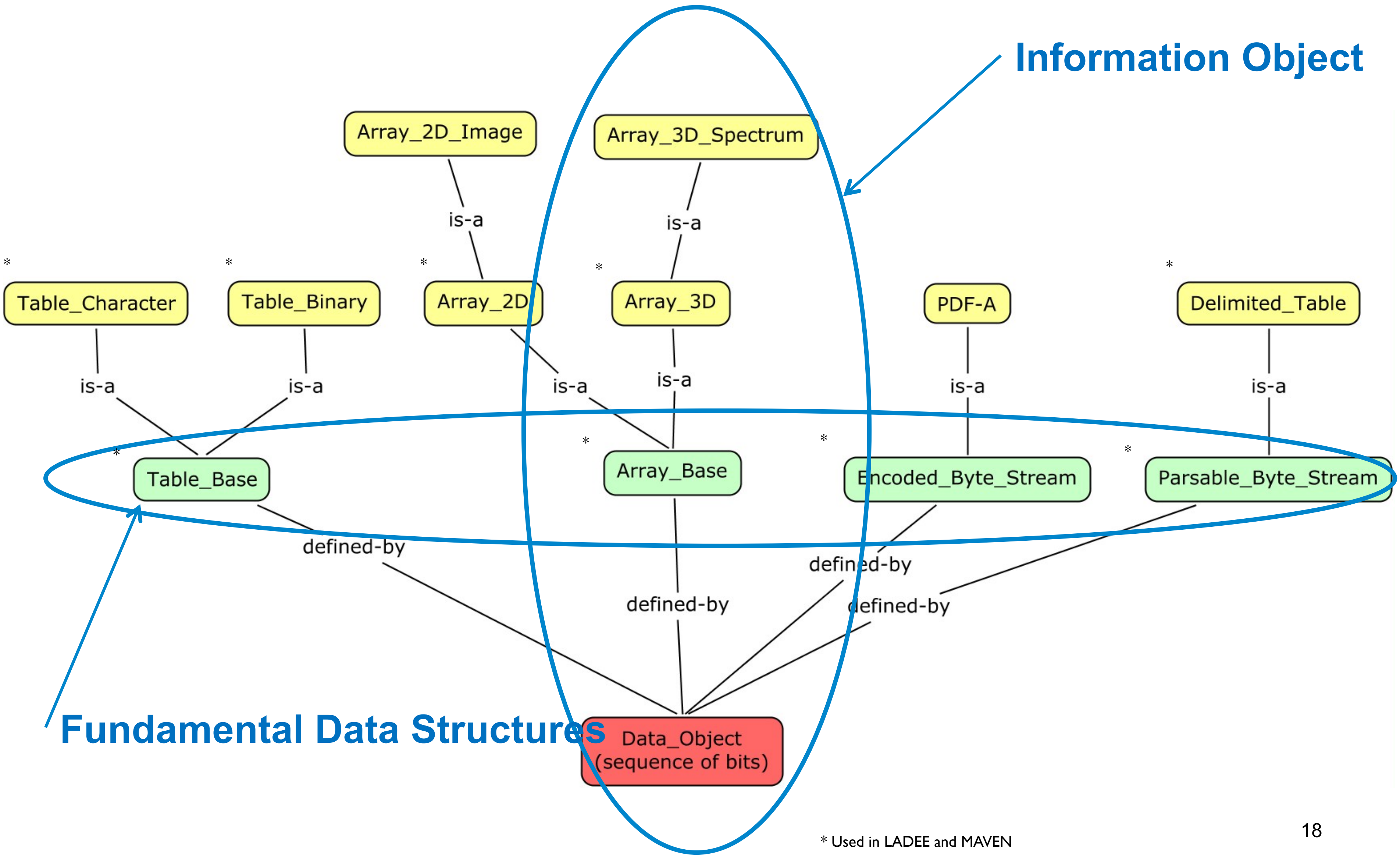
- conceptual object: An object which is intangible – for example the Cassini mission and NASA's strategic plan for solar system exploration.

[1] Open Archival Information System (OAIS) Reference Model - ISO 14721:2003

National Aeronautics and
Space Administration

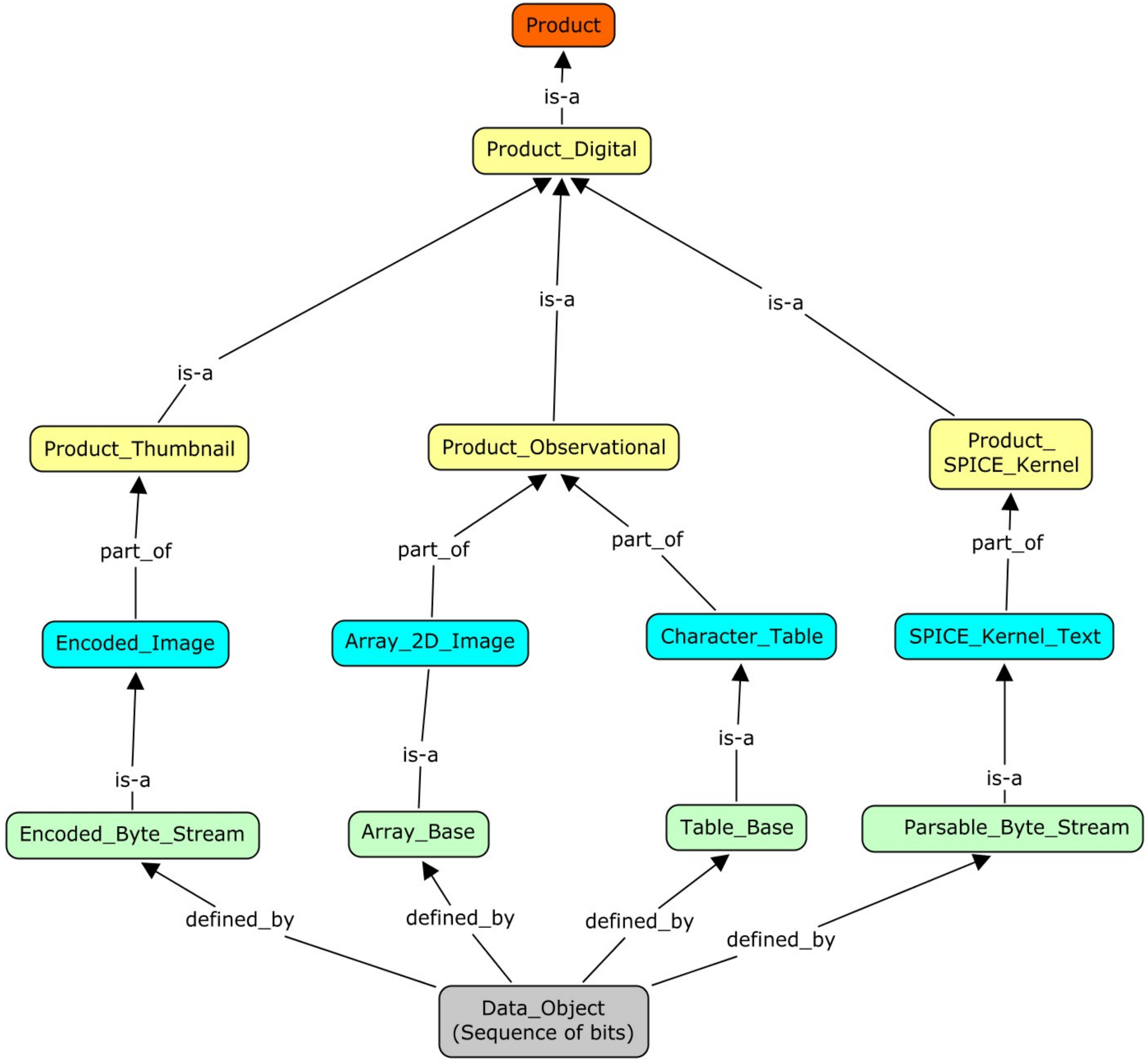
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

PDS4 Data Formats



* Used in LADEE and MAVEN

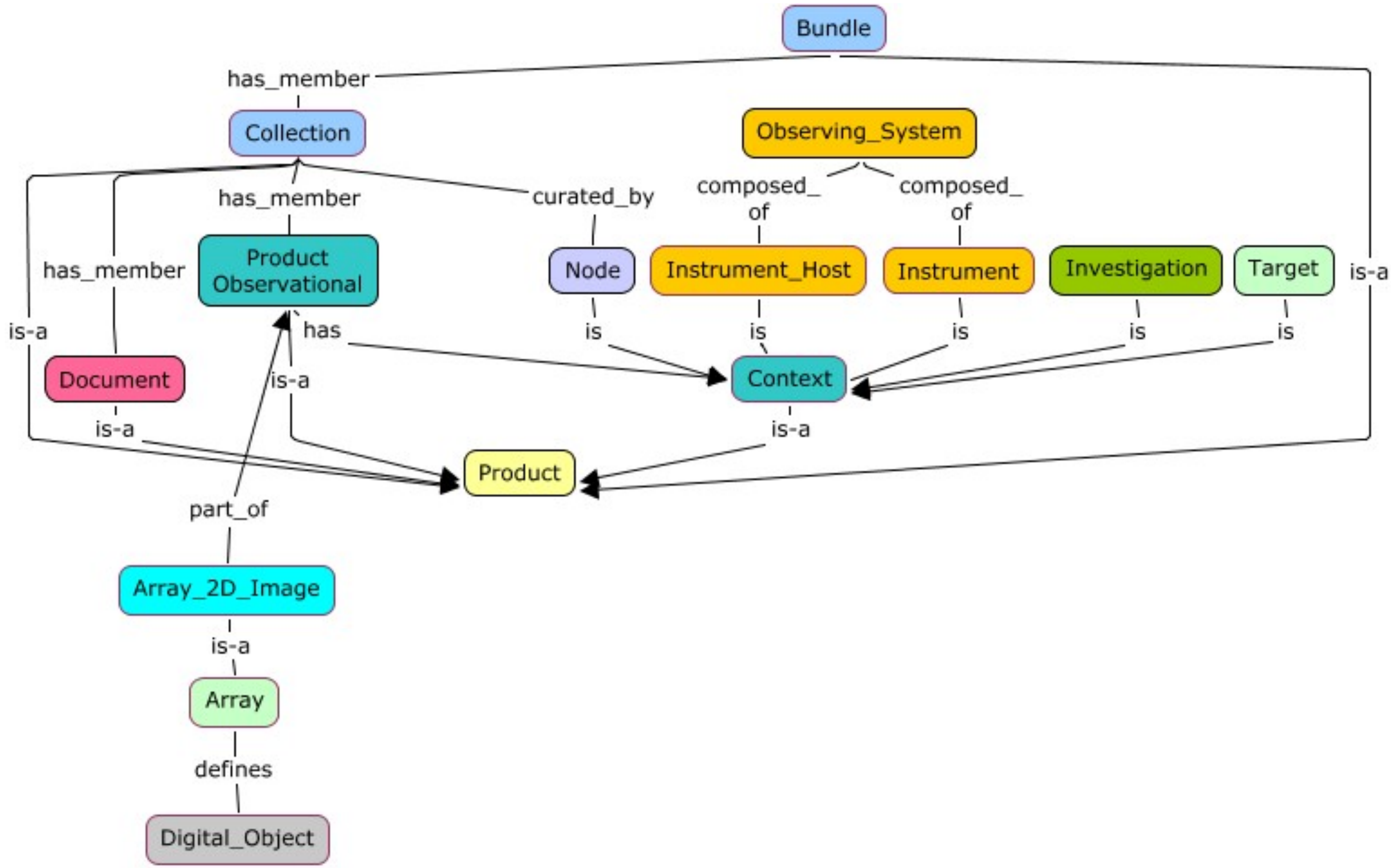
PDS4 Product Model



National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

PDS4 Information Model Concept Map



National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

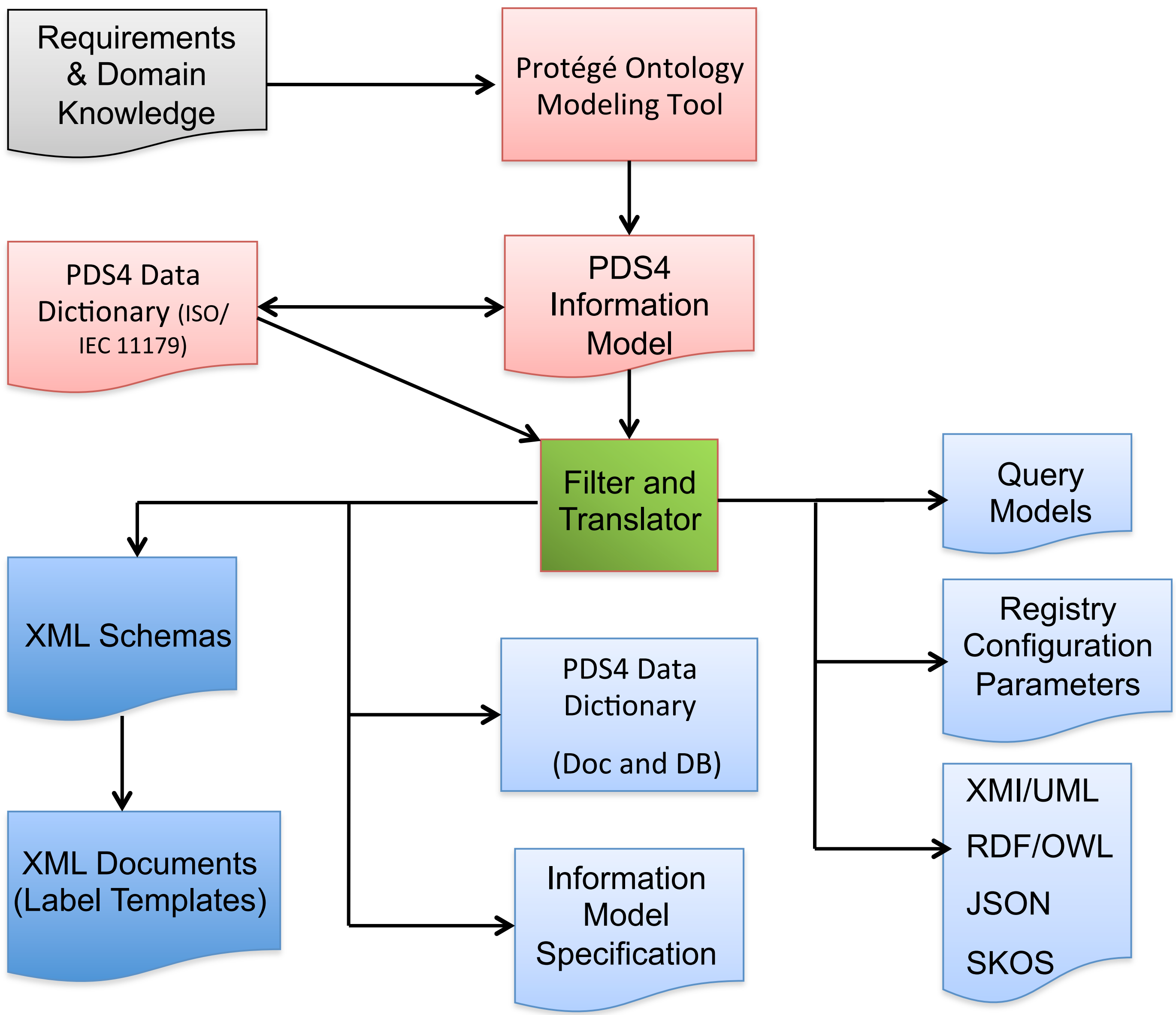
The Information Model Driven Process

- The model is updated frequently to reflect design decisions.

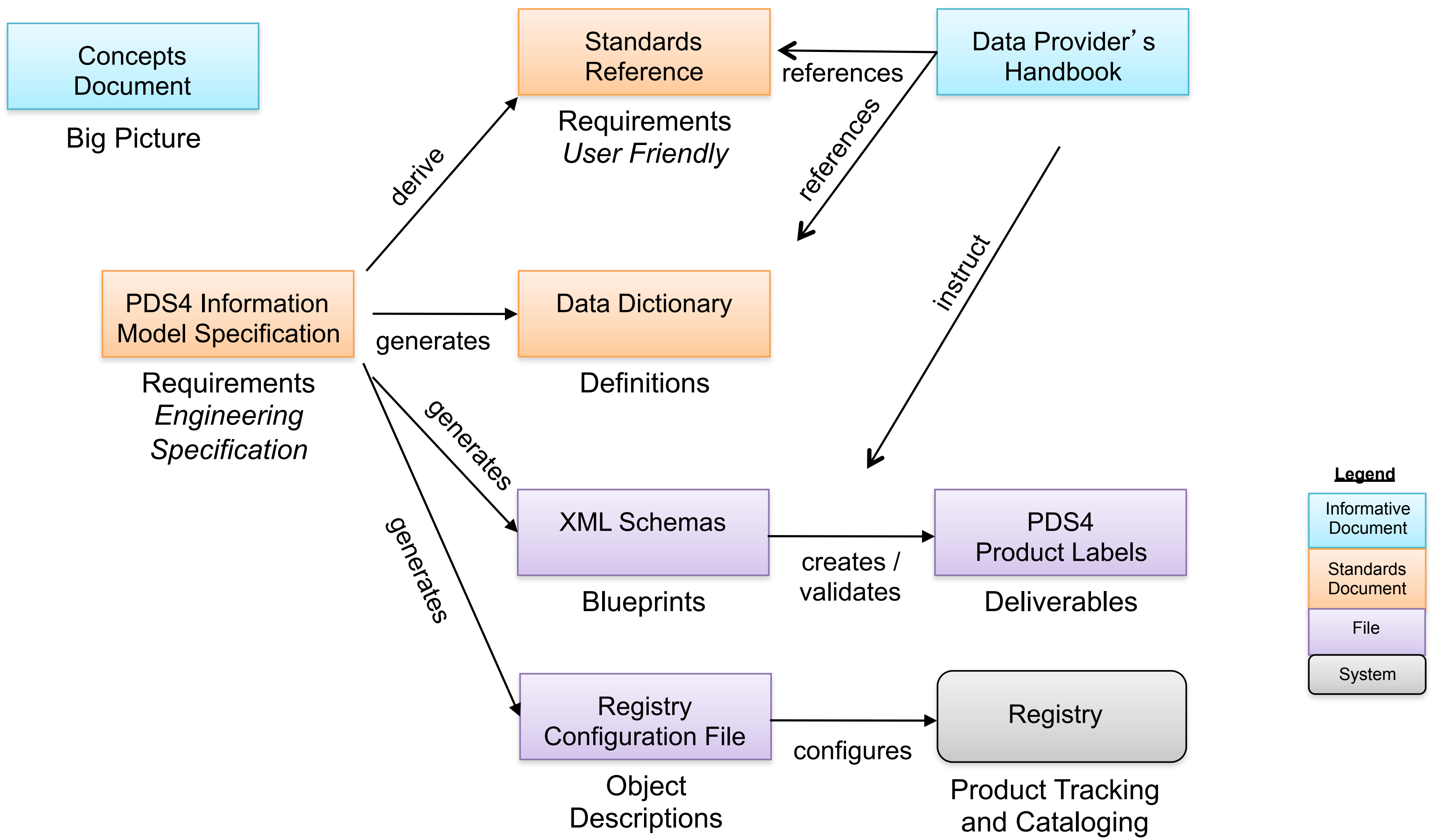
  Using Protégé as a modeling tool

- 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.



PDS4 Documents, Artifacts, and their Relationships



Product Label Template

Identification_Area
    Logical_Identifier
    Version_Id

Observation_Area
    Time_Coordinates
    Primary_Result_Summary
    Investigation_Area
    Observing_System
    Target_Identification

Discipline_Area
Mission Area

Reference_List
    Internal_Reference
    External_Reference

File_Area_Observational
    File
        Header
        Array_2D_Image

Industry Standards* Referenced and Controlling

- ISO 14721:2003 - Open Archival Information System (OAIS) Reference Model - Provides a standard for information objects.
- ISO/IEC 11179:3 Registry Metamodel and Basic Attributes specification - Adopted for the data dictionary schema.
- Reference Architecture for Space Information Management (RASIM) - CCSDS 312-0.G-1 – Provides the overarching architectural principles.
- 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

National Aeronautics and Space Administration

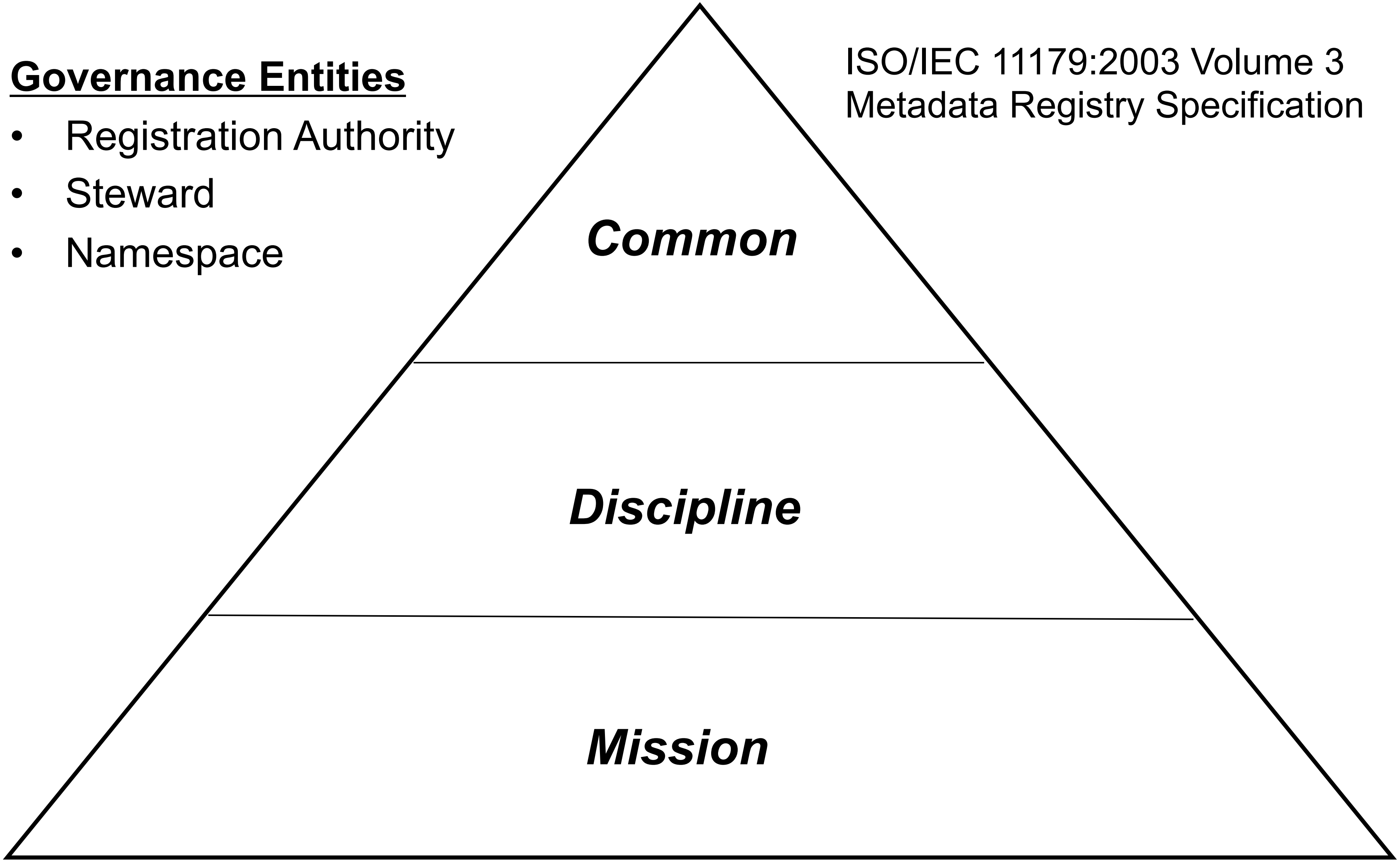
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Data Dictionary Governance

Governance Entities

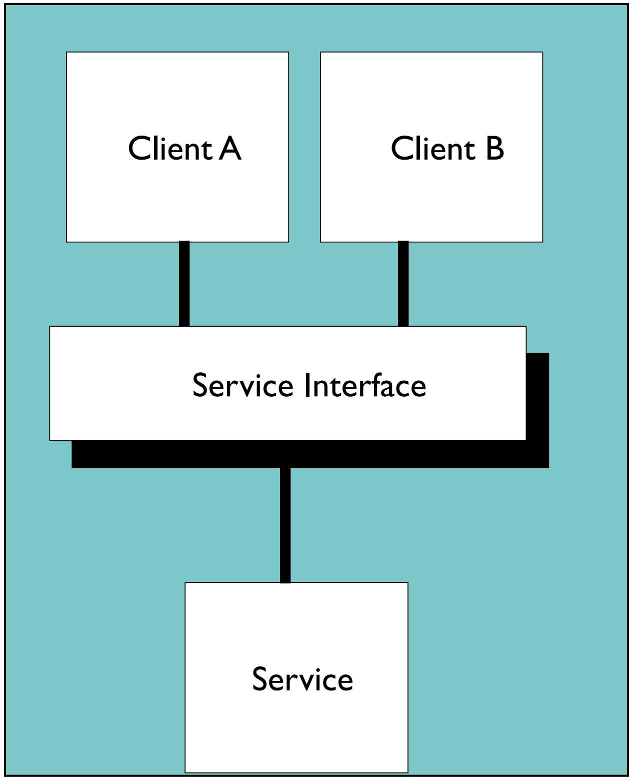
- Registration Authority
- Steward
- Namespace

ISO/IEC 11179:2003 Volume 3
Metadata Registry Specification

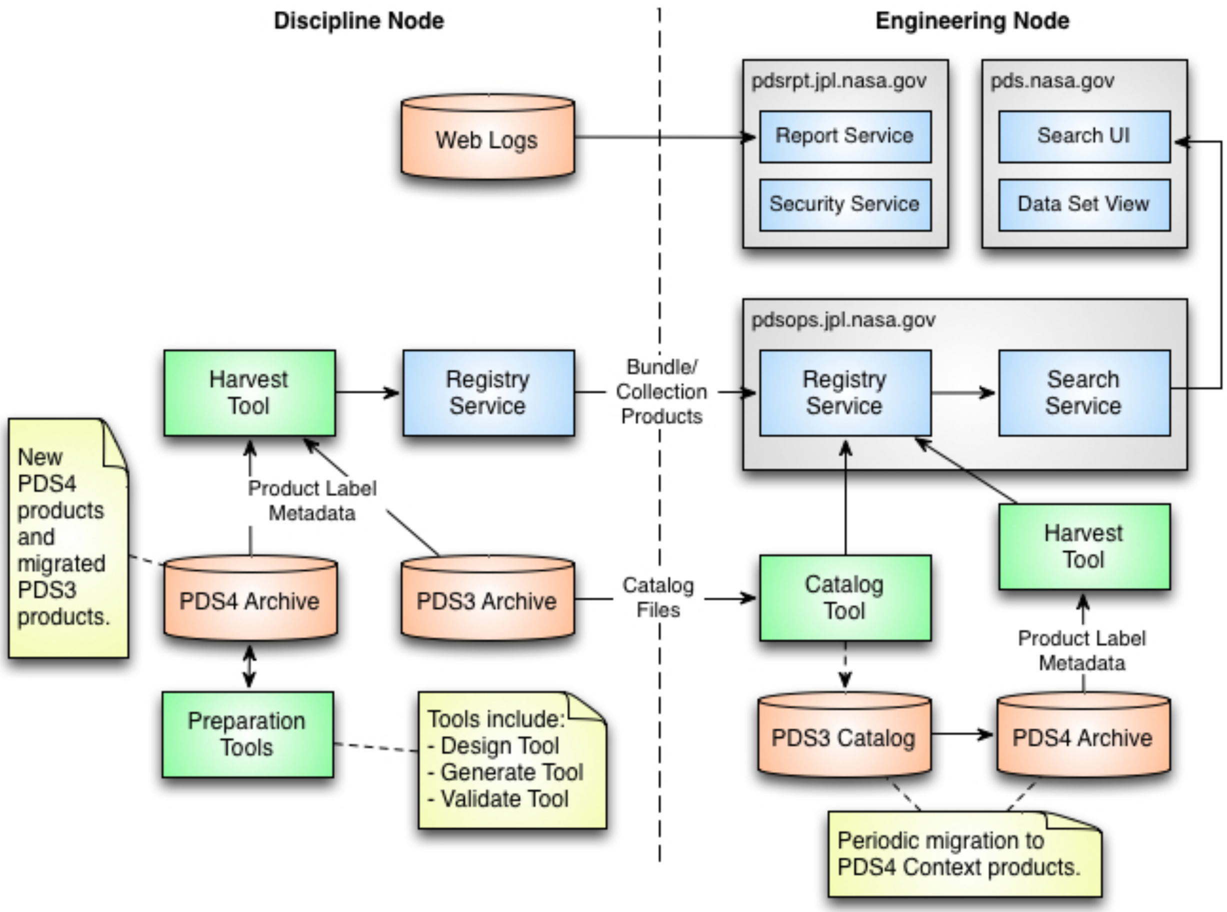


System Design Approach

- **Based on a distributed information services architecture (aka SOA-style)**
  - *Allow for common and node specific network-based services.*
  - *Allow for integrating with other international systems*
- **System includes services, tools and applications**
- **Use of online registries across the PDS to track and share information about PDS holdings**
- **Implement distributed services that bring PDS forward into the online era of running a national data system**
- **Use and contribute back to open source (e.g., Apache OODT, Apache SOLR, Apache Tika, etc)**



PDS Deployment



OODT: An Open Source Framework for Building Data Intensive Science Systems

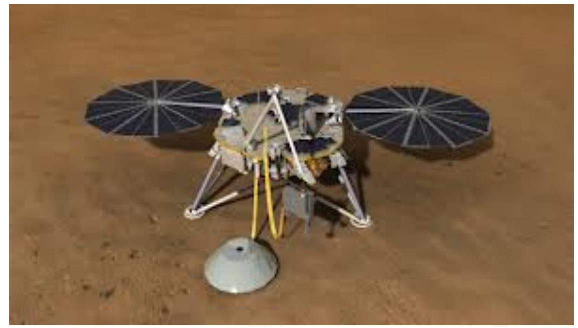
http://oodt.apache.org

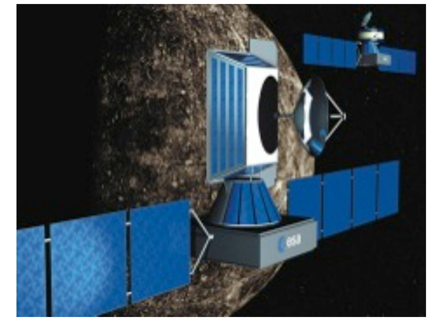
- An open source data management framework to support science data system implementation
  - Developed at NASA/JPL
  - Top Level Project at the Apache Software Foundation (2011)
  - Used across multiple centers (JPL, GSFC, Langley)
  - Used across multiple agencies (NASA, NIH, NSF, DARPA, NOAA)
  - Integrates with an information architecture (e.g., earth science, biomedicine, etc)
  - Significantly reduces cost and increases performance of science data processing and management systems
- Applied to multiple Earth Science missions
  - Seawinds, OCO-2, SMAP, NPP Sounder Peate, JPSS
  - CARVE, Airborne Snow Observatory
- Applied to Earth science, planetary science, astronomy, biomedicine, defense

National Aeronautics and Space Administration

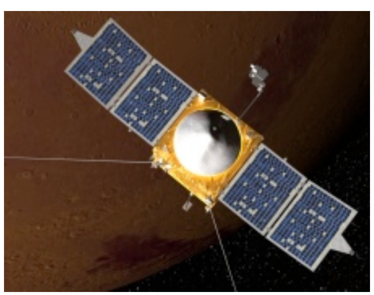
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

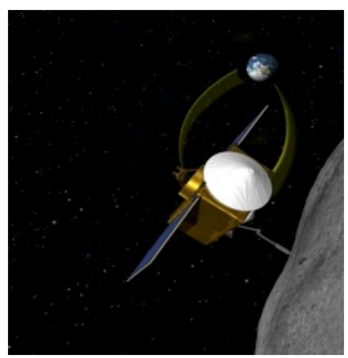
Internationalization of Massive Planetary Science Data: Architecture and Standards


InSight (NASA)


BepiColombo (ESA/JAXA)

Planetary Data System Version 4
International, distributed, model-driven data architecture for capturing, managing and distributing planetary science data results to the world-wide science community.*
2000: 4 TBs; 2014: 720 TBs


MAVEN (NASA)


Osiris-REx (NASA)


LADEE (NASA)


ExoMars (ESA)

* Endorsed by the **International Planetary Data Alliance** in July 2012 – https://planetarydata.org/documents/steering-committee/ipda-endorsements-recommendations-and-actions

Some Features

- **Data Reuse**
  - *Designed a few simple formats for 80% of the data*
  - *All things are formally defined once*
  - *Everything that is registered as a product*
  - *Multi-level governance*

- **Model Driven**
  - *Model evolves with changes in the science discipline*
  - *Implementation technologies evolve at their own speed.*
  - *Improves interoperability at the information level*

- **Subsumes legacy archive**
  - *Proxy labels exist for each legacy product*
  - *High value data sets are migrated as needed*

Recommendations

- **Invest in capturing and maintaining data in well-annotated, accessible, structured data repositories**
  - ***Based on rigorous data/information architectures***
- **Computer Scientists, Statisticians/Data Scientists, Domain Experts (Scientists) must systematize the analysis of massive data**
  - *Significant efficiencies may be achieved by thinking of data analysis and data access together rather than thinking of them as serial operations.*
  - *We need new statistical methods and algorithms optimized for this type of environment.*
- **Develop computing infrastructures for sharing and analyzing highly distributed, heterogeneous data**
  - *This requires coordination (international, cross-agency)*
  - *It requires a software architecture*
- **Sustainability in both the data and the software infrastructures are critical**
  - *Although they can be on different evolutionary paths*

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
Stephanie McLaughlin
Thomas Morgan
Lynn Neakrase
Paul Ramirez
Anne Raugh
Mark Rose
Elizabeth Rye
Boris Semenov
Dick Simpson
Susie Slavney

David Heather
Santa Martinez

Peter Allan
Michel Gangloff
Thomas Roatsch
Alain Sarkissian

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

Thank You

Questions and Answers

Backup