Planetary Data System

Architectural Properties for Data Reusability

S. Hughes, D. Giaretta,
R. Downs, R, Joyner, J. Garrett

May 3, 2023 – 08:30
503/1-001 - Council Chamber

PV 2023 Conference
May 2 – 4, 2023
CERN
Geneva, Switzerland

Jet Propulsion Laboratory
California Institute of Technology

Introduction

- A fundamental requirement for long-term digital repositories is to ensure the reusability of data that have been entrusted to the repositories.
- Of the FAIR principles (Findable, Accessible, Interoperable, Reusable), Reusable is proving to be the most challenging.
- Two mature ISO level standards[1,2] exist that provide guidance for the long-term preservation of digital data.
  - *Several important architectural properties have been identified that help enable data reusability.*

[1] Open Archival Information System (OAIS) Reference Model (ISO 14721)
[2] Metadata Registry Specification (ISO/IEC 11179)

Shared Understanding

- Reusability depends on the existence of a shared understanding of the data.

- In 2001, Uschold [1] argued that a “single shared ontology” is critical for developing a digital library that enables semantic interoperability across disciplines.

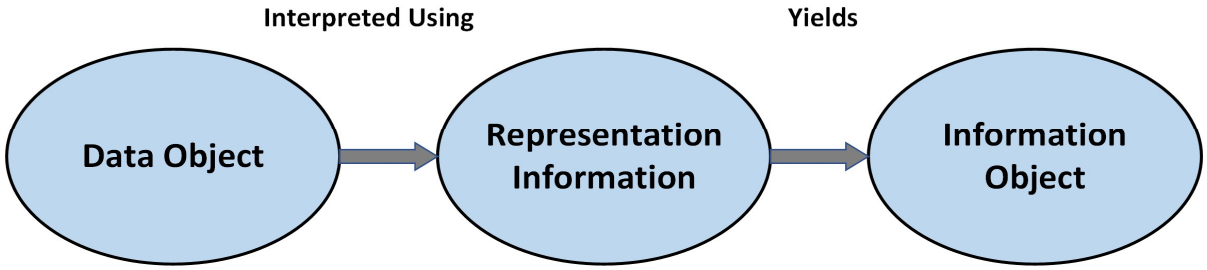
  – *A single shared ontology by definition promotes a shared understanding.*

[1] M. Uschold and Gruninger. M., "Ontologies and Semantics for Seamless Connectivity," SIGMOD Record, vol. 33, 2004.

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

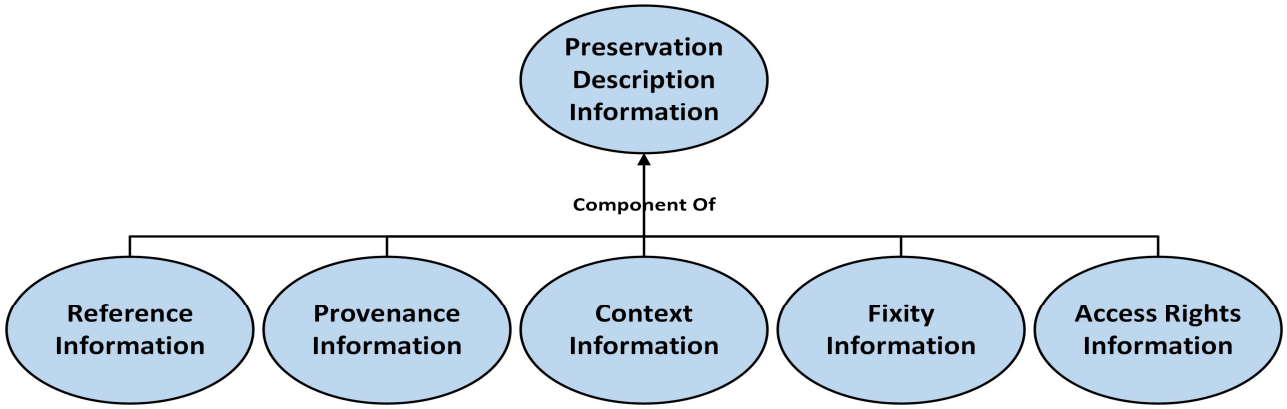
The Information Object[1]



- Information Object = Data Object + Representation Information
  - *Data Object = Physical Object or a Digital Object*
  - *Representation Information is the information that maps a Data Object into more meaningful concepts so that the Data Object may be understood.*
  - *It is a fundamental building block in the development of a common understanding.*

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

Preservation Description Information PDI

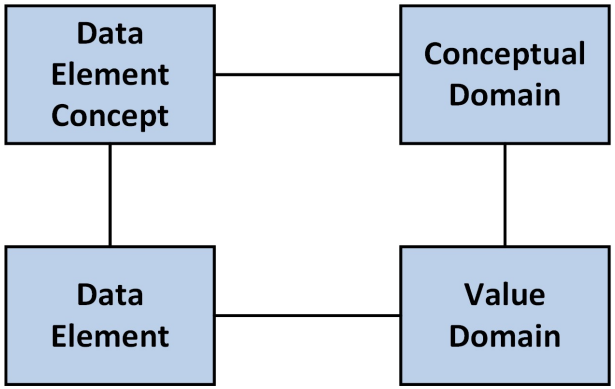


- Reference Information is necessary for referencing this data as well as referencing data that is in a meaningful relationship with this data.
- Provenance Information provides the history of the data and is essential for authenticity.
- Context Information is the information that helps orient the data within an environment.
- Fixity Information is required to ensure that data in general has not been unintentionally altered
- Access Rights Information identifies the access restrictions pertaining to the data, including the legal framework, licensing terms, and access control.

PDI (cont)

- The information, which along with Representation Information which is necessary for adequate preservation of the Data Object
- Each category of PDI is itself an Information Object
  - *This ensures that each has its own Representation Information to ensure that that it can be interpreted.*
  - *For example, each instance of Provenance Information has its own Representation Information so that the consumer can understand it.*
- PDI is an information object.
- Context Information in particular has an important role in enabling data reusability
  - *It can be used to define the relationships of the Data Object to the things within its environment*
  - *Adds semantic information.*

Metadata Registry[1]



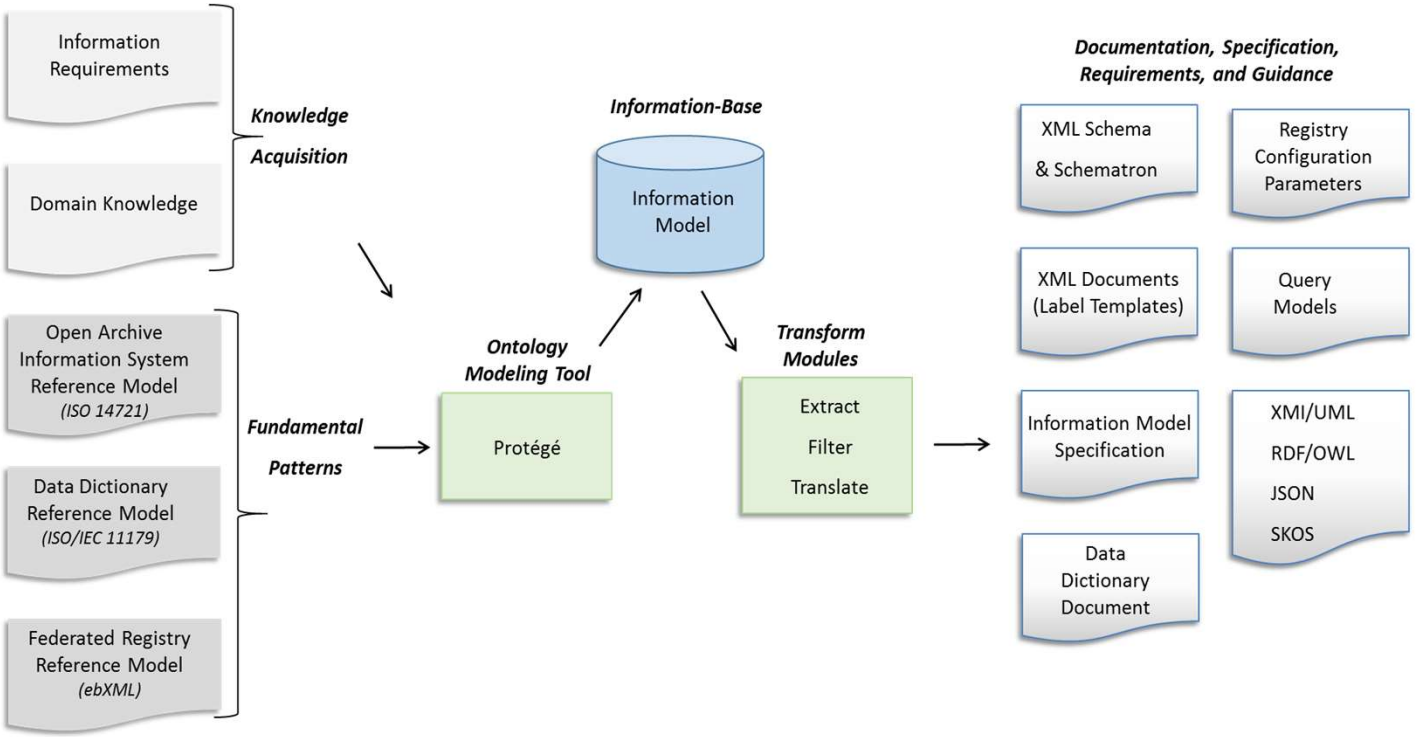
- A data dictionary schema for "data element" information
  - *Example the "start_time" of an event.*
  - *alternate names, definition sources, definitions in other languages, effective dates, submitting organization, and stewardship.*
  - *data representation, units of measurement, effective dates, submitter, and steward.*
  - *permissible values and value meanings*

[1] Metadata Registry Specification (ISO/IEC 11179)

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

PDS4 Information Model



- The PDS4 Information Model is implemented and maintained in the Protégé ontology modeling tool.
- Two separate instances
  - *Entity Model – ISO 14721 - OAIS RM*
  - *Data Dictionary Model – ISO/IEC 11179 – Metadata Registry*

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Trustworthy Digital Repository (TDR)

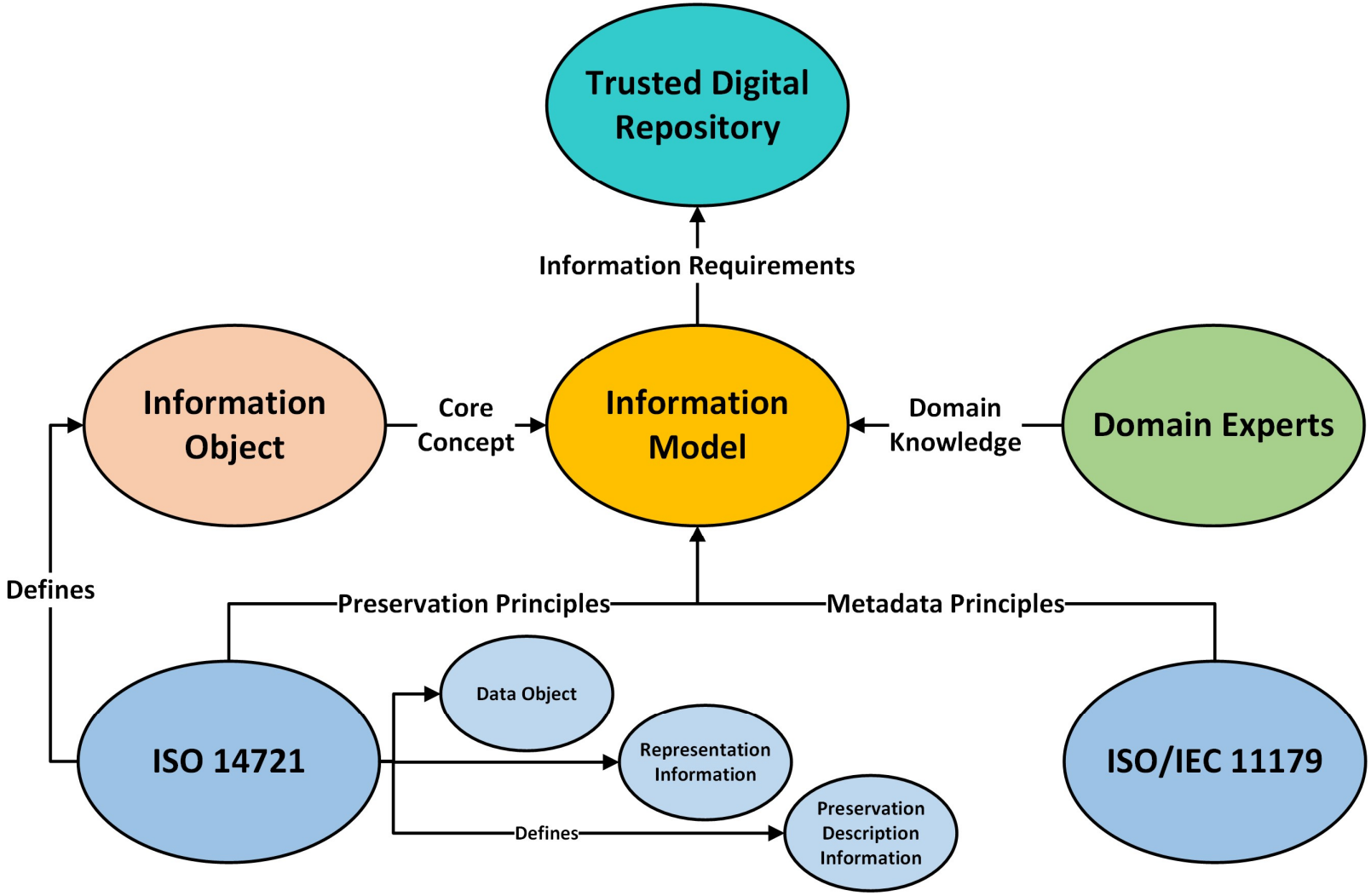
- A Trustworthy Digital Repository (TDR) is an organization or system responsible for the long-term preservation and access to digital materials, such as data, records, documents, and other digital objects.
  - *A TDR must be able to guarantee the authenticity, integrity, and usability of its digital content over time.*

- The aim of a TDR is to ensure that digital content remains accessible, usable, and reliable for as long as it is needed.
  - *This is particularly important for materials that have long-term value, such as cultural heritage objects, scientific research data, and government records.*

National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Architectural Properties as a Graph



National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Conclusion

- A framework for capturing the information required to support data reusability has been developed using principles adopted from two ISO information systems standards.
- Key properties have been identified that support reusability.
- However this is simply a framework.
  - *The really hard work involves acquiring the appropriate knowledge from domain experts and populating an domain information model using the framework as a guide.*

Acknowledgements

- The authors would like to thank the PDS Data Design Working Group (DDWG) and other PDS Discipline Node staff for their individual roles in the development of the PDS4 Information Model.

Questions / Answers

National Aeronautics and Space Administration

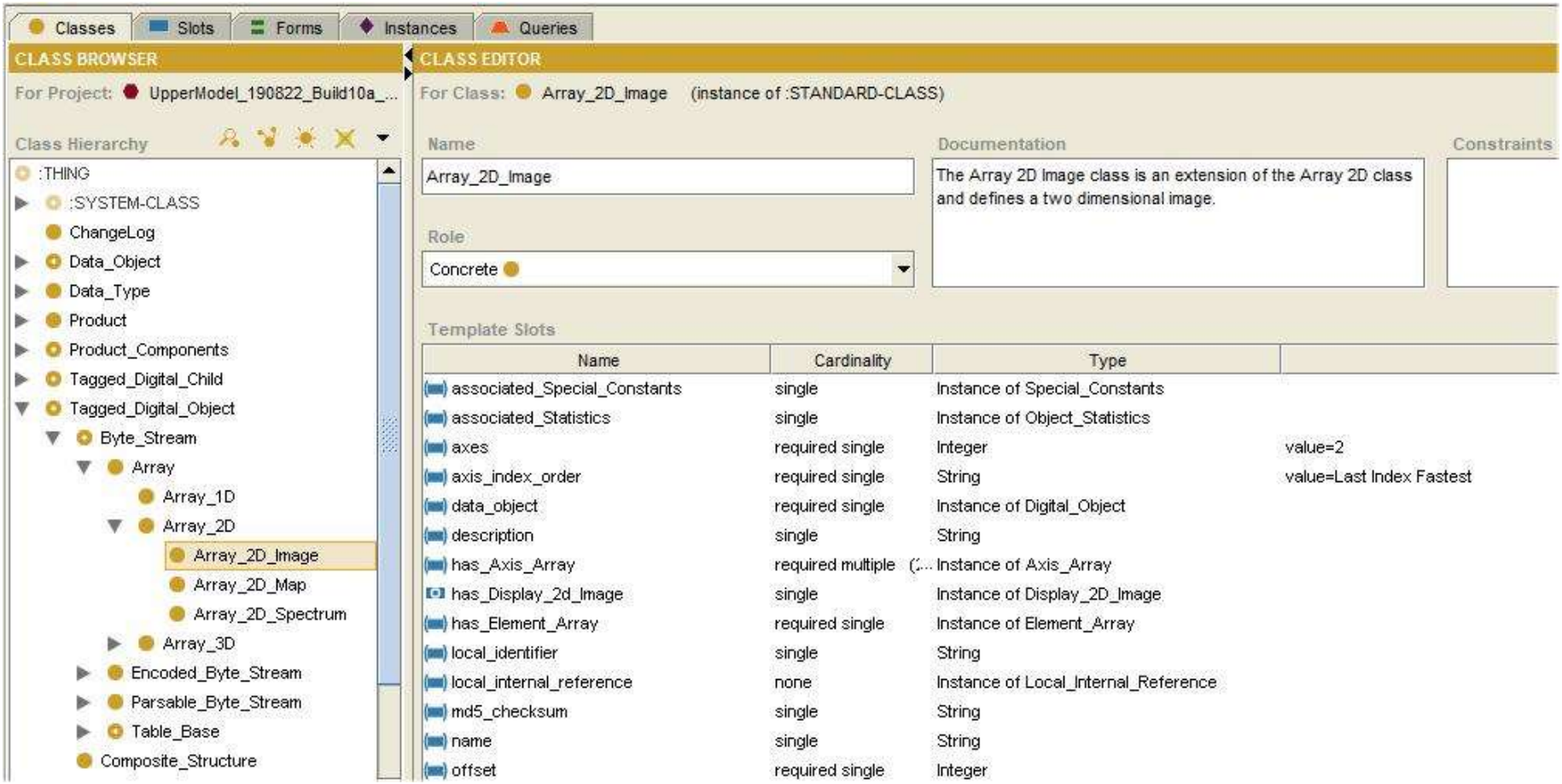
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Backup

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

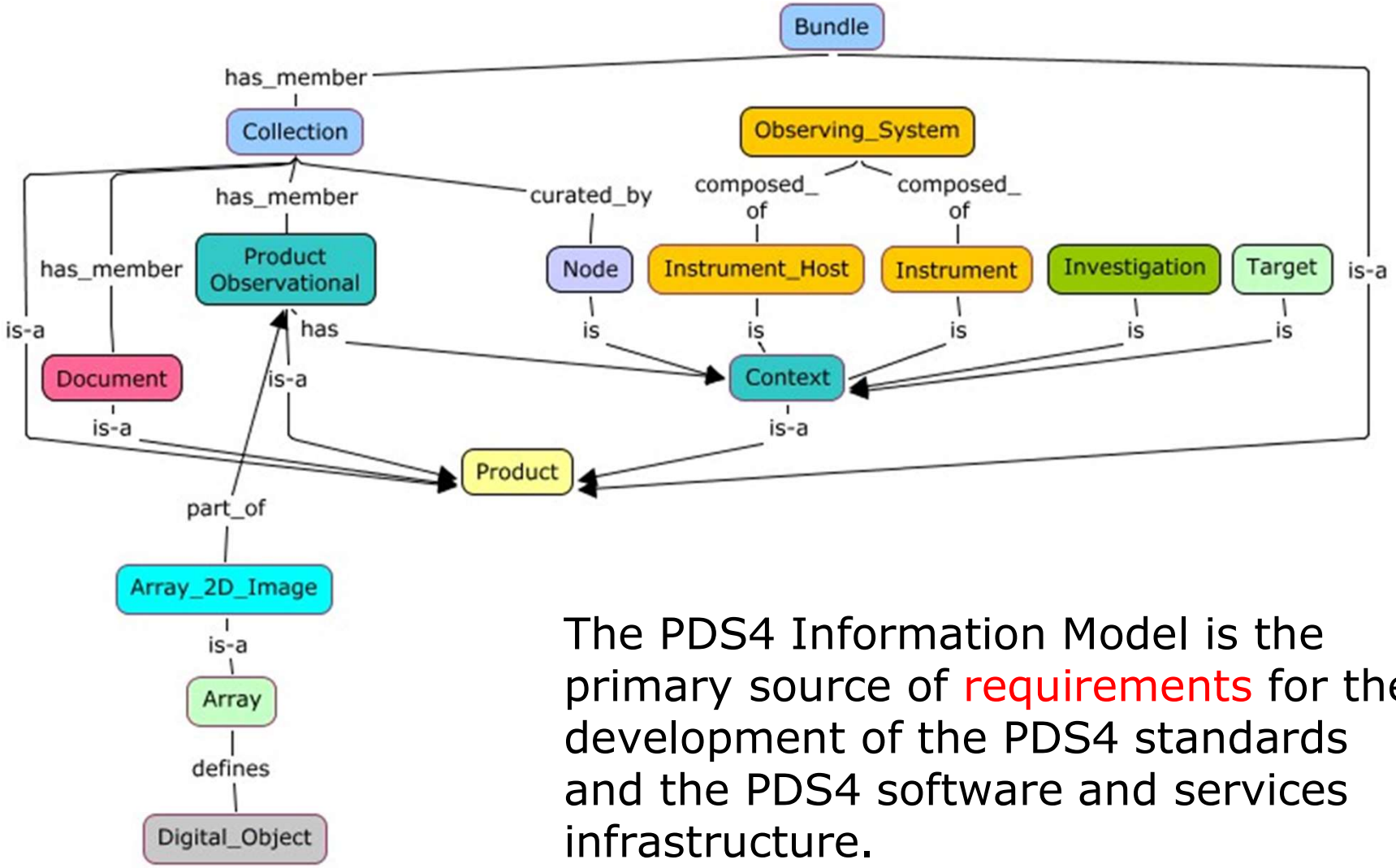
Ontology



National Aeronautics and Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

General Concept Map



The PDS4 Information Model is the primary source of requirements for the development of the PDS4 standards and the PDS4 software and services infrastructure.