An Ontology-Based Archive Information Model for the Planetary Science Community

J. Steven Hughes[1], Daniel J. Crichton[1], and Chris Mattmann[1]
Jet Propulsion Laboratory, Pasadena, California, 91016

The Planetary Data System (PDS) information model is a mature but complex model that has been used to capture over 30 years of planetary science data for the PDS archive. As the de-facto information model for the planetary science data archive, it is being adopted by the International Planetary Data Alliance (IPDA) as their archive data standard. However, after seventeen years of evolutionary change the model needs refinement. First a formal specification is needed to explicitly capture the model in a commonly accepted data engineering notation. Second, the core and essential elements of the model need to be identified to help simplify the overall archive process. A team of PDS technical staff members have captured the PDS information model in an ontology modeling tool. Using the resulting knowledge-base, work continues to identify the core elements, identify problems and issues, and then test proposed modifications to the model. The final deliverables of this work will include specifications for the next generation PDS information model and the initial set of IPDA archive data standards. Having the information model captured in an ontology modeling tool also makes the model suitable for use by Semantic Web applications.

I. Introduction

In 2007 an ontology-based information model was compiled for the International Planetary Data Alliance (IPDA) from core components of the NASA's Planetary Data System (PDS) archive data standards. The IPDA is a joint effort by national space exploration agencies, research institutions, and universities to enable global access and exchange of high quality planetary science data, and to establish archive standards that make it easier to share the data across international boundaries. The charter of the International Planetary Data Alliance states, "The data standards within the IPDA, including the data models and derived dictionaries, are based on the NASA Planetary Data System (PDS) standard that is the de-facto standard for all planetary data at the time of the IPDA founding".

The PDS has been serving the NASA planetary science community for almost 20 years and has a mature set of archive data standards that are well suited for archiving. The PDS has been a pioneer as a science data archive and has significantly influenced the archiving efforts of the larger space science community. However the current archive data standards are documented in over 800 pages of requirements, recommendations, and narrative that are often interpreted in diverse ways. One of the first requirements of the IPDA information modeling task was to write a formal specification of the PDS archive data standards and identify the core components that could guide data archive and information system development tasks in the IPDA community[1,6].

The information modeling project team chose an ontology modeling tool[9,10] to capture the PDS archive data standards. These tools are designed to help convert unstructured knowledge into a formal specification that describes the "things" in the domain. Domain knowledge is captured in these tools in a language or notation that is expressive enough to capture the majority of the information that can be depicted in a Unified Modeling Language (UML) class diagram. This was especially important for the PDS information model since historically it has been dependent on the Object Description Language (ODL), a weak object-based language develop by the PDS in the 1990's. After capturing the information model it can then be exported into the appropriate languages for implementation and documentation purposes.

In the following sections we describe the current work to identify the core information model elements and document problems and issues. We will also describe the framework of tools that are used to generate the

[1] Computer Scientist, Division 38 - Instruments and Science Data Systems, 4800 Oak Grove Dr., MS 169-414.

deliverables for this project, including formal specification documents for the PDS and the IPDA. This work is also now considered the starting point for the next version of the PDS archive data standards.

II. The Information Model Ontology

The Information Model as documented in the PDS archive data standards documentation consists of three related data models and the data dictionary. These three data models are the upper level or catalog model, the data format model, and the data product model. The catalog model is used to describe the things such as Instrument and Mission that provide a context within which data products are collected. The data format model is used to describe the logical and physical structure of the digital data to be archived. Finally the data product model describes the package that contains the data formats, file descriptions, identification information, and miscellaneous descriptive elements that comprise the data product. The document sources used to compile the Information Model includes the Planetary Science Data Dictionary (PSDD)[2], the PDS Standards Reference[3], and the PDS Catalog Design Document[4].

The development of a concise and consistent ontology required the identification of core elements that were common to each of the data models. Since the primary uses of the PDS data models are to describe things to be archived, the key element has been identified as the description class. The identification of the description class provides a means to unify the three models under the concept of an Open Archival Information System (OAIS) Reference Model[5] Information Object.

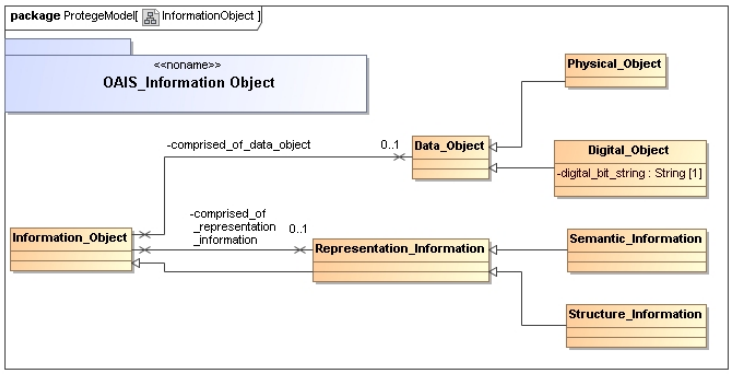

Figure 1. OAIS Information Object. *The OAIS Information Object consists of a data object and representation information. Representation information is also a subclass of information object allowing the creation of a recursive representation network.*

The OAIS Reference Model defines an information object as an aggregation of a "data object" and "representation information". The data object is the thing to be described and may be a digital sequence of bits, a physical thing, or a conceptual thing. The representation information is the description of the data object and may include both syntactic and semantic information. Fig. 1 shows a UML class diagram for an OAIS Information Object. As further specified in the figure, representation information is a subclass of an information object and suggests the development of a representation network where each instance of representation information has its own representation information. This recursive structure typically ends with some natural language narrative in the role of representation information.

A PDS data product consists of at least one aggregate consisting of an instance of a Data Format class and the actual data being described. This aggregate is easily mapped to the OAIS information object. First, the data, a sequence of digital bits maps directly to the OAIS Digital Object. It has been modeled as the Data Object Class in the ontology. As a generic class with only one attribute with data type "bit_string", it collects all the individuals, or "sequences of bits", in the archive. Another way to say this is that the Data Object Class was used to instantiate all data instances or "sequences of digital bits" in the archive.

A PDS Data Format description class map to the OAIS Representation Information. For example, the PDS Image is used to describe an image as a two-dimensional array of lines and line-samples. It has been modeled as the Image Description class in the ontology.

The OAIS Reference Model does not specify how the representation information is to be linked to the data object. In the PDS, a pointer, consisting of a file name and an offset is included in the description to link it to the data. A pointer class was defined in the ontology for this purpose.

The aggregate of a description class, the pointer class, and the data class is modeled in the ontology as the tagged data class. For example, the image description class, pointer class, and the data class is an aggregate called a tagged image. This concept is illustrated in Fig. 2.

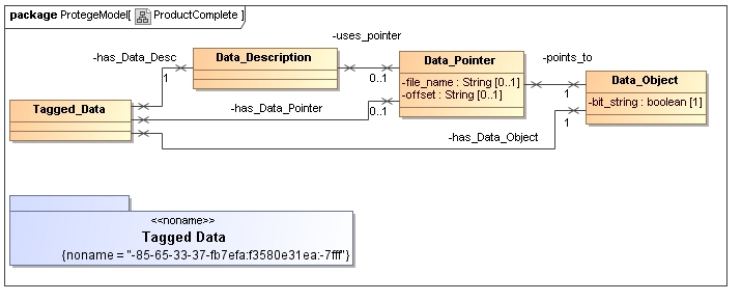

Figure 2. Tagged Data. *The Tagged Data Class consists of a data object, data pointer, and data description and map to the OAIS Information Object.*

The tagged data class, consisting of a triple, the data description, data pointer, and data object classes, represents a core essential element of the PDS Information Model. Since the data object class conceptually instantiates all data instances in the archive, the tagged data class conceptually instantiates all described data instances in the archive. Also since the tagged data class is the key component of the data product class it conceptually instantiates all products in the archive.

The OAIS Information Model can be specialized in a simple way to unify most of the components across the PDS data models. This specialization consists of generating the eight possible permutations formed by considering whether or not each element of the triple exists. For example, the tagged image class described above represents the case where the description, pointer and data all exist. The tagged data class is the parent class for essentially all data description classes.

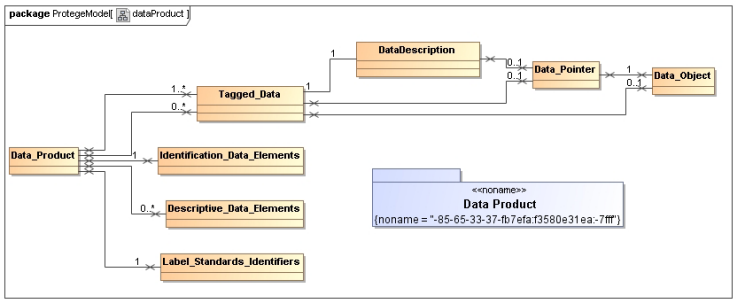

Figure 3. Data Product. *The Data Product Class aggregates the Tagged Data Class with classes for identification and ancillary information.*

A catalog class however, such as Instrument, represents the case where a description exists but not the pointer or the data. This is consistent with the PDS model when one considers that when a physical or conceptual object is being described, the data instance does not actually exist in the archive. This specialization of the OAIS Information Model is the parent class for essentially all catalog description classes. There are a couple of other permutations used for items in the archive such as where a pointer and data exist but without any description. However these are considered exceptions.

Finally, the remaining classes needed to complete the information model are modeled. For example, data products are modeled as aggregates that comprise tagged data classes as well as other miscellaneous elements. The partial UML Class Diagram for the PDS data product is shown in Fig. 3.

III. Presenting the Information Model to End Users

The information model is captured in an ontology modeling tool. This tool provided the means to capture information in a generalized notation that can be transformed as needed to other notations and languages for documentation and implementation purposes. The ontology model tool also allows the capture of information that might not have a place in traditional modeling tools. This is accomplished by simply creating a model for the additional information and then capturing it. For example, the PDS data dictionary includes data dictionary attributes not defined as typical ontology slot attributes. For example the PDS value formation rules have no direct mapping to an ontology slot attribute. A model for the PDS data dictionary was created in the ontology modeling tool and the PDS data dictionary content was ingested. This approach made all data dictionary attributes as well as the data dictionary content available for use in the framework.

The outstanding problem that remains after capturing the information model is presenting the information model in a form suitable for one or more audiences. Since the current task was to produce a data engineering specification, the object-oriented paradigm was chosen to present the information model. Mapping ontology concepts to object-oriented concepts is relatively straight forward. Ontology classes and their slot map to object-oriented classes and attributes. The behavior aspect of the object-oriented paradigm is ignored. However, where a relationship between classes is modeled as a slot of type "instance" in an ontology, a relationship in the object-oriented paradigm is modeled as a separate construct. The ontology class browser view of the tagged data class is shown in Fig. 4. As can be seen the Tagged Image is comprised of a Data Object, Data Pointer, and the Image description.

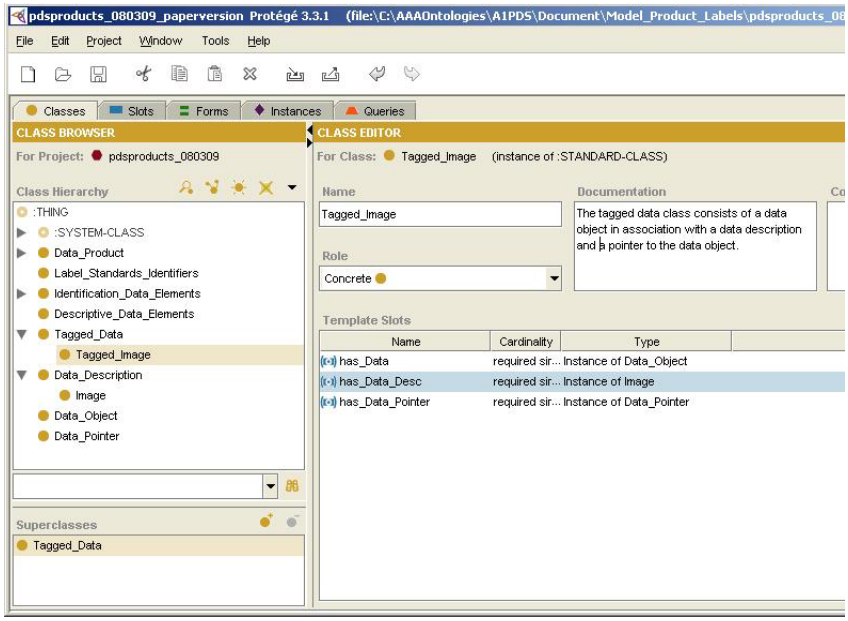

Figure 4. Tagged Data Class . *The Tagged Data Class as shown in the ontology modeling tool.*

Two notations were chosen to depict the information model in the specification. Unified Modeling Language (UML) class diagrams were chosen for the ontology classes and their relationships. The process for creating the UML class diagrams consisted of exporting the ontology modeling tool content to an XML Metadata Interchange (XMI) file. This file is then imported into a UML modeling tool where a visualizing wizard was used to create and refine the diagrams. Since a class diagram depiction of the complete information model would be very large and complex, it is partitioned into logical grouped. Examples of these diagrams can be seen in Figs. 2, and 3. The second notation was developed and is a table format that simply lists the class description, its attributes, relations, hierarchy, and other characteristics. An example of the table format for the Tagged Data Object is shown in Fig. 5.

The information model specification document consists of several sections consisting of logically grouped classes. The UML class diagrams, class tables, a class hierarchy and a short narrative make up each section. A dictionary is included at the end of the document that includes all class attributes and relationships together with their definitions.

Finally, a LaTEX and HTML versions of the information model specification document are generated simultaneously. The LaTEX version in turn allows the generation of a PDF version of the specification. Since the UML modeling tool generates Encapsulated PostScript (eps) versions of the UML Class diagrams, these are included to allow zooming into the diagrams. The HTML version of the document includes the diagrams in jpeg format but has links between all named elements in the document.

9.10 Tagged_Data_Object

Root Class: Tagged_Data_Object
Class Description: A tagged data object consists of a data object in association with a data object description and optionally a pointer to the data object.

	Entity	Card	Value
Hierarchy	Tagged_Data_Object		
Subclass	Tagged_Image		
Attribute	none		
Inherited Attribute	none		
Association	has_Data_Object has_Data_Object_Description has_Data_Object_Pointer	0..1 1 0..1	Data_Object File Data_Object_Pointer
Inherited Association	none		
Referenced from	Data_Product		

Figure 5. Tagged Data Class in Table Format. *The Tagged Data Class in table format shows all attributes, relationships, and references.*

IV. Current Work

The information Model Specifications generated for both the IPDA and PDS are currently under assessment. Working groups have been formed to validate the initial requirements and that the information model meets those requirements. In particular the requirement that the specifications be unambiguous is being validated. In addition, the PDS working group is validating that the specification represents the PDS archive data standards as well as possible. All clear requirements are must be captured. In addition where the PDS standards are ambiguous, the most common interpretations are being modeled. The validation of the specification is also identifying problems, issues and gaps in requirements that should be addressed for the next version of the archive data standards. Since these issues have been identified and located within the specification it is expected that fixes can be easily made. In addition, since the ontology modeling tool allows the ingestion of data, the model and any changes can be tested. In addition the export of the tool contents to RDF/XML allows the configuration of web-based semantic browsers[6,7] for navigating through both the data and metadata.

V. Conclusion

The generation of a formal specification for the PDS information model was prompted by a request from the International Planetary Data Alliance (IPDA) to adopt the PDS archive data standards. The mature but complex model was captured in an ontology model tool to formalize the relatively unstructured archive data standards. The identification and mapping of key components of the model to the OAIS Information Object helps formalize and unify the model's components. It also helps confirm the model's role as an international standard.

The resulting knowledge based is used to generate specification documents using commonly accepted modeling notations. The availability of this formal data engineering specification for distribution to data producers and application developers world-wide will significantly minimize dependence on natural language to understand the model. The development of archive data products and archive information systems compliant to this archive data standard will result in truly interoperable planetary science data archives world-wide.

Acknowledgments

The authors would like to acknowledge the many members of the Planetary Data System (PDS) Technical Staff and planetary science domain experts for developing and maintaining the PDS data model. This work only formalizes and makes more accessible the results of years of effort and countless hours of discussion. This work was performed under contract to NASA by the Jet Propulsion Laboratory, California Institute of Technology.

References

[1] Hughes, J.S., Crichton, D.J., Beebe, R., Preliminary Definition of the Core Archive Data Standards of the International Planetary Data Alliance (IPDA), PV 2007 International Conference, http://pv2007.dlr.de/Papers/Hughes_CoreDataStd.pdf, October, 2007.

[2] Planetary Science Data Dictionary, JPL D-7116, Rev. D, Jet Propulsion Laboratory, California Institute of Technology.

[3] Planetary Data System Standards Reference, Version 3.7, JPL D-7669, Part 2, Jet Propulsion Laboratory, California Institute of Technology, March 20, 2006.

[4] Planetary Data System Catalog Design Document, Version 2.0, JPL D-1152, Jet Propulsion Laboratory, California Institute of Technology, 1990.

[5] Reference Model for an Open Archival Information System (OAIS), Consultative Committee for Space Data Systems, CCSDS 650.0-B-1, Blue Book, January 2002.

[6] J. Hughes, D. Crichton, C. Mattmann and P. Ramirez, "Data Model Management for Space Information Management Systems", Jet Propulsion Laboratory, Pasadena, CA, AIAA-2006-5720, SpaceOps 2006 Conference, Rome, Italy, June 19-23, 2006.

[7] Hughes, J. S., Crichton, D. J., Kelly, S., Mattmann, C., The Semantic Planetary Data System, Ensuring Long-term Preservation and Adding Value to Scientific and Technical data, The Royal Society, Edinburgh, November 2005.

[8] H.Wache, T. Vogele, U. Visser, H. Stuckenschmidt,G. Schuster, H. Neumann and S. Hubner , Ontology-Based Integration of Information — A Survey of Existing Approaches, Proceedings of IJCAI-01 Workshop:Ontologies and Information Sharing, Seattle, WA, 2001, Vol. pp. 108-117.

[9] Models as a Basis for Ontologies, Ed Barkmeyer, NIST, Ontolog Forum, April, 2007.