

Ontology-Based Information Model Development for Science Information Reuse and Integration

J. Steven Hughes¹, Daniel J. Crichton¹, Chris A. Mattmann^{1,2}

¹*Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109, USA
{jshughes,crichton,mattmann}@jpl.nasa.gov*

²*Computer Science Department
University of Southern California
Los Angeles, CA 90089, USA
mattmann@usc.edu*

Abstract

Scientific digital libraries serve complex and evolving research communities. Justifications for the development of scientific digital libraries include the desire to preserve science data and the promises of information interconnectedness, correlative science, and system interoperability. Shared ontologies are fundamental to fulfilling these promises. We present a tool framework, a set of principles, and a real world case study where shared ontologies are used to develop and manage science information models and subsequently guide the implementation of scientific digital libraries. The tool framework, based on an ontology modeling tool, has been used to formalize legacy information models as well as design new models. Within this framework, the information model remains relevant within changing domains and thereby promotes the interoperability, interconnectedness, and correlation desired by scientists.

Keywords: Digital Library, Ontology, Information Model, Interoperability, Science Data, Science Metadata

1. Introduction

Scientific Digital Libraries are the key to advancing science through scientific collaboration. The advent of the Web and languages such as XML have brought an explosion of online science data repositories and the promises of correlated data and interoperable systems. However there have been relatively few real successes since research [1] suggests that just having physical and syntactic connectivity is not adequate. To achieve seamless connectivity between repositories, not only must the semantic issues be addressed, but important assumptions must be made about the ontologies being used to address the semantic issues. These assumptions include the need for a “single shared ontology” and the need for human assistance in the development of the

ontology. Without these assumptions the effort to achieve seamless connectively across pre-existing repositories is essentially “cryptography”, and rapidly becomes intractable. Our experiences support Uschold’s [1] assumptions about the need for a single shared ontology and the need for human assistance to address the semantic issues required for interoperability to occur. This paper will present a tool framework and a set of principles that have been used to develop shared ontologies for several scientific digital library projects.

2. Background

The problem of bringing together heterogeneous and distributed information systems is known as the “interoperability problem” [2]. As recent research suggests physical and syntactic interoperability without semantic interoperability does not solve the general interoperability problem. To address semantic interoperability, Uschold [1, 3] suggests the use of shared ontologies with the following assumptions.

1. All parties should use a single language for representing their ontologies.
2. All members in a given community should use:
 - a. a single shared ontology, or
 - b. a single shared upper ontology, with distinct domain ontologies, or
 - c. a shared interlingua ontology to map individual ontologies to and from.
3. The semantic mapping among ontologies should be human-assisted, rather than fully automated.
4. The mapping will be done between lightweight ontologies, with a limited role for automated reasoning.
5. Adequate infrastructure support will exist for community repositories of both ontologies and interontology mappings.

Uschold and Gruninger [3] also suggest the following phases for the ontological engineering process.

1. Identify the purpose and scope including specialization, intended use, scenarios, set of terms including characteristics and granularity.
2. Build the ontology.
3. Evaluation: Verification and Validation.

In the following section we present a case study where a shared ontology was developed under these assumptions.

3. The Planetary Data System

The Planetary Data System (PDS) was developed to archive and distribute scientific data from NASA planetary missions, astronomical observations, and laboratory measurements. The PDS data standards [4] were developed in the late 1980's to define the concepts and terms needed for archiving science data in the planetary science domain. Even though the data standards were innovative [4-7] for their time, ambiguity and many assumptions have crept in over almost two decades of use and have caused significant problems for PDS operations, data providers, and end-users.

In 2008 the PDS formed a team to review the data standards and create an ontology [6]. The team configured a tool framework, based on an ontology modeling tool, to manage the ontology and produce specifications for developers and documentation for end-users.

The scope of the PDS data standards is one of the broadest in the space sciences and covers several planetary science sub-domains. They define and describe the data structures, data formats, and contextual information needed to make the science data useful to current and future planetary scientists. Each sub-domain has their own domain of discourse but simultaneously desires collaboration with the other sub-domains. The sub-domains also share data types ranging from images to binary tables.

Uschold's assumption that all members in a given community should use a single shared upper ontology, with distinct domain ontologies fits the planetary science domain. For example the PDS upper ontology defines mission, instrument, and target, things that are common across the sub-domains. The sub-domains, namely Imaging, Atmospheres, Geosciences, Rings, Planetary Plasma Interactions, Small Bodies, and Radio Science subsequently define their own ontologies. This hierarchy continues to the mission level where things specific to a

mission, such as operational concepts are defined. Interoperability between PDS disciplines is subsequently realized using the upper ontology. The sharing of discipline level ontologies, especially imaging for example, also supports interoperability between the sub-domains that use imaging data. Uschold's other assumptions are also supported in that the PDS uses a single common language, human assistance is the preferred method used to determine the semantic mappings, automated reasoning is used infrequently and primarily only for testing the ontology, and the ontologies are configured as PDS resources.

To validate the ontology the system's functional requirements were referenced to identify the "things" that the implemented services act on to perform their functions. These are either explicitly mentioned as nouns or simply implied in the requirements. The resulting list is considered to be the "information modeling" response to the system's functional requirements and validates that the ontology contains the classes needed to support system services.

4. Information Modeling

As an information system, a scientific digital library requires an information model. An ontology is very useful in the development of an information model since it focuses on defining "what is", as in, what is an asteroid and how is it different from a comet? Once the domain is defined in an ontology, the classes and relationships typically associated with an information model can then be added, for example an asteroid's associations to the asteroid's images and related science publications.

Information modeling also focuses on metadata. In fact, many things in the domain, for example the asteroid, are only represented by metadata. In contrast, a data model typically focuses on describing a digital structure.

In a scientific digital library, especially those with requirements for long-term usability and persistence, the metadata is on equal footing in significance to the data. For example, a digital image is essentially useless to a planetary scientist unless information about the locations of the light source, the imaging instrument, and the target body are all known within a single frame of reference. Also since imaging observations during space flight are often non-repeatable, it is in the best interest of science to collect as much information as possible about the observation and the context within which it was performed. There are also requirements for complex classification schemes to enable searching within large volumes of data and to support correlative science.

An information model will typically comprise several data models for the actual data as well as models for physical and conceptual things in the domain. As mentioned earlier, the context within which data is collected is as important as the data itself. For example in addition to a digital image model, a model will also be needed to describe the physical instrument and the mission that is managing the project. Therefore in general, information models for science domains must describe digital, physical, and conceptual things. These three general classes can be unified under the concept of the Open Archival Information System (OAIS) “Information Object” [8].

In general an information object is defined as comprising a data object and its descriptive or “representation information”. A “data object” can either be a digital object, a black box containing a sequence of bits, or a physical object that can be touched, for example a moon rock. The OAIS data object can be extended to add the conceptual object. The representation information contains the structural, semantic, and other information needed to understand and use the data object.

The need for structural information in a science digital library again supports Uschold’s distinct domain ontologies assumption, the domain in this case being data structure. In the PDS one of the most important ontologies developed was a data structure ontology. This ontology promotes interoperability across the sub-domains as well as a stable long-term archive. For example, the definition of an n-dimensional homogeneous array as a fundamental data structure promotes interoperability at the basic I/O level for data processing. The extension of this class to 2-D and 3-D arrays and then to Image_Grayscale and 3-D Spectrum respectively, promote interoperability during data analysis.

An information model is used throughout the development and operation of an information system and referenced by software developers to project managers. Few of these users understand or need the details of an information model. To be useful the information model must be filtered and presented in notations that are suitable to the target audience. The Zachman Framework [9], a classification structure that is used in information technology development, defines several viewpoints and

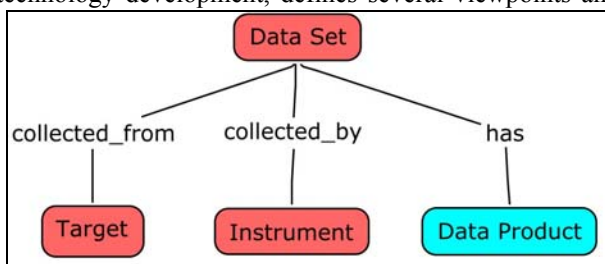


Figure 1. A portion of the PDS Conceptual Model depicted as a concept map.

associated models that address this problem.

A conceptual model defines the community model of data from a manager’s point of view and is concerned with the language of the community. Figure 1 illustrates a small portion of the PDS conceptual model as a concept map.

A logical model defines the system model from a designer’s point of view and is concerned with entity classes, attributes, and relationships. Figure 2 provides a portion of the logical model for two PDS classes, data set and instrument.

Other models in the Zachman hierarchy include the contextual model that provides a high level strategic view and other more detailed views associated with specific implementation choices. For example, the implementation of a model into a relational database system requires the logical model to be mapped to a relational physical model.

5. Principles

Wache [2] cites a “striking lack of sophisticated methodologies supporting the development and use of ontologies.” The following principles were used during development of the PDS ontology and point to guidelines for shared ontology development.

5.1. Model Independence

The model should remain independent of its implementation. During the development of the PDS ontology, it was assumed that the ontology would remain independent of the target languages into which it was to be expressed. This is important since implementation models often add constraints. For example, XML is currently a popular language for implementation. However, if used as a modeling language, the hierarchical nature of XML/Schema would skew the model of a domain that was not intrinsically hierarchical.

Once an ontology model is captured, the model can be filtered and exported to other often less expressive languages. The special treatments needed to “shoe horn” an ontology into a target language are then located within the local implementation and subsequently are not propagated to other implementations or versions of the ontology model. The PDS experience is that many requests for change are often overly impacted by the limitations and quirks of the implementation language. The change process is therefore easier managed by considering the model and its implementation separately.

5.2. Model Driven

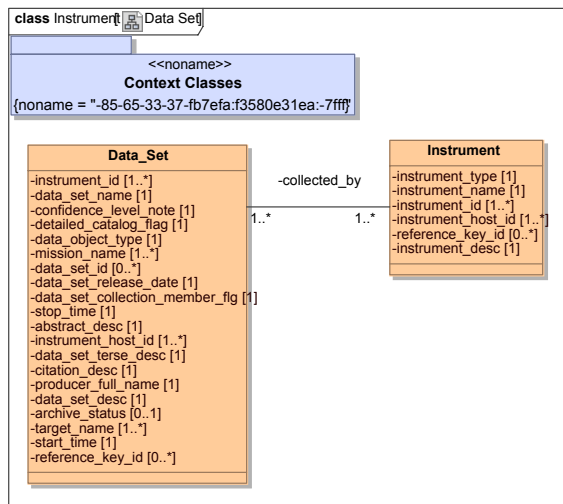


Figure 2. PDS Logical Model – UML Class Diagram

Developing scientific digital libraries for diverse and complex science domains such as the PDS poses two special challenges. First because the PDS supports a research community, the information model must keep pace with advancements in the science domain and periodic changes in geographical and political boundaries. Second, the technology used to implement the underlying information system will change at a different speed, typically faster, than the science domain.

As recommended earlier the ontology should remain independent of the implementation technology. A changing environment suggests that the ontology should guide both the development and long term management of the information system.

As Uschold suggests, adequate infrastructure support must exist for community repositories for the ontologies to ensure they are available to meet the needs of a model driven architecture.

5.3. Semantic Richness

The ontology modeling language should be semantically richer than the other languages in the framework. This is suggested by the model independence principle since the ontology contents will typically be filtered and exported to less expressive languages. Although not explicitly stated this would seem to be consistent with Uschold’s suggestion to use a single language for representing ontologies.

5.4. Data Object Class Unification

The dichotomy in a scientific digital library between descriptions of “actual” data and descriptions of physical and conceptual things in the domain can be unified under the OAIS Information Object. Since digital, physical, and conceptual objects all have representation information but not actual data, then interoperability is dependent on shared ontologies that define the representation information.

5.5. Ontology Development and Maintenance

Information modeling is a highly difficult problem that requires time and both domain and information modeling experts. For even the smallest domains the development of an ontology can take years to complete since getting consensus on what something is, especially in the science domains, can be extremely arduous. In addition the domain continues to evolve even after an ontology is completed and continual updates will be required throughout the life of the system. Our experience suggests that automated tools can help, but human assistance is ultimately required, especially to finalize the mapping between ontologies, as Uschold suggests.

6. Tool Framework

The tool framework is based on the Protégé ontology modeling tool [10]. Other tools include the CMapTools Knowledge Modeling Kit [11] which is used to generate visual depictions of conceptual models. The MagicDraw modeling tool [12] is used to generate UML models, class diagrams and code for defining and accessing Java classes. A Java application was written to generate specification documents that present several views of the information model using class definition tables, UML class diagrams, and concept maps. The specification also includes the data dictionary that describes the attributes in detail. The ontology content is both exported to XMI for use by the modeling tools and parsed by the Java application for the generation of the specification document. HTML and LaTeX versions of the specification document are generated. This framework and the process flow are illustrated in Figure 3.

Besides the PDS, the tool framework is being used to support information modeling tasks for several other projects including the International Planetary Data Alliance (IPDA), the Early Detection Research Network (EDRN) Knowledge Environment (EKE) [13], and the Consultative Committee for Space Data Systems (CCSDS) Registry Reference Model specification work.

The IPDA is a close association of international partners with the aim of improving the quality of planetary science data and services to the end users from space based instrumentation. In particular it seeks to improve interoperability between planetary science archives by developing common data and technology architectures. The IPDA adopted [14] the PDS information model as the de-facto data standard for the planetary science community and is now developing the technology architecture including a set of standard protocols.

The EDRN is a research network of collaborating scientists from over 40 institutions focused on identifying and validating cancer biomarkers (biological indicators of cancer) at their earliest stages. The EDRN Knowledge Environment (EKE) serves as an online, distributed resource of data and information that helps improve scientific research by enabling real-time access to cancer-research information that crosses institutional boundaries at a national level. The EDRN core ontology [15] defines this data and information.

The CCSDS is an organization of Space Agencies and produces recommendations and standards mainly for ground systems and their interface to space systems. The CCSDS Registry Reference Architecture includes an information model for a general purpose registry. This information model is being managed by the tool framework.

7. Data Dictionary

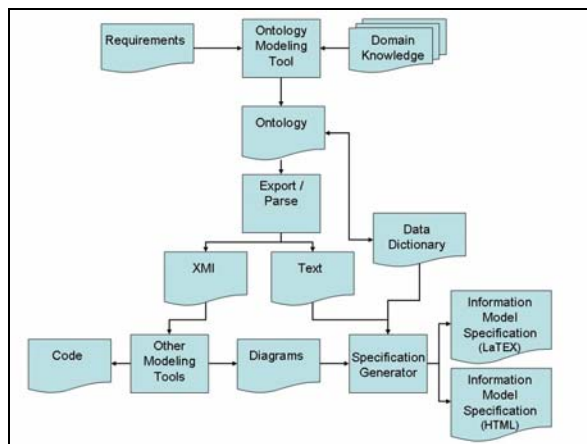


Figure 3. Tool Framework and Process Flow

Ontology modeling tools provide the means to capture a domain model in terms of classes and relationships. However a scientific digital library such as the PDS needs a rich set of data about class attributes and their values. For example the definition of an image pixel must include the data type of the value, the value's minimum and maximum bounds, whether the value is signed, and the order of the bytes in storage. In addition during the design of a new image, data engineers will want to know what pixel definitions have been previously used that are similar in concept, who defined them, and who is allowed to make changes. The PDS is adopting the ISO/IEC 11179 2003 [16] Metadata Registry Specification for its data dictionary and integrating it into its tool framework.

8. Related Work

Wache et. al. [2] summarize that reasonable results have been achieved on the technical side of using ontologies for intelligent information integration. The typical information integration system uses ontologies to explicate the contents of an information source, mainly by describing the intended meaning of table and datafield names. For this purpose, each information source is supplemented by an ontology which resembles and extends the structure of the information source. Noy [17] states that many issues that ontology researchers in semantic integration grapple with are very similar to the issues that database and information-integration researchers have been addressing. Some of the approaches are also similar although the ontology community relies more heavily on the higher expressive power of ontology languages and on reasoning techniques. Knublauch [18] also examine the benefit of using ontologies to support information modeling. Finally Singh et al. [19] suggest the need for ontologies to support the development of metadata catalogs for the sophisticated data-intensive applications resulting from advances in computational, storage and network technologies and data grid infrastructures.

9. Conclusion

The advent of the Web and languages such as XML has brought an explosion of online science data repositories and the promises of correlated data and interoperable systems across science domains. However there have been relatively few successes at providing useful scientific collaboration. Research suggests the need for information models that provide the missing semantic information. Experience suggests however that the development of an information model requires significant input from hard-to-get domain experts and years of effort for even the simplest science domain. To help address this

issue, the authors have developed a tool framework based on an ontology modeling tool for developing and managing information models. The information model is maintained independent of its implementation and allows the model to evolve with the domain. In a model driven architecture, changes to the model will also drive changes to the implementation.

Our experiences supports Uschold's suggestion and assumptions regarding the use of shared ontologies to enable interoperability. In particular a hierarchy of ontologies, one shared upper ontology and several sub-domain ontologies are required in a domain as complex as planetary science.

The use of the framework with the Planetary Data System as well as several other science information systems has proven the usefulness of the tool framework and provided a set of principles and some lessons-learned for improving information reuse and integration.

10. Acknowledgments

The authors wish to acknowledge the PDS Data Modeling for developing the original PDS data model and the PDS Technical staff who have performed heroically in attempting to keep the PDS data standards viable in the continually evolving planetary science domain. This work was supported by the Jet Propulsion Laboratory, managed by the California Institute of Technology under a contract with the National Aeronautics and Space Administration.

11. References

- [1] M. Uschold and G. M., "Ontologies and Semantics for Seamless Connectivity," *SIGMOD Record*, vol. 33, 2004.
- [2] H. Wache, et al., "Ontology-Based Integration of Information — A Survey of Existing Approaches," In Proc. IJCAI-01 Workshop: Ontologies and Information Sharing, 2001.
- [3] M. Uschold and M. Gruniger, "Ontologies: Principles, methods and applications," *Knowledge Engineering Review*, vol. 11, pp. 93-155, 1996.
- [4] J. S. Hughes and S. K. McMahon, "The Planetary Data System. A Case Study in the Development and Management of Meta-Data for a Scientific Digital Library.," In Proc. ECDL, 1998.
- [5] J. S. Hughes, et al., "A Planetary Data System for the 2006 Mars Reconnaissance Orbiter Era and Beyond," In Proc. 2nd ESA Symposium on Ensuring the Long Term Preservation and Adding Value to Scientific and Technical Data (PV-2004), Frascati, Italy, 2004.
- [6] J. S. Hughes, et al., "An Ontology-Based Archive Information Model for the Planetary Science Community," In Proc. Spaceops, Heidelberg, Germany, 2008.
- [7] S. Hughes, et al., "The Semantic Planetary Data System," In Proc. 3rd Symposium on Ensuring Long-term Preservation and Adding Value to Scientific and Technical Data, The Royal Society, Edinburgh, UK, 2005.
- [8] "Reference Model for an Open Archival Information System (OAIS)," *CCSDS 650.0-B-1*, 2002.
- [9] J. A. Zachman, "A framework for information systems architecture," *IBM Syst. J.*, vol. 26, pp. 276-292, 1987.
- [10] H. Eriksson and M. Musen, "Metatools for Knowledge Acquisition," *IEEE Softw.*, vol. 10, pp. 23-29, 1993.
- [11] A. Cañas, et al., "Managing, Mapping, and Manipulating Conceptual Knowledge," In Proc. AAAI-99 Workshop on Exploring Synergies of Knowledge Management and Case-Based Reasoning, 1999.
- [12] NoMagic, "Magic Draw, <http://www.magicdraw.com/>," 2009.
- [13] D. Crichton, et al., "A Distributed Information Services Architecture to Support Biomarker Discovery in Early Detection of Cancer," In Proc. 2nd IEEE International Conference on e-Science and Grid Computing, Amsterdam, the Netherlands, 2006.
- [14] J. S. Hughes, et al., "Preliminary Definition of the Core Archive Data Standards of the International Planetary Data Alliance (IPDA)," In Proc. PV 2007, 2007.
- [15] J. S. Hughes, et al., "An Information Model for Biomarker Research," In Proc. 5th EDRN Scientific Workshop, Bethesda, MD, 2008.
- [16] ISO/IEC, "ISO/IEC 11179: Information Technology -- Metadata registries (MDR), <http://metadata-standards.org/11179/>," 2008.
- [17] N. Noy, "Semantic Integration: A Survey of Ontology Based Approaches," *SIGMOD Record*, vol. 33, pp. 65-70, 2004.
- [18] H. Knublauch, "Ontology-Driven Software Development in the Context of the Semantic Web: An Example Scenario with Protege/OWL," in International Workshop on the Model-Driven Semantic Web. Monterey, CA, 2004.
- [19] G. Singh, et al., "A Metadata Catalog Service for Data Intensive Applications," in Proceedings of the 2003 ACM/IEEE conference on Supercomputing: IEEE Computer Society, 2003, pp. 33.