Architectural Properties for Data Reusability

John S. Hughes
Architecture Sytems Engineering
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA USA
John.S.Hughes@jpl.nasa.gov

David Giaretta
Giaretta Associates
Dorchester, United Kingdom
david@giaretta.org

Robert Downs
Center for International Earth Science Information Network (CIESIN)
Columbia University
Palisades, NY USA
rdowns@ciesin.columbia.edu

Ronald Joyner
Architecture Systems Engineering
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA USA
Ronald.Joyner@jpl.nasa.gov

John Garrett
Garrett Software
Columbia, MD USA
garrett@his.com

***Abstract*—A fundamental requirement for long-term digital repositories is to ensure the reusability of data that have been entrusted to the repositories. This is especially true for scientific studies involving repeated observations of the same variables during different time periods. The data must be preserved for future use and remain usable, even as science discipline practices change and implementation technologies evolve.**

The Planetary Data System (PDS) has over four decades of experience building a digital repository for diverse and evolving science domains. The PDS has been successfully providing its designated community with high-quality peer-reviewed digital data that are being reused regularly in scientific studies, cross-disciplinary research, and data science.

The PDS4 Information Model was designed using principles and standards developed for the long-term preservation of digital data. Several important architectural properties have been identified as important for promoting data reusability. These will be described and presented in a knowledge graph.

Keywords—ontology, architecture, property, data, reuse, digital, trust, repository

I. Introduction

Most efforts to implement the FAIR data principles concentrate on making data Findable and Accessible. However, the two principles of Interoperability and Reusability are often neglected and achieving Reusability is proving to be the most challenging. Ensuring that data is reusable is crucial, especially for cross-discipline studies that require data to be preserved over decades and that must survive multiple technological changes. The increasing trend towards cross-discipline research necessitates data reuse for purposes other than their original intent and the rise of Data Science and Analytic applications require data to be machine-readable and reusable in compute-intensive environments. Reusability is also crucial for maximizing the impact of the data in the community and ensuring the sustainability of data infrastructures.

II. Data Reusability

It is fair to say that the FAIR data principle of Reusability depends on the existence of a shared understanding about the data. At a minimum there must be shared understanding between the data producers and the data consumers so that the consumers can understand the data that they find and access. If there are two or more data producers producing data, then the producers should agree on common classifications for the data, standard data formats, effective search parameters, etc.

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. Typically, a knowledge representation framework is used to create the ontology. This framework consists of an ontology modeling tool and standard frameworks for classifying information. The results are a set of formal and sharable set of information requirements for an information system that promote data reusability.

In general the development process should include a) a well-defined scope and sufficient resources, b) the adoption of a systems architecture that allows development of the data architecture separate from the software/services architecture, c) the development and maintenance of an information model that remains independent of the implementation, and d) the use of "agile development" to provide early delivery of prototypes to users and evolutionary development and adaptive planning based on feedback.

The PDS4 Information Model [2, 3, 4, 5] is implemented and maintained in the Protégé ontology modeling tool. Two standard frameworks were used during the knowledge acquisition phase to classify and organize the information captured. Several key concepts supporting data reusability have been identified.

A. Information Object

The Open Archival Information System (OAIS) Reference Model (RM), ISO 14721:2012 [6] introduces the concept of the Information Object which is defined as a Data Object together with its Representation Information. A Data Object in turn is either a 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. Since the purpose of Representation Information is to ensure that the Data Object is understood, the OAIS Information Object becomes the fundamental building block in the development of a common understanding of the data.

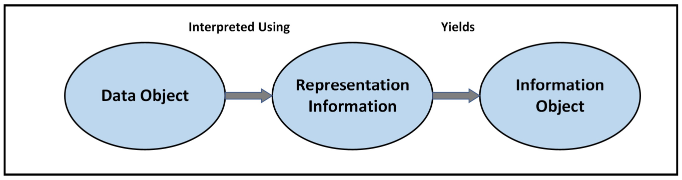

Figure 1 Information Object

B. Preservation Description Information

The OAIS RM also introduces five categories of metadata that promote a common understanding of the data and subsequently improves the reusability of the data.

1) Reference - Reference Information is necessary for referencing this data as well as referencing data that is in a meaningful relationship with this data.

2) Provenance - Provenance Information provides the history of the data and is essential for authenticity.

3) Context – Context Information is the information that helps orient the data within an environment.

4) Fixity – Fixity Information is required to ensure that data in general has not been unintentionally altered

5) Access Rights- Access Rights Information identifies the access restrictions pertaining to the data, including the legal framework, licensing terms, and access control.

In the OAIS RM Information Model each of these categories of information are themselves Information Objects. This ensures that each has Representation Information to ensure that that it can be interpreted. For example, Provenance Information has its own Representation Information so that it can be understood by the user of the data.

The information categories of Reference, Provenance, Context, Fixity, and Access Rights Information are grouped together into what is called Preservation Descriptive Information (PDI). PDI along with Representation Information, is necessary for adequate preservation of the Data Object. Also, PDI is modeled as an Information Object.

Context Information in particular has an important role in enabling data reusability since it can be used to define the relationships of the Data Object to the things within its environment and provide additional semantic information. Context Information is best captured in a formal model to promote a common understanding because of the complexity of the relationships between the things in the environment.

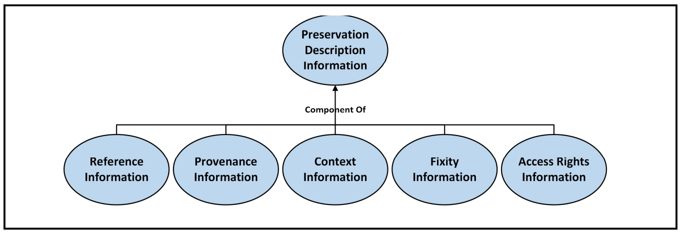

Figure 2 - Preservation Description Information

C. Metadata Registry

The ISO/IEC 11179:3 Metadata registries (MDR) [7] provides a data dictionary schema for data elements, for example the "start_time" of an event. This schema is necessary since where ontology development is focused on capturing concepts in the domain, additional detailed information about attributes of an objects is also needed, for example, alternate names, definition sources, definitions in other languages, effective dates, and stewardship. An attribute also has associated values, each with their own value type, data representation, units of measurement, effective dates, submitter, and steward. If the attribute value is an enumerated type, it will have permissible values. In turn each value in an enumerated list could have its own definition and effective dates. Again, an ontology modeling tool can be used to capture the metadata models in terms of classes, their attributes and relationships.

Finally, data engineers will typically want to be able to find all attributes that are similar in concept to determine if an attribute could be reused or whether a new attribute has to be defined. This scenario also applies to valid values and enumerated lists. In figure 3 a high-level depiction is provided of an ISO 11179 data element, it value domain and how they are classified.

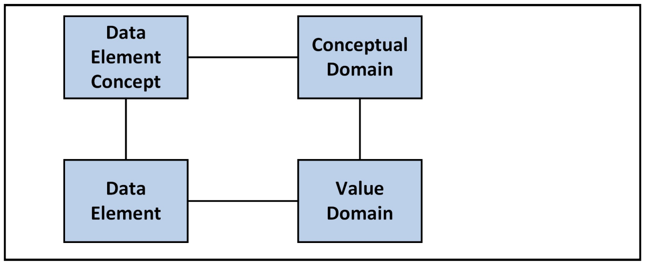

Figure 3 - ISO/IEC 11179 High-Level Metamodel

D. Registry

The ebXML federated registry model provided the concepts of the registry object, object classification, identification, and versioning. Object classification and identification support reusability and defining these in the model allows a registry to be configured from the model.

E. *Information Model Governance*

The PDS4 Information Model captures the knowledge about the planetary science community's digital repositories at several levels of specificity and provides a means for managing changes and their impact. Multi-level governance, depicted in figure 5, is instituted in the model at the common, discipline, and mission levels to enable interoperability at the appropriate level and simplify the management of the model as it evolves over time.

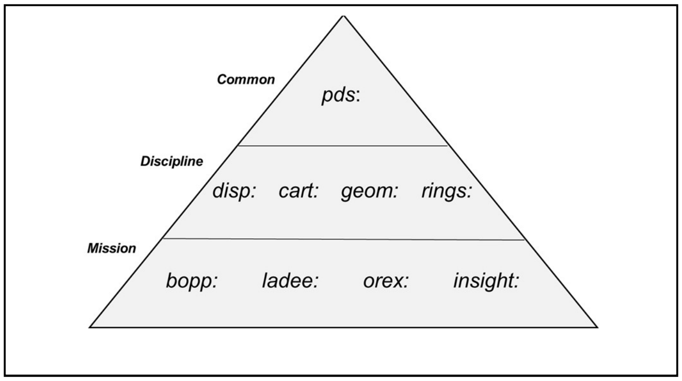

Figure 4 - Multi-Level Governance

At the common or upper level of an information model resides the knowledge about the "things" (digital objects and products can be located and retrieved and how they are identified, referenced, and packaged. Data objects in the repository must also have representation information provided in logical and well-defined terms so that they can be properly interpreted for scientific studies.

At the next level, shared knowledge in specific disciplines must be available to understand and advance science, for example standard geometry models are needed to determine locations and standard cartography models are required for describing maps. Finally, at the project or mission level, standard vocabularies are defined for local teams to communicate and effectively support the investigation.

A fundamental principle used in the development of the PDS4 Information Model is that the model remains independent from its implementation. In a classic enterprise architecture such as that presented in the Zackman Framework of Enterprise Architecture, the architecture is partitioned into architectural elements, for example "why", "how", "what", "who", "where" and "when".

The PDS4 Information Model encompasses the "what" element in the framework, that is, the data being processed or archived. The model is agnostic to the other elements, especially the "how", or the implementation element of the architecture. A model that is independent of the implementation is inherently more stable because it can more readily change as the science discipline changes. Concurrently it is shielded from technology changes which typically changes at more rapid pace.

To further manage complexities during the developmental and evolutionary phases of the PDS4 Information Model, the multi-level governance scheme is instituted at the common, discipline and mission levels. The common model is governed under a formal change control process where a change control board (CCB) decides whether to approve each change request based on the change's potential impact on the overall PDS enterprise. At the discipline and mission levels similar governance but with contextually limited scope are instituted.

F. *The Information Model in Context*

The PDS4 Information Model is generated from the ontology and its contents are extracted and exported to system files in various formats as can be seen in figure 4.

During the knowledge acquisition process, each "thing of interest" in the domain, for example a planetary image, is defined to the extent necessary to meet the functional requirements of the system. These definitions comprise the Domain Knowledge, the primary information captured in the ontology. The ontology also encodes general information requirements and the guidance provided by the ISO standard frameworks.

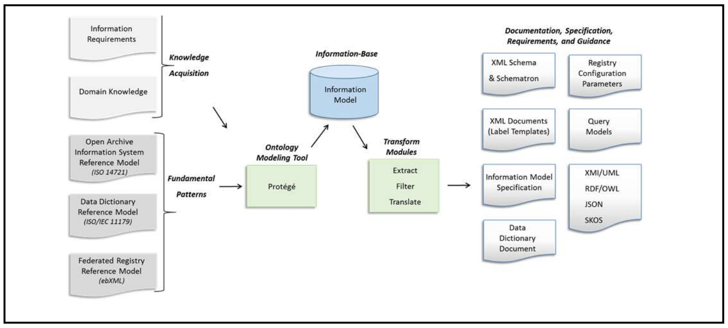
Information Requirements
Domain Knowledge
Knowledge Acquisition
Open Archive Information System Reference Model (ISO 14721)
Data Dictionary Reference Model (ISO/IEC 11179)
Federated Registry Reference Model (ebXML)
Fundamental Patterns
Ontology Modeling Tool
Protégé
Information-Base
Information Model
Transform Modules
Extract
Filter
Translate
Documentation, Specification, Requirements, and Guidance
XML Schema & Schematron
Registry Configuration Parameters
XML Documents (Label Templates)
Query Models
Information Model Specification
XMI/UML
RDF/OWL
JSON
SKOS
Data Dictionary Document

Figure 5 - The Information Model in Context

G. *Operational Artifacts*

The PDS4 Information Model remains independent of the implemented information system. Therefore, the contents of the ontology must be exported to operational artifact. The PDS chose XML Schema [9, 10] as its implementation language. This required the careful selection of a subset of the possible XML Schema constructs that carry the necessary semantics. For example, classes and their properties that were defined in the ontology were implemented in XML Schema as either xs:simpleType or xs:complexType depending on the complexity of the definition. They were also defined as using XML Schema xs:element for use in XML documents.

Once the Information Model had been translated to XML Schema, an issue arose that the Information Model seemed to "exist" in two languages. Designers who had become comfortable with an "objected-oriented" model now encountered the same objects in XML Schema, but defined and organized in a significantly different way.

To ensure reusability of the data, the information model independence principle was reapplied and the model and the system continued developed on parallel paths. Once developed the system services and tools responded as expected to the Information Model via extracted configuration files. The following operational artifacts are now generated for PDS4 label templates, label validation, user documentation, software configuration files, and semantic applications.

1) XML Schema – PDS4 Label Templates and label validation
2) Schematron – Extended PDS4 label validation
3) XML/XMI – UML documentation
4) JSON – Software configuration files
5) RDF/OWL – Semantic applications
6) DocBook – Documentation

III. Trusted Data Repository

A Trustworthy Digital Repository (TDR) is commonly defined as "a repository whose mission is to provide reliable, long-term access to managed digital resources to its customers, now and in the future." As regards data reusability, this includes aspects such as the use of standard file formats and encoding schemes, the preservation of high-resolution digital data, and the provision of appropriate licenses and permissions for the use of the data. Reusability is important because it ensures that digital information can be used and repurposed for different purposes, including research, education, and creative activities, over the long term. Providing appropriate documentation and training to stakeholders ensure that they are able to use and interpret the digital information correctly.

A. Knowledge Graph

A knowledge graph represents a network of real-world entities, objects, or concepts, and illustrates the relationship between them. The nodes of the knowledge graph in figure 6 depict a subset of the design principles and standards that have been discussed above. Node properties that promote data reusability in long-term digital repositories are identified along the named connections between the node. For example, the Information Object is the "core concept" within the Information Model that in turn provides the "information requirements" for a Trustworthy Digital Repository. The Information Model also requires "domain knowledge" from domain experts and the "principles" for data preservation and managing metadata are provided by the ISO 14721 and ISO/EIC 11179 standards.

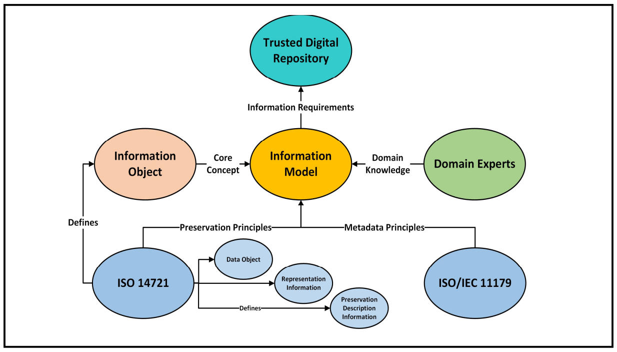

Figure 6 - Knowledge Graph - Properties for Data Reusability

IV. Conclusion

The FAIR data principle of Reusability continues to remain a challenge but is essential if the data is to be used in the future. To address this challenge, an Information Model was developed that provides a shared understanding of the data. The Information Model was captured using an ontology modeling tool to provide a formal, standard, and machine-readable specification. The ISO 14721 framework provided the archival principles required for long-term digital preservation. It also defined the Information Object and Preservation Description Information that guides the knowledge acquisition process. The Information Model was developed and is maintained independent of the implemented information system and the export of its contents to system files provide the information requirements necessary to promote data reusability.

Acknowledgment

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

References

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

[2] Crichton, D., Hughes, J.S., Hardman, S., Law, E., Beebe, R., Morgan, T., Grayzeck, E., A Scalable Planetary Science Information Architecture for Big Science Data, 10th IEEE e-Science conference, 2014.

[3] Crichton, D., Beebe, R., Hughes, S., Stein, T., Grayzeck, E., "PDS4: Developing the Next Generation Planetary Data System", EPSC Abstracts, Vol. 6, EPSC-DPS2011-1733, EPSC-DPS Joint Meeting 2011.

[4] Hughes , J.S., Crichton, D., Hardman, S., Law, E., Joyner, R., Ramirez, P., PDS4: A Model-Driven Planetary Science Data Architecture for Long-Term Preservation, IEEE 30th International Conference on Data Engineering (ICDE), Chicago, IL USA, 2014.

[5] Hughes, J.S., Crichton, D. J., Mattmann, C. A., "Ontology-Based Information Model Development for Science Information Reuse and Integration", 10.1109/IRI.2009.5211603, IEEE International Conference on Information Reuse & Integration, 2009.

[6] ISO 14721:2012: Reference Model for an Open Archival Information System (OAIS), ISO, 2012.

[7] ISO/IEC 11179: Information Technology -- Metadata registries (MDR), ISO/IEC, 2008.

[8] Extensible Markup Language (XML) 1.0 (Fifth Edition), W3C Recommendation, 26 November 2008.

[9] XML Schema Part 1: Structures Second Edition, W3C Recommendation, 28 October 2004.

[10] XML Schema Part 2: Datatypes Second Edition, W3C Recommendation, 28 October 2004.

[11] [W3C RDF/XML Syntax Specification (Revised), W3C Recommendation, 10 February 2004.

[12] (2013) The Protégé Ontology Editor and Knowledge Acquisition System website. [Online]. Available: http://protege.stanford.edu/.

[13] Reference Architecture for Space Information Management (RASIM), CCSDS 312-0.G-1.