



Fundamental Data Standards for Science Data System Interoperability

Earth and Space Science Informatics Workshop

J. Steven Hughes, Daniel Crichton
Chris Mattmann, Ron Joyner, Elizabeth Rye
August 2-4, 2010

Introduction

- With the advent of the internet, there has been an explosion of online science data repositories.
- However there have been relatively few successes in meeting the expectations of modern science users.
 - Systems do not interoperate at the levels promised
 - Data can only be correlated on surprisingly few attributes.
- The ability to interoperate and correlate is dependant on a relatively few types of fundamental data standards.

Shared Models

- Research¹ indicates that shared models are required to enable system interoperability and data correlation.
- Shared information models provide a common understanding of the science domains.
- Shared meta-models provide a common understanding of the information models.
 - Provide the framework for an information model.
 - Define the semantics (i.e. meaning) within an information model.

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

A Simple Example

- ISO 8601 - Representation of Dates and Times
 - Provides meaning of and standard formation rules for date and time based values.
 - E.g. `yyyy-mm-ddThh:mm:ss.sssZ`
- Used in the definition of `start_date` and `start_date_time`.
 - Provide the date a mission started or the time an observation started.
- Enables system interoperability and data correlation.
 - A query for science data products with `start_date_time` as a constraint can be distributed to all conforming repositories and the products returned can be correlated by time.

Some Fundamental Data Standards

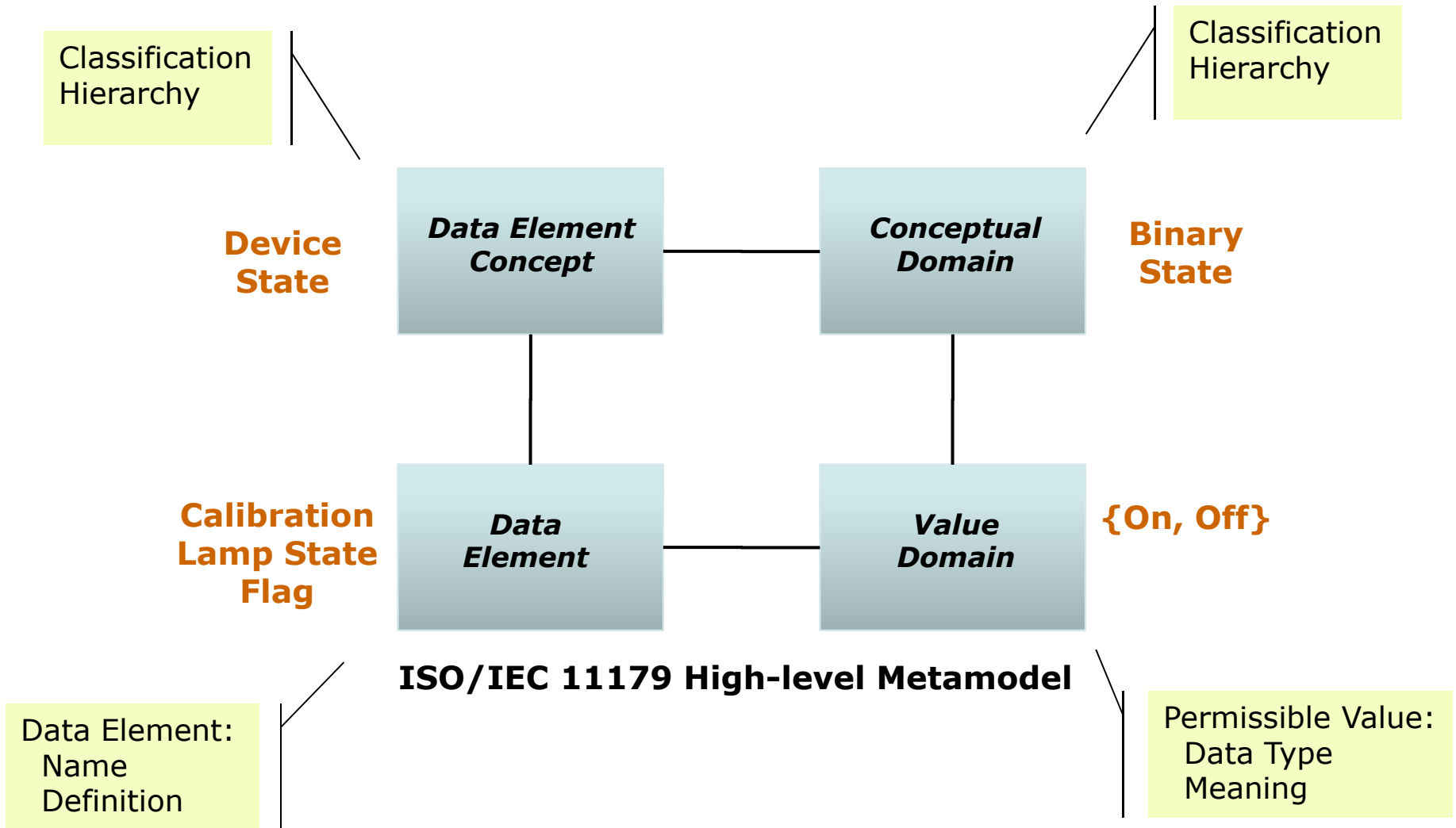
- ISO/IEC 11179:3 Registry Metamodel and Basic Attributes specification - Provides for a common schema for data dictionaries.
- ISO/IEC 11404:2007(E) - Provides the specification for language-independent data types.
- ISO 8601:2004 - Provides meaning of and standard formation rules for date and time based values.
- Open Archival Information System (OAIS) Reference Model - Provides a standard for the unified modeling of digital, conceptual, and physical data objects.
- Electronic Business XML (ebXML) federated registry/repository information model – Provides a standard to support federated registry/repository functions.
- The Dublin Core Metadata Initiative – Provides a global vocabulary of fifteen basic properties for use in resource description.

ISO/IEC 11179:3

Registry Metamodel and Basic Attributes

- A data element such as `start_date_time` is decomposed into four key components.
- Each component is administered separately.
 - Data Elements have names and definitions.
 - Permissible values have data types and meanings.
 - Each data element and permissible value exists in a classification hierarchy.
- Enables interoperability at a fundamental level.
 - A potential collaborator can understand the definitional framework and semantics of another repository's controlled vocabulary.

The Core Meta-Model



Important ISO/IEC 11179 Meta-Attributes

Data Element

- Name
- Submitter, Steward
- Definition
- Namespace
- Source of definition
- Change log
- Version
- Concept
- Alternate Names
- Definition in multiple natural languages
- Classification
- Unit of measurement
- Effective Dates

Value Domain

- Permissible Value
- Value Meaning
- Submitter, Steward
- Definition
- Cardinality
- Source of definition
- Change log
- Version
- Concept
- Character Set
- Representation
- Minimum and Maximum Value
- Minimum and Maximum Length
- Alternate encodings
- Effective Dates

Registry Reference Models

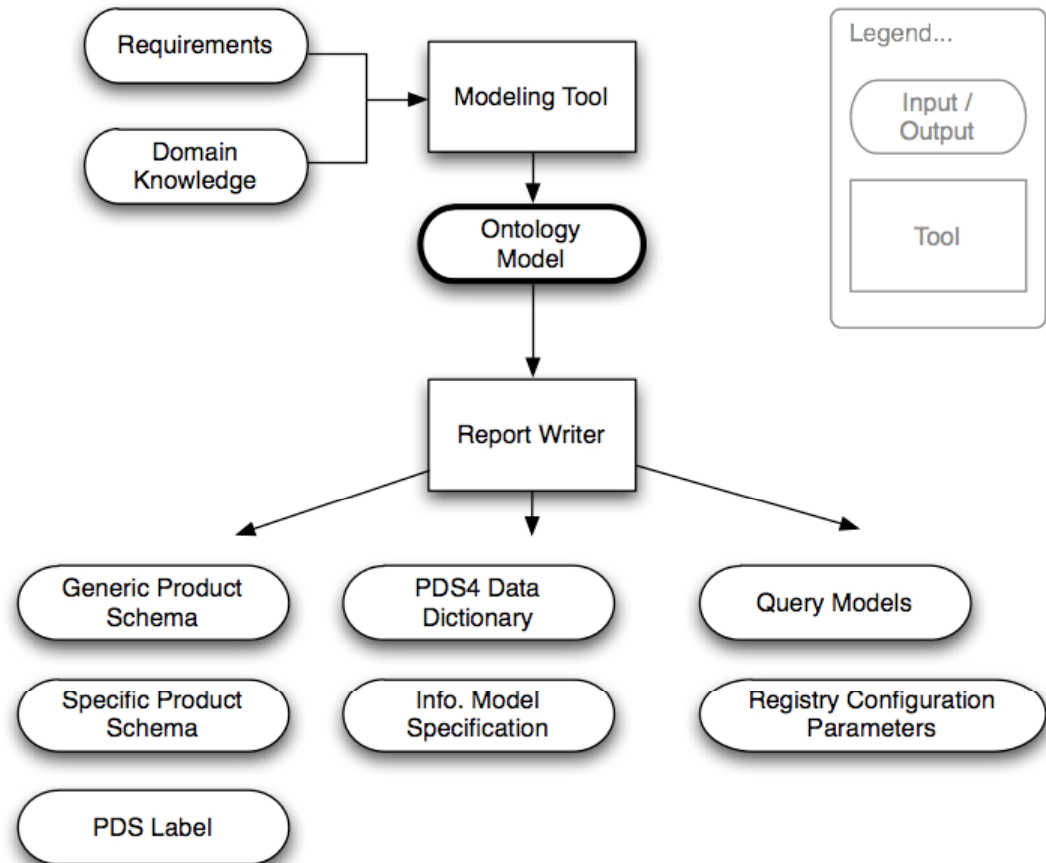
- Provides a standard to support federated registry/repository functions.
 - Electronic Business XML (ebXML) federated registry/repository information model – Version 3
 - CCSDS Reference Model for a façade API.
 - Publish, Version, StoreObject, Federated Query and Replicate
 - Intrinsic RegistryObjects – Content known
 - Services, Associations, Classification Schemes, Slots
 - Extrinsic RegistryObjects – Content not known
 - Domain objects – Images, Time Series, Missions, instruments
 - Registry configured for domain object
 - Suggests attributes for domain information model
 - Identifier, logical identifier, version identifier

Case Study - PDS 2010

- PDS 2010 is a PDS-wide project to upgrade PDS from PDS3 to PDS4
- A transition from a 20-year-old collection of standards and tools to a modern system constructed using best practices for data system development.
- Fewer, simpler, and more rigorously defined formats for science data products.
- Use of XML, a well-supported international standard, for data product labeling, validation, and searching.
- A hierarchy of data dictionaries built to the ISO 11179 standard, designed to increase flexibility, enable complex searches, and make it easier to share data internationally.

PDS Model Driven Process

- The ontology defines the things in the domain resulting in an information model.
- The 11179 compliant data dictionary defines data elements.
- The report writer uses the ontology's content and the data dictionary to project documents.



Summary

- The vast amount of science data now available in online data repositories provides a huge potential resource for scientific discovery.
- Scientists now expect systems that interoperate and data that can be correlated.
- Many technologies hold promise.
 - Domain interlinguas
 - Shared information models
 - Text and facet based search
 - Google-like technologies
 - Semantic technologies
- A few fundamental data standards lay the foundation and provide significant leverage for each of these approaches.

THANK YOU!