PDS Data Dictionary Discussion and Planning

PDS 2010 Tech Session

June 10, 2009

PDS 2010 Data Design WG

Topics

- **Purpose**
- **Concepts**
- **Recommendations**
- **Preliminary Design Decisions**
- **Plan**

Purpose

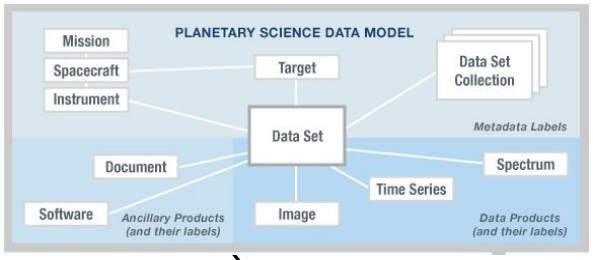
- **Present recommendations for a new data dictionary model.**
- **Discuss a few preliminary design decisions.**
- **Introduce a plan.**

Important Concepts

- **A data dictionary contains a set of defined data elements**
- **Each data element is defined by a set of attributes**
- **The Data Dictionary expresses the information model**
  - Structural Metadata
  - Descriptive Metadata
- **The Data Dictionary has a registration authority that is responsible for its maintenance.**
- **Data elements can be grouped and classified**
  - Example - namespaces
  - Allows a hierarchy of namespaces, from PDS common, to *discipline specific*, to local mission data dictionaries.

PDS4 Data Dictionary in Context

Information Model



Used to Create

Product Schema

Validates

Expressed As

Extracted/Specialized

Data Element

has

Class

Planetary Science Data Dictionary

Conforms to the ISO/IEC 11179 Specification

Product

Tagged Data Object (Information Object)

```
OBJECT = IMAGE_GRAYSCALE
  DATA_LOCATION          = ("N2075WE02R.FIT", 0)
  FIRST_ELEMENT          = TOPLEFT
  MIN_INDEX              = 0
  NUMBER_OF_AXES         = 2
  AXES_ORDER             = FAST2SLOW
  BYTE_ORDER             = MSBF
  ELEMENT_BYTES          = 2
  ELEMENT_TYPE           = DECIMAL_INTEGER
  ELEMENT_UNIT           = "DATA NUMBER"
  AXIS_NAME              = ("LINE", SAMPLE")
  AXIS_LENGTH            = (248, 256)
  AXIS_SCALE_TYPE        = ("N/A", "N/A")
  AXIS_UNIT              = ("N/A", "N/A")
END_OBJECT = IMAGE_GRAYSCALE
```

Describes

Data Object

Why a new data dictionary model?

- **Current model is simplistic and is limited by the ODL grammar.**
- **There is no constraint language.**
  - ***Required*** and ***Optional*** constraints are not enough.
- **Less than half of the current data dictionary requirements are met by the current model.**
- **We tried to invent the wheel once, let's not try to reinvent it.**

Recommendation #1
ISO/IEC 11179 – Metadata Registry Specification

- **Standard model for data dictionaries**
  - Addresses the semantics of data, the representation of data, and the registration of the descriptions of that data
- **Provides the following**
  - Standard description of data
  - Common understanding of data across organizational elements and between organizations

ISO/IEC 11179 Structure Summarized

- **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

- **Valid Value**
  - Value
  - 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

- **Class**
  - Attributes (Data Elements)

Recommendation #2
Namespace Management

- **Classify data elements in the dictionary into namespaces**
  - Local data dictionaries have already started down this path

- **Use namespaces to minimize global changes to the data dictionary and to coordinate at multiple levels (PDS-wide, International, Node, Mission)**

Namespace Examples

PDS Common	**target_name**
Discipline/ Agency/ Organization	***Imaging:emission_angle*** ***iau:feature_name*** ***dc:alternate***
Mission/ LDD	**mer:INST_DECOMP_STAGES**

Recommendation #3:
Set up a Data Dictionary Service

- **Export data dictionary content in a variety of formats.**
  - Label Schema/Tool Data Dictionary – Offline Validation
  - Hard Copy PSDD Document
  - Data Set Data Dictionary for archive volumes.

- **Manage and audit data dictionary changes.**

- **Enable online validation against the Data Dictionary via this service**

Proposed Design Decisions – 1 of 2

- **Adopt the concept of "steward" as "an organization that manages an administered item".**
  - For the PDS, an administered item can be a data element, or a valid value.
  - A steward could be a discipline node, the PDS, or an external agency.

- **Design a generalized namespace that can be implemented in multiple grammars.**
  - For example an ordered list of namespace_ids (references to dictionaries) could be provided and the attributes are located by searching the dictionaries in the order specified.

- **Allow attribute and class references to the current dictionary and any dictionary higher in a hierarchy.**
  - For example a mission data dictionary with GEO as the lead node could reference items in the *discipline node dictionaries* and the PDS common dictionary.

- **Do not allow lateral references between local data dictionaries.**
  - For example, Mission A can not reference Mission B's local data dictionary since Mission B could changed its dictionary without notifying mission A.

Proposed Design Decisions – 2 of 2

- **The sets of classes, relationships, and attributes grouped by a namespace will have the same steward and registration authority.**
  - For example, the attribute target_name with its set of standard values will be in the PDS Common data dictionary with the PDS Standards Coordinator as the steward and the PDS as the registration authority.
  - If the SBN wishes to change the name of an ASTEROID, then the SBN will submit a change request to target_name's steward, the PDS Standards Coordinator, who will process the change request in accordance with the change control process.
  - Any other node could also submit a request to change the name of the Asteroid.
- **All but the local data dictionaries will be federated.**
  - For example, even though the PDS common data dictionary is governed by the PDS standards process and *the Discipline Data Dictionaries* will be governed by the PDS Discipline Nodes, they will all be globally accessible through the data dictionary service.

Plan

- **Build the Planetary Science Data Dictionary (PSDD) using the ISO/IEC 11179 standard**
- **Organize data elements around a PDS namespace standard**
- **Setup an online data dictionary***
  - The complexity of the model will not necessarily propagate to users.
- **Develop tools for managing data dictionaries***

*** Systems Design.**

Backup

Some Questions a Data Dictionary Should Answer

- What is the organization that is responsible for managing the data element Sample_Type? Who can be contacted about its management? What is the source of the definition?
- What changes have been made to the definition and by whom? What is its version?
- What is the rationale for the data element? Is this a common concept shared by other data elements?
- What alternate names does the data element have?
- What natural languages are used in the definition?
- How is the data element classified? E.g. science vs operations, science discipline, etc.

More Questions

- **Who controls the valid value “ieee_real? Who can be contacted about its management? What is the source of the valid value?**
- **What is the definition of “ieee_real”?**
- **What changes have been made to the definition of “ieee_real” and by whom? What is its version? What is the window within which this value was valid?**
- **What is the rationale for the set of valid values? Is this a common valid vale concept that other data elements might use?**
- **What character set is used for the valid values? How many characters can be used for a valid value?**
- **How is the valid value represented and what is the unit of measurement?**
- **What are the minimum and maximum values allowed?**
- **What values signify Missing, Not Applicable, and Unknown for this data element.**

Status

- Use cases have been defined.
- Requirements have been written.
- Several data dictionary models were reviewed.
  - CCSDS Data Entity Dictionary Specification Language (DEDSL)
  - SPASE Data Dictionary
- Three data dictionary models have been captured in an ontology.
  - Planetary Science Data Dictionary (PSDD) – PDS Data Dictionary
  - ISO/IEC 11179-3:1994 – Specification and standardization of data elements
  - ISO/IEC 11179-3:2002 – Metadata registries (MDR)
- An ISO/IEC 11179-3:2002 database is being prototyped
  - Ontology modeling tool allows data to be ingested.
  - Planetary Science Data Dictionary content is being loaded.
  - Validation of model is not yet complete.

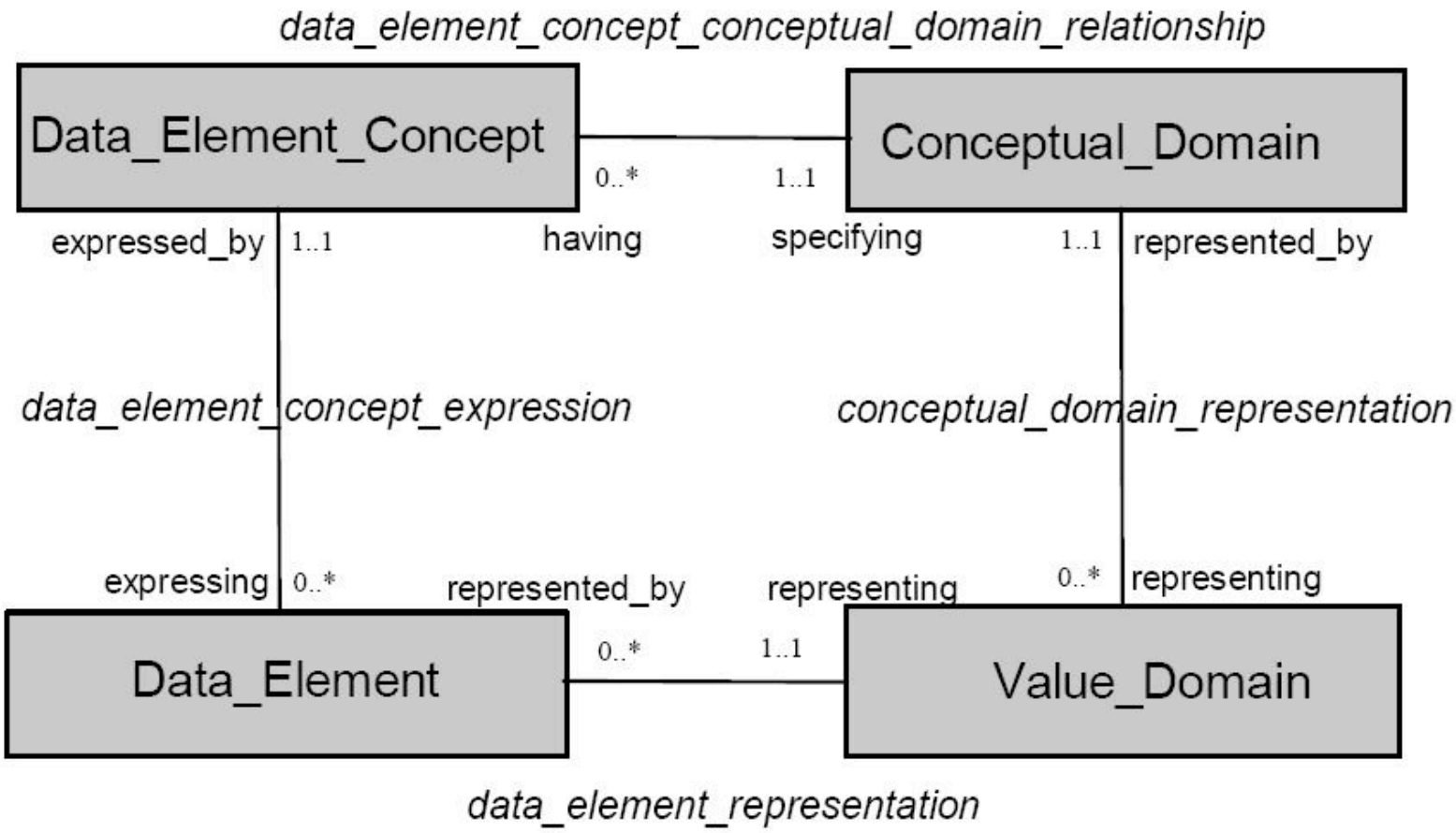
Requirements (1 of 2)

1. *The Data Dictionary shall define data elements in compliance with the data dictionary model .
2. *The Data Dictionary shall define a “units-of-measurement” value set in compliance with the data dictionary model.
3. The Data Dictionary shall define a “special-values” value set in compliance with the data dictionary model.
4. The Data Dictionary shall define a “data-element-value-alias” value set in compliance with the data dictionary model.
5. *The Data Dictionary shall define a “data-element-value-formation-rule” value set in compliance with the data dictionary model.
6. *The Data Dictionary shall define a “standard-value” value set in compliance with the data dictionary model.
7. *The Data Dictionary shall define a “namespace” value set in compliance with the data dictionary model.
8. The Data Dictionary shall define linked-in-kind data element relationships in compliance with the data dictionary model.
9. *The Data Dictionary shall have a naming standard for the data element title (common name).
10. The Data Dictionary shall provide a namespace attribute for indicating control authorities for groups of data elements.
11. The Data Dictionary shall provide a general data type attribute for classifying a data element according to a non-implementation-specific list of data types.
12. *The Data Dictionary shall provide a general classification type attribute for classifying groups of data elements according to common characteristics
13. *The Data Dictionary shall provide a system classification type attribute for classifying groups of data elements according to the data system that uses it

Requirements (2 of 2)

1. *The Data Dictionary shall provide an alias attribute for specifying one or more aliases that are applicable to the referenced data element.
2. *The Data Dictionary shall provide a standard value type attribute for specifying the type of standard value that is appropriate for the referenced data element.
3. *The Data Dictionary shall provide a minimum and maximum column value attribute for specifying the minimum and maximum numeric values that are applicable to the referenced data element.
4. *The Data Dictionary shall provide a minimum and maximum length value attribute for specifying the minimum and maximum permissible length of the character values that are applicable to the referenced data element.
5. The Data Dictionary shall provide two identifier attributes for specifying the unique instance of the data element and a locally defined instance of the referenced data element.
6. *The Data Dictionary shall provide a textual-description attribute for defining the referenced data element.
7. *The Data Dictionary shall provide a data-element-formation-rule attribute that supplies a rule that is to be applied during the creation of a value for the data element (e.g., the values supplied for reference_key_id must conform to the rules used by a specific professional journal for referencing citations).
8. The Data Dictionary shall provide a series-set attribute for specifying if it is permissible or not permissible to specify values in a series or set.
9. The Data Dictionary shall provide a data-element-partial-label attribute that specifies if it is permissible or not permissible for a data element to exist within a partial-label (e.g., a FMT file).
10. The Data Dictionary shall provide a has-units attribute that specifies if it is permissible or not permissible to associate a unit with the referenced data element.
11. The Data Dictionary shall provide a can-be-locally-defined attribute that specifies if it is permissible or not permissible for a data element to be locally defined (i.e., overwrite the attributes of a data element in the PSDD.
12. The Data Dictionary shall provide a can-take-on-identifiers attribute that specifies if it is permissible or not permissible to pre-pend an identifier to a data element.
13. The Data Dictionary shall provide a can-be-a-pointer attribute that specifies if it is permissible or not permissible for the data element / object to be characterized as one of the three types of pointers (e.g., (1) data location pointer, (2) include pointer, and (3) related information pointer).
14. The Data Dictionary shall provide attributes for defining object classes.

ISO/IEC 11179 Concepts



ISO/IEC 11179
Data Element Information

- **Data Element**
  - Data Element, Data Element Concept
  - Concept Domain, Value Domain
- **Administration and Registration**
  - Status, Creation_Date, Effective_Date, Last_Change_Date, Registration_Authority, Steward, Submitter, Language
- **Classification Scheme - User defined**
  - Scope – e.g. Common, Imaging, Atmospheres
  - Mission – e.g. MEX, MER
- **Other**
  - Description, Data Type, Enumeration, Unit of Measurement, Value Meaning, Alternate Names, Language, Reference Documents, Relationships, Terminology
  - Effective begin and end date

Planetary Science Data Dictionary Example

- **Data Element_Name:** Sample_Type
- **Data_Element_Description:** The sample_type element indicates the data storage representation of sample value (within an image).
- **Value_Type:** Identifier (Enumerated)
- **Valid_Values:** ieee_real, lsb_integer, …

PDS4 Data Dictionary Goals

- **Build it on a standard structure**
- **Streamline keyword changes, when needed**
- **Partition the data dictionary into name spaces**
  - International, PDS Common, Node, Mission, etc
  - Local Data Dictionaries
- **Make the model explicit.**
- **Significantly clean up the DD content**
- **Use it to drive interoperability**
- **Make creation/management of local data dictionaries more efficient**

Group Discussion: What are others?

Current Challenges

- **Underlying structure of the data dictionary is inadequate for our needs**
- **Limited namespace management (e.g., classification of keywords/data elements)**
- **Keyword changes are time consuming – No data element context.**
- **Much of the model is implicit.**

Group Discussion: What are others?

ISO/IEC 11179

- ISO/IEC 11179 - Metadata registries (MDR)
  - Addresses the semantics of data, the representation of data, and the registration of the descriptions of that data.
  - It is through these descriptions that an accurate understanding of the semantics and a useful depiction of the data are found.
- The purposes of ISO/IEC 11179 are to promote the following:
  - Standard description of data
  - Common understanding of data across organizational elements and between organizations
  - Re-use and standardization of data over time, space, and applications
  - Harmonization and standardization of data within an organization and across organizations
  - Management of the components of data
  - Re-use of the components of data
- In ISO/IEC 11179 the basic container for data is called a data element. It may exist purely as an abstraction or exist in some application system.

Sample Type – Data Element

Instance: de_sample_type

Types
- DataElement

Own Slots

	Slot Name	Value
▬	administrationRecord	Test_Load
▬	classification	
▬	contextEntry	Context_Entry_NASA_PDS
▬	dataElementPrecision	
▬	dataIdentifier	de_sample_type
▬	derivation	
▬	exemplifiedBy	
▬	expressedBy	dec_sample_type
▬	reference	
▬	registeredBy	RA_0001_NASA_PDS_1
▬	registrationAuthorityIdentifier	0001_NASA_PDS_1
▬	representationClassQualifier	
▬	representing1	
▬	steward	Steward_PDS
▬	submitter	Submitter_PDS
▬	typeBy	
▬	versionIdentifier	0.080410

^ back to top

Sample Type – Data Element Concept

Instance: dec_sample_type

Types
- DataElementConcept

Own Slots

	Slot Name	Value
	administrationRecord	Test_Load
	classification	
	contextEntry	Context_Entry_NASA_PDS
	dataIdentifier	dec_sample_type
	expressing	de_sample_type
	objectClass	
	objectClassQualifier	
	property	
	propertyQualifier	
	reference	
	registeredBy	RA_0001_NASA_PDS_1
	registrationAuthorityIdentifier	0001_NASA_PDS_1
	relationship	
	specifying	ecd_sample_type
	steward	Steward_PDS
	submitter	Submitter_PDS
	versionIdentifier	0.080410

^ back to top

Sample Type – Concept Domain

Instance: ecd_sample_type

Types
- EnumeratedConceptualDomain

Own Slots

	Slot Name	Value
	administrationRecord	Test_Load
	classification	
	containedIn2	vm_sample_type_IEEE_REAL, vm_sample_type_LSB_INTEGER, vm_sample_type_LSB_UNSIGNED_INTEGER, vm_sample_type_MSB_INTEGER, vm_sample_type_MSB_UNSIGNED_INTEGER, vm_sample_type_PC_REAL, vm_sample_type_UNSIGNED_INTEGER, vm_sample_type_VAX_REAL
	contextEntry	Context_Entry_NASA_PDS
	dataIdentifier	ecd_sample_type
	dimensionality	
	having	dec_sample_type
	reference	
	registeredBy	RA_0001_NASA_PDS_1
	registrationAuthorityIdentifier	0001_NASA_PDS_1
	relationship	
	representing2	evd_sample_type
	steward	Steward_PDS
	submitter	Submitter_PDS
	versionIdentifier	0.080410

^ back to top

Sample Type – Value Domain

Instance: evd_sample_type

Types
- EnumeratedValueDomain

Own Slots

	Slot Name	Value
▬	administrationRecord	Test_Load
▬	classification	
▬	containedIn1	pv_sample_type_IEEE_REAL, pv_sample_type_LSB_INTEGER, pv_sample_type_LSB_UNSIGNED_INTEGER, pv_sample_type_MSB_INTEGER, pv_sample_type_MSB_UNSIGNED_INTEGER, pv_sample_type_PC_REAL, pv_sample_type_UNSIGNED_INTEGER, pv_sample_type_VAX_REAL
▬	contextEntry	Context_Entry_NASA_PDS
▬	dataIdentifier	evd_sample_type
▬	datatype	IDENTIFIER
▬	maximumCharacterQuantity	255
▬	reference	
▬	registeredBy	RA_0001_NASA_PDS_1
▬	registrationAuthorityIdentifier	0001_NASA_PDS_1
▬	relationship	
▬	representedBy1	ecd_sample_type
▬	representedBy2	de_sample_type
▬	steward	Steward_PDS
▬	submitter	Submitter_PDS
▬	typeBy	
▬	unitOfMeasure	N/A
▬	valueDomainFormat	
▬	versionIdentifier	0.080410

^ back to top

Sample Type – Value Meaning and Permissible Values

Instance: pv_sample_type_IEEE_REAL

Types
- PermissibleValue

Own Slots

	Slot Name	Value
▬	aggregate	
▬	beginDate	1990-06-01
▬	containing1	evd_sample_type
▬	endDate	2019-12-31
▬	usedIn	vm_sample_type_IEEE_REAL
▬	value	IEEE_REAL

^ back to top

Instance: vm_sample_type_IEEE_REAL

Types
- ValueMeaning

Own Slots

	Slot Name	Value
▬	beginDate	1990-06-01
▬	containing2	ecd_sample_type
▬	description	TBD Description
▬	endDate	2019-12-31
▬	has	pv_sample_type_IEEE_REAL

^ back to top

Sample Type – Admin Record

Instance: Test_Load

Types
- AdministrationRecord

Own Slots

	Slot Name	Value
	administeredItem	
	administrativeNote	Test load from Planetary Science Data Dictionary (PSDD)
	administrativeStatus	Final
	changeDescription	No changes have been made to PSDD content.
	creationDate	2008-07-07
	effectiveDate	2008-07-07
	explanatoryComment	This is a test load of a ISO/IEC 11179 Data Dictionary using PSDD content.
	lastChangeDate	2008-07-07
	origin	Planetary Data System
	registrationStatus	Preferred
	unresolvedIssue	Issues still being determined.
	untilDate	2018-04-10

^ back to top

Sample Type – Registration Authority

Instance: RA_0001_NASA_PDS_1

Types
- RegistrationAuthority

Own Slots		
	Slot Name	**Value**
	documentationLanguageIdentifier	Language_Identification_English
	languageUsed	Language_Identification_English
	organizationMailingAddress	4800 Oak Grove Drive
	organizationName	NASA Planetary Data System
	registering	
	registrar	PDS Registrar
	registrationAuthorityIdentifier_v	0001_NASA_PDS_1

^ back to top

Sample Type – Registrar

Instance: PDS Registrar

Types
- Registrar

Own Slots

	Slot Name	Value
	contact	Elizabeth Rye
	registrarIdentifier	
	represents	RA_0001_NASA_PDS_1

^ back to top

Instance: Elizabeth Rye

Types
- Contact

Own Slots

	Slot Name	Value
	contactEmailAddress	Elizabeth.Rye@jpl.nasa.gov
	contactInformation	TBD
	contactMailingAddress	Jet Propulsion Laboratory, MS-169-315, 4800 Oak Grove Dr., Pasadena, CA 91109
	contactPhone	818.354.6135
	contactTitle	Standards Coordinator
	contractName	Elizabeth Rye

^ back to top

Terminology

- A Data Model defines the entities to be processed, their attributes, and the relationships that add meaning.
- An attribute has alternate names.
  - Data Element
  - Vocabulary Term
- The set of all attributes in a data model is also called its vocabulary and is collected into a data dictionary.
- When defining an attribute, a set of meta-attributes or “attributes about attributes” are used.
  - The *name* of an attribute is a meta-attribute.
  - For example when defining the data element *sample_type*, the meta-attribute, *attribute_name*, has the value “sample_type”.

Problem 2 – Enabling Interoperability

- Option 1 - The adoption of a common vocabulary by two or more repositories enables inter-operability between those entities.
  - **For example, the PDS Imaging and Geoscience node both use imaging vocabulary.**
- Option 2 - Where a single vocabulary is not in common between two repositories, inter-operability is dependant on the identification of commonalities between vocabulary terms.
  - **For example, two different instrument teams having similar instruments but separated by time and mission, should be at least using a common schema for their new vocabularies.**
  - The use of a common mechanism for defining vocabulary terms makes it easier to identify commonalities between vocabulary terms.
  - A common mechanism for defining vocabulary terms provides interoperability at a deeper level.

Key Issues

- Hierarchy (Governance) - Governance is the exercise of authority or control. This addresses the issue of who has control over each registered_item where registered_item = {class, attribute, attribute value, ...}.
  - Assign a registration authority and a steward to every registered_item.
  - A steward is defined as an organization (individual) that maintains stewardship of an administered item. Examples of stewards are the PDS as a whole, a discipline node, a mission, an instrument team, etc.
  - Any steward could have governance over one or more registered_items..
- Namespace - A set of names in which all names are unique.
  - Namespaces are a commonly accepted way to define a collection of things.
  - Namespaces provide a means for defining collections of registered_items, resulting in distinct data dictionaries.
  - Subsequently a PDS local data dictionary is logically a unique namespace where all registered_items in that namespace have a common steward.
    - Actually do we need to require a single common steward?
- Contents of data dictionary
  - Allow for both attributes (data elements) as well as classes.
  - For PDS4 this seems to be an important feature for supporting our model.
    - The PDS3 data dictionary supports both.

Key Issues

- Hierarchy (classes and their attributes)
  - The data dictionary will need to support the model. So for example, a local data dictionary may reference a class in the global dictionary and extend it.
  - Any new attributes or restrictions on an inherited attribute are in the local data dictionary namespace and governed by the local data dictionary steward.
    - This requires that proposed changes to a referenced dictionary can be allowed only after all possible impacts in other dictionaries are resolved.
    - Question: Why not allow a reference to a lateral dictionary, say from the GEO to the Imaging dictionary?
    - Question: Why not allow a reference from the GEO to a specific Imaging Instrument dictionary?
    - Is a network as opposed to a hierarchy is needed?
    - What if a locally defined data dictionary becomes globally useful? Will we then promote the local data dictionary or is it simply global since it is referenced globally.
  - Data modeling rules should primarily determine what can reference what.
  - The data dictionary content should capture the PDS4 data model and add administration functionality.
- Hierarchy (Globally-used versus Used-once)
  - This hierarchy will exist implicitly but will probably need to be explicitly defined.
  - It is important to help determine the level of impact a change might make.
    - Could the level of impact be represented by some type of metric on the number of references
- Implementation and Deployment
  - Implementation and deployment is a no-brainer from an engineering perspective.
    - i.e. implement and deploy a centralized data dictionary with services that meet the federated PDS requirements.
  - If politically untenable, then a federated data dictionary will need to be implemented with its additional complexities.