Project no. 018340

Project acronym: EDIT

Project title: Toward the European Distributed Institute of Taxonomy

Instrument: Network of Excellence

Thematic Priority: Sub-Priority 1.1.6.3: “Global Change and Ecosystems”

C5.80 Review of CDM v.1 and model for descriptive data in CDM v.2

Due date of component: Month 35
Actual submission date: Month 35

Start date of project: 01/03/2006 Duration: 5 years

Organisation name of lead contractor for this component: 2 MNHN

Draft for revision

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	**X**
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

C5.80 Review of CDM v.1 and model for descriptive data in CDM v.2

Relying on recent experience acquired through the use of the model for descriptive data in CDM v1, we reviewed the existing Java classes. We propose some improvements to ensure an optimal handling of data, as well as compatibility with existing descriptive standards and tools. This report provides a summary of the comments made on the first version of the CDM descriptive data model and propositions of evolutions for an improved management of descriptive data in the CDM.

Overview of CDM v1 model for descriptive data

The CDM v1 descriptive data model has been designed to allow the expression and structuring of descriptions of taxa, specimens or taxon names. Descriptions are expressed through two main Java abstract classes of the eu.etaxonomy.cdm.model.description package (cf. description diagram at http://wp5.e-taxonomy.eu/cdm/v14/):

- the DescriptionBase class represents a description as a whole: either a taxon description, a specimen description, or a taxon name description. It is composed of DescriptionElementBase objects.
- the DescriptionElementBase class represents a unit of descriptive information. This unit is associated with a feature (or character, or descriptor). It is expressed according to the type of information recorded (quantitative, categorical, textual, distribution, common name, association with another individual, interaction with another taxon).

The way characters are defined and handled in a descriptive model can vary a great deal. The CDM bias for the Feature class is original and makes the descriptive model very flexible. Contrary to a descriptive model such as SDD, the TDWG international XML standard for structured descriptive data, features (or characters, or descriptors) are not typed. The CDM Feature object can support any type of data whether categorical, quantitative, textual, or any other defined type. The handling of characters is thus performed through the unique Feature class. The typing of data is relegated to the level of the DescriptionElementBase.

Finally, in the CDM, features can be organised in hierarchy using the classes FeatureNode and FeatureTree. Hierarchy and dependencies between characters, as well as inclusive conceptual groups, can be stored in these objects.

Issues raised by the CDM v1 model for descriptive data.

The setting up of tools such as the import/export between CDM and SDD allowed the handling and testing of the CDM model for descriptive data. Some weaknesses stood out. The main ones are described below:

- concerning the **Feature** class: the available attributes to express what type of data is supported by a certain Feature are not homogeneous. The Feature object cannot be typed or multi-typed unambiguously; information is only available at the DescriptionElementBase level.
- concerning the links between **DescriptionBase** objects: the fact that structured descriptions are not always linked with a scientific taxonomic name raises problems for regrouping related descriptions. If the only possibility to regroup descriptions is by using the association with an existing taxonomic hierarchy, it limits the possibility of extracting sets of descriptions from the CDM. In addition, when importing data into the CDM, the information on potential connections between descriptions other than taxonomic is lost if not structured identically (e.g. use of the Scope class). A model such as SDD uses a Dataset object which contains a set of descriptions that can be tagged with a name, a description and media objects.

- concerning **general properties** shared by CDM objects: some essential general properties (title, description, media and original sources) are available only for a restricted number of objects. As examples, an instance of Feature cannot be linked with an OriginalSource object or a Media object; a FeatureNode cannot have a title, a description, or be linked with an OriginalSource object or a Media object.

Proposed changes to the model for CDM v2.

To suppress the issues presented above, some evolutions to the CDM v1 model for descriptive data are proposed hereafter. These suggestions as well as other less structurally impacting changes are discussed into more details in the following wiki page:
http://dev.e-taxonomy.eu/trac/wiki/CdmVersionTwoDiscussion.
Some of the proposed modifications were tested to evaluate potential consequences across the CDM.

- ### Evolution of existing classes and new objects
  - ### Feature

The objective of this proposed modification is to clarify the multi-typing of the Feature class and to separate the Feature object itself from the values that can be used associated with a feature in a particular context (e.g. taxonomic scope). A new boolean attribute that indicates if the Feature supports CategoricalData would be created. In parallel, all attributes related to supported categories, modifiers, statistical measures, etc. would be moved to a new abstract class. From this abstract class, named PossibleValues, would inherit new classes such as PossibleStates or PossibleStatisticalMeasures (see diagram 1). Feature gains a new 'recommendedPossibleValues' attribute in which can be listed all the different set of PossibleValues, independently from the typing of these values.

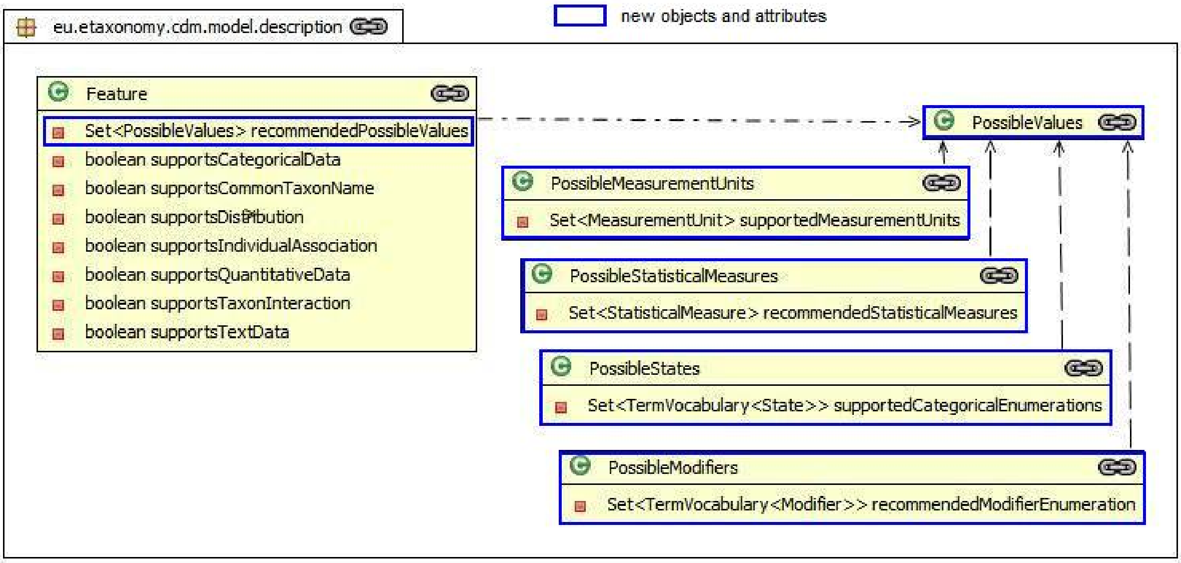
eu.etaxonomy.cdm.model.description
new objects and attributes
Feature
Set<PossibleValues> recommendedPossibleValues
boolean supportsCategoricalData
boolean supportsCommonTaxonName
boolean supportsDistribution
boolean supportsIndividualAssociation
boolean supportsQuantitativeData
boolean supportsTaxonInteraction
boolean supportsTextData
PossibleValues
PossibleMeasurementUnits
Set<MeasurementUnit> supportedMeasurementUnits
PossibleStatisticalMeasures
Set<StatisticalMeasure> recommendedStatisticalMeasures
PossibleStates
Set<TermVocabulary<State>> supportedCategoricalEnumerations
PossibleModifiers
Set<TermVocabulary<Modifier>> recommendedModifierEnumeration

Diagram 1 - Proposed modified Feature class

The PossibleValues object brings possibilities of distribution and exchange because it can be distinctly identified from the object Feature.

- ### WorkingSet

The objective of this evolution is to be able to store groups of descriptions together but not based on a scientific taxon names. A new class called WorkingSet could be created containing a set of DescriptionBase objects and a descriptive system. The descriptive system attribute would point to a new type of object (DescriptiveSystem) containing a set of features which use is shared by the concerned WorkingSet (see diagram 2). WorkingSet becomes an equivalent of the SDD Dataset element.

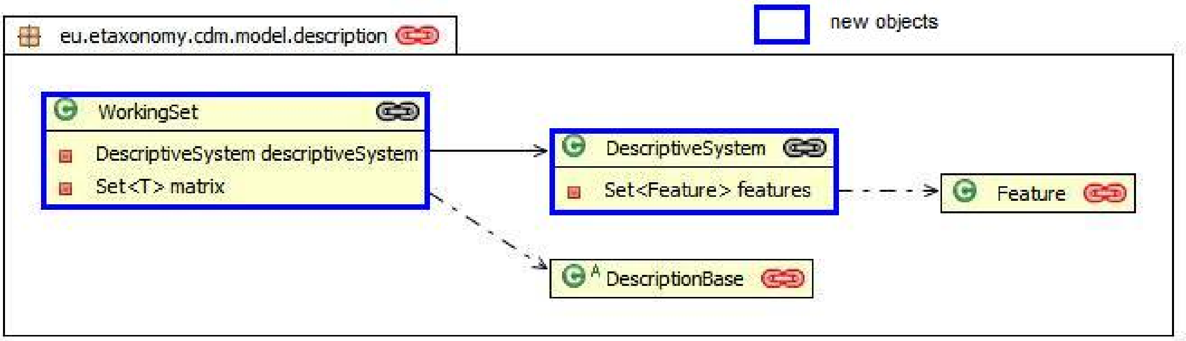

Diagram 2 - Proposed new WorkingSet and DescriptiveSystem classes

- **Status**

A new object would be created to be able to express a generic status information about a DescriptionBaseElement. For example, if a feature is "color of wings" but the described taxon has no wings, the description element associated with this feature is "not applicable". It needs to be distinguished from a case where the description information is missing because "not available". A DescriptionBaseElement could have a status attribute corresponding to a Status object expressing this information. This object would be similar to the DataStatus element in SDD and allow the recording of standardized reasons why data are missing. UBIF terminology (common foundation for several TDWG/GBIF standards like SDD) would be used (see diagram 3).

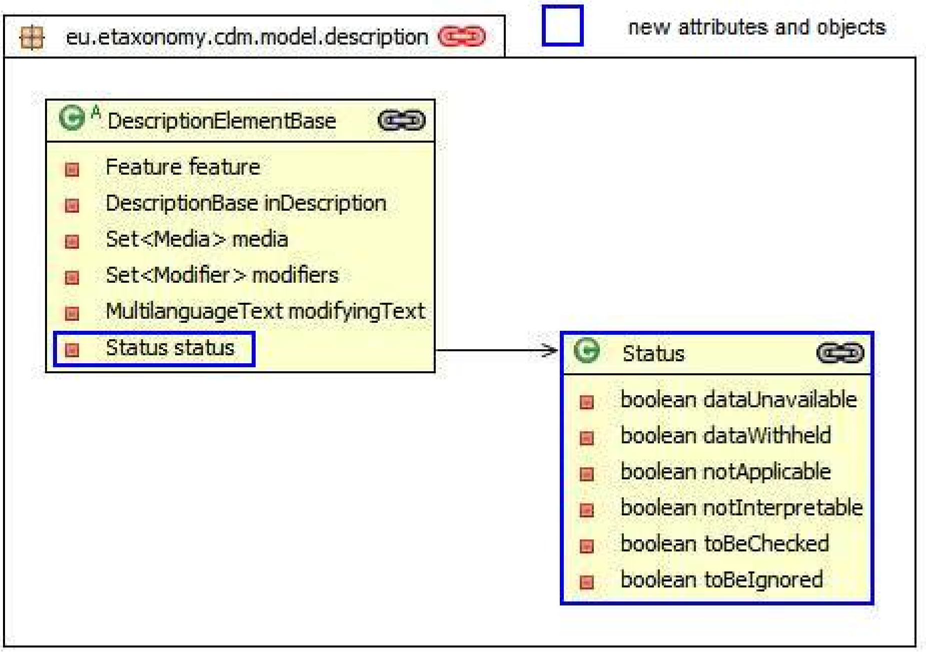

Diagram 3 - Proposed modified DescriptionElementBase class

- **Proposed structural change**

To make generic properties available to a large number of CDM objects, a solution could be to set these properties higher in the CDM hierarchy of objects. The existing Representation object would need to be modified and made independent from VersionableEntity. The sources property could be carried at the level of the CdmBase as it should not evolve with the different versions of a CDM object. The modified Representation class, carrying the other properties (title, description and media objects, would be available as a VersionableEntity attribute (see diagram 4). Moreover, the handling of languages could be centralized and simplified thanks to the use of the MultiLanguageText class for textual properties (title and description).

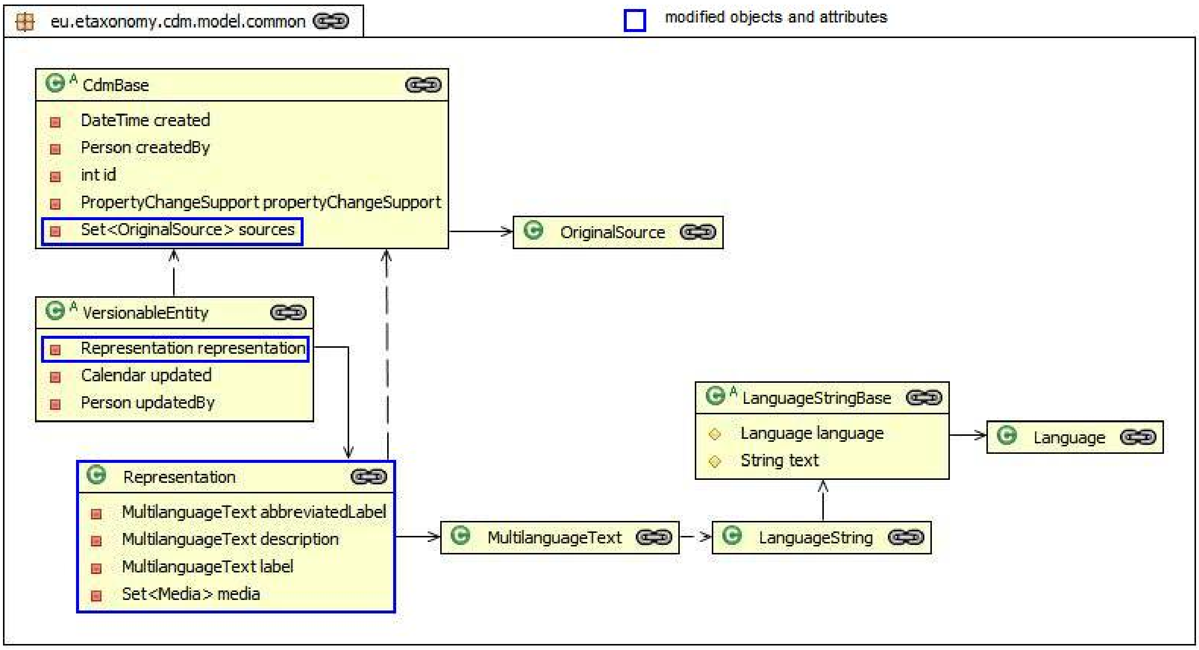

Diagram 4 - Proposed modified Representation class

The diagram is simplified. This proposition implies further changes to the model as Media, OriginalSource and Language classes should not depend any more from VersionableEntity. Impacts versus benefits should be evaluated.
Representation becomes an equivalent of the SDD Representation element.