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"

C5.117 Specification of a CDM Library functionality for the output of structured descriptive content in structured and textual form

Due date of component: Month 40
Actual submission date: Month 40

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)	

1. Desirable types of outputs for formalized descriptive content stored in the CDM

The cdmlib-model.description classes are designed to manage and store structured descriptive content. In order to display this content, different types of outputs need to be specified. As for today, the main use-case is the necessity to create output to the EDIT portals and potentially to the Taxonomic Editor. More generally, a flexible output for display on Web pages or in a format targeted to be read by another system would cover most use-cases.

The taxonomist will need to be able to choose which information is to be displayed: group of interest, selection of taxa and features (**CONTENT** and **FILTERING**); and to decide how information should be displayed (**LAYOUT** and **FORMAT**).

Formalized descriptive content can be represented either through **structured** output, showing the original formalization of the content, or through **natural language** textual output, presenting content in a readable, user-friendly way. To make a clear distinction between these two types of output, a very simple example is presented below.

Out of concern for clarity, let us consider 1 taxon described by 1 categorical character with 3 states. In reality, sets of descriptions will of course contain multiple taxa, and multiple categorical, quantitative or textual characters.

Fictitious example inspired from the taxonomic knowledge base: Anatomy of the Old-World palms (Romain Thomas, http://lis.snv.jussieu.fr/apps/xper/data/Palm-ID/)

The formalization of descriptive data used below is shared by commonly-used descriptive data models including the CDM:

- taxon: *Phoenix dactylifera* L.
- character: "phloem organization"
- states associated with the character: "with a single phloem strand, cells not sclerotic" (s1), "with a single phloem strand and conspicuously sclerotic" (s2), "phloem divided into 2 separate strands" (s3)
- available modifiers: "often" (m1), "in the spring" (m2)
- in addition, the character is included in the group "Leaf axis"

Description of *Phoenix dactylifera* L.: phloem organization: s2 [m1] + s3 [m2]

1.1. Structured textual output

a) In the form of enumeration

Phoenix dactylifera L.

Leaf axis

- phloem organization : with a single phloem strand and conspicuously sclerotic [often] ; phloem divided into 2 separate strands [in the spring]

b) In the form of a taxa x characters matrix

Taxa \ Characters	phloem organization [**Leaf axis**]
Phoenix dactylifera L.	phloem organization : with a single phloem strand and conspicuously sclerotic [often] ; phloem divided into 2 separate strands [in the spring]

1.2. Natural language textual output

Phoenix dactylifera L.

"The phloem in the leaf axis is often organized in a single phloem strand and conspicuously sclerotic, or divided into 2 separate strands in the spring."

2. Global activity diagram of a request for a descriptive content output by a taxonomist

The following activity diagram represents the different options that could be made available to the taxonomist to customize an output for descriptive content available in the CDM. It details a use case which was already specified in the component C5.081 Use case model and functional description of CDM descriptive data editor.

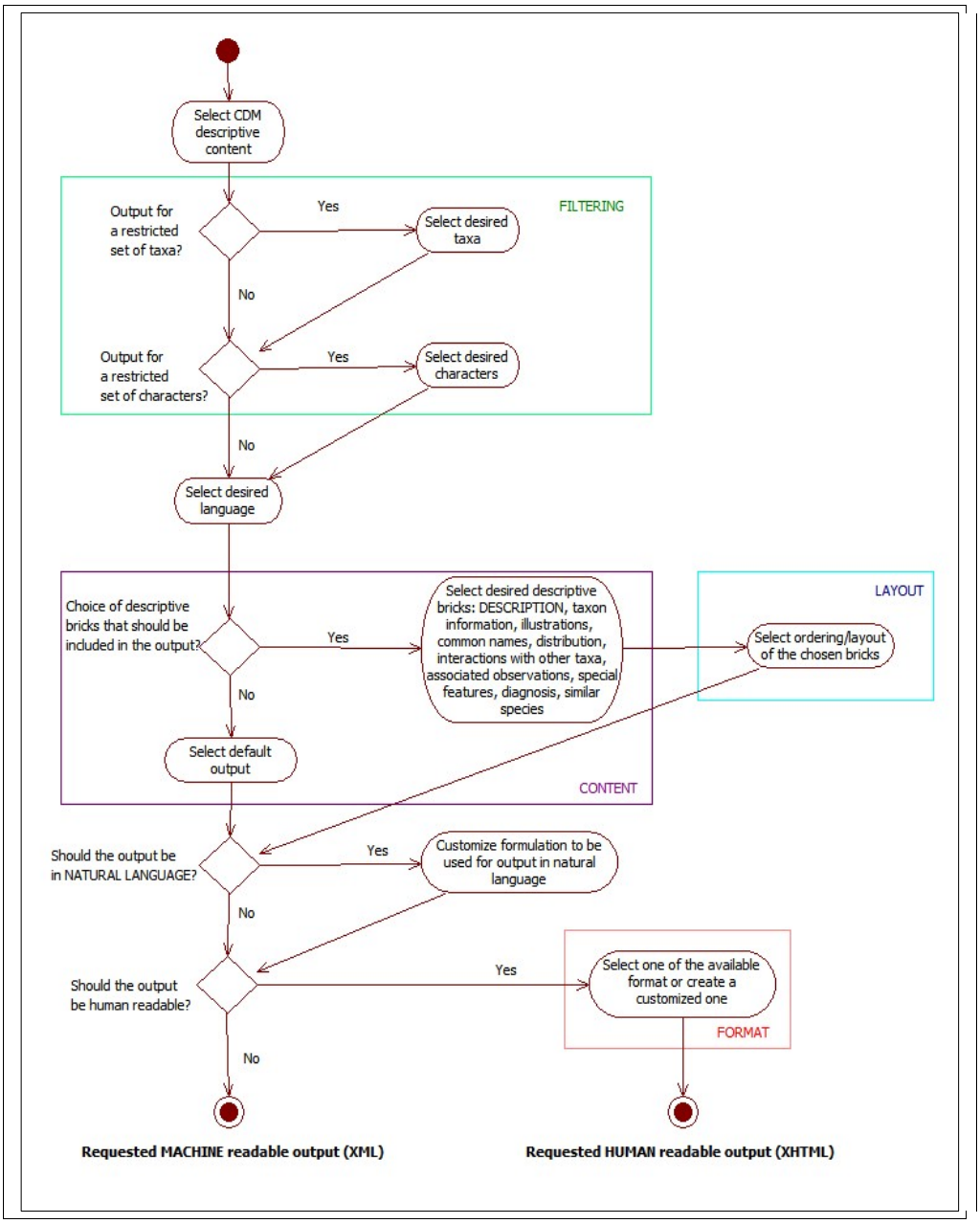

Diagram 1: Request for CDM descriptive content output

The diagram concentrates on descriptive content output requested by the taxonomist. However, such an output can also be combined with other types of data associated with taxa (possibly coming from the CDM too).

3. New and modified CDM objects to provide descriptive outputs

To provide the described output to taxonomists, the creation of two **new packages**, as well as the **modification of existing CDM objects** is proposed. The modification concerns objects from the library **cdmlib-model** / package eu.etaxonomy.cdm.model.description with the creation of new attributes to store alternative formulation for features, states and modifiers for the generation of natural language. One new package will be part of the library **cdmlib-services**: eu.etaxonomy.cdm.api.service and will contain classes to transform CategoricalData and QuantitativeData into natural language. The second new package will be part of **cdmlib-io**: eu.etaxonomy.cdm.io.description and will contain **classes which create text representations for descriptive objects**, from small modules returning representations for description elements to more substantial combination of descriptive content such as a complete taxon description or a default descriptive form.

The following diagram represents the impacted CDM objects, the proposed modifications and creations.

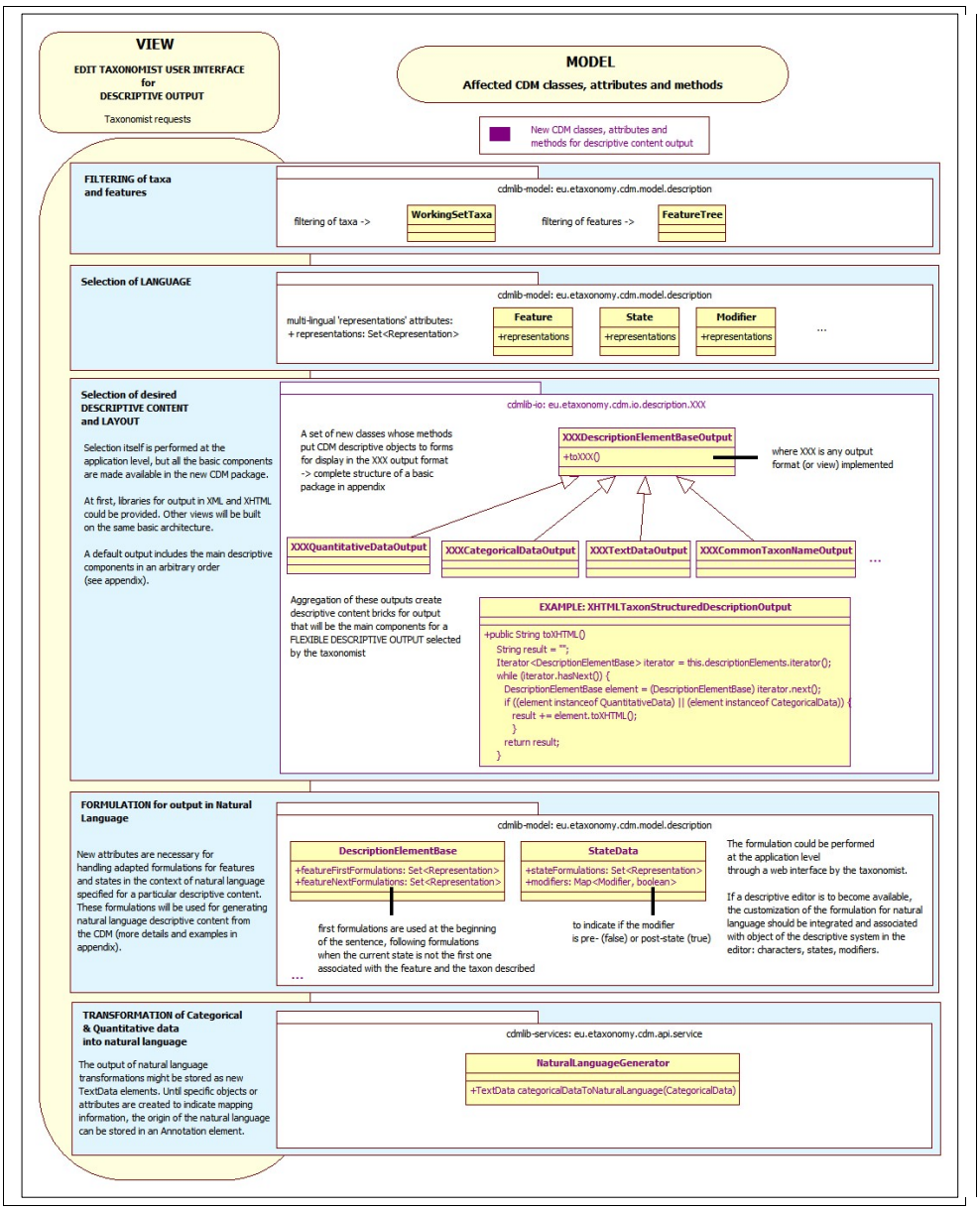

Diagram 2: CDM classes for descriptive content output

The new classes will allow the building of output **components representing fragments of a descriptive form** such as: description name, description itself, information, illustrations, common names, distribution information, associated specimens and observations, interactions with other taxa, and potentially other information resulting from treatment of the descriptive content such as special features, diagnosis, similar species. The output may of course also be combined with other types of data from the CDM or from another origin. Descriptive views can be associated with other views at the application level.

To clarify with an example the new attributes created for output in natural language, let us consider anew our previous example (see 1.). Here is the information previous and new text attributes would contain.

Feature	f1
representations (English)	"phloem organization"
featureFirstFormulations (English)	"The phloem in the leaf axis is"
featureNextFormulations (English)	", or"

StateData	s2	s3
representations (English)	"with a single phloem strand and conspicuously sclerotic"	"phloem divided into 2 separate strands"
stateFormulations (English)	"organized in a single phloem strand and conspicuously sclerotic"	"divided into 2 separate strands"
modifiers (English)	"often", false (= tagged as a pre-modifier)	"in the spring", true (= tagged as a post-modifier)

On the basis of the formulations attributes that are customized by the taxonomist through the user interface, adapted natural language can be generated. Here the result would be: "The phloem in the leaf axis is often organized in a single phloem strand and conspicuously sclerotic, or divided into 2 separate strands in the spring."

4. Appendices

More details on the implementation are available at the following Wiki page:

http://dev.e-taxonomy.eu/trac/wiki/StructuredDescriptiveContentOutput