CMDI 1.2 Taskforce Vocabularies

21.02.2014

Oddrun Pauline Ohren
CLARINO
National Library of Norway

The issue

- Utilising external vocabularies as value domains for CMDI elements and CMDI attributes
- More specifically:
  - How to do this by using the CLAVAS vocabulary service
  - Necessary provisions in the CMDI 1.2 model
  - Support by tools
- Working group
  - Twan Goosen, CLARIN ERIC
  - Oddrun Pauline Ohren, CLARINO
  - Thomas Eckart, ASV Leipzig

Value domains in CMDI 1.1: Elements

- Simple value types (string, integer, etc)
  - Attribute @ValueScheme
- Restricted value types
  - Subelement <ValueScheme> with
    - subelement <pattern> : Value domain defined by regular expression
- Controlled vocabularies
  - Subelement <ValueScheme> with
    - Subelement <enumeration>:
      - List of items with
        - » optional @ConceptLink = <URI, link to some concept registry>
        - » optional @AppInfo = Display string of item

CMDI Elements Examples (1)

Using @ValueScheme:

```
<CMD_Element name="AccessRights" ConceptLink="
http://purl.org/dc/terms/accessRights" ValueScheme="string" CardinalityMin="1"
CardinalityMax="1" Documentation="Information about who can access the resource
or an indication of its security status." Multilingual="false"/>
```

Using <ValueScheme> with <pattern>:

```
 <CMD_Element name="DateCreated"
ConceptLink="http://purl.org/dc/terms/created" CardinalityMin="1"
CardinalityMax="1" Documentation="Date of creation of the resource.">
        <ValueScheme>
           <pattern>[0-9]{4}(-[0-9]{2}(-[0-9]{2})?)?</pattern>
        </ValueScheme>
      </CMD_Element>
```

CMDI Elements Examples (2)

Using <ValueScheme> with <enumeration>:

```
<CMD_Element name="Language" ConceptLink="http://purl.org/dc/terms/language" CardinalityMin="0"
CardinalityMax="unbounded" Documentation="A language of the resource.">
    <ValueScheme>
      <enumeration>
        <item ConceptLink="http://cdb.iso.org/lg/CDB-00138580-001" AppInfo="Dutch (nld)">nld</item>
        <item ConceptLink="http://cdb.iso.org/lg/CDB-00138502-001" AppInfo="English (eng)">eng</item>
        <item ConceptLink="http://cdb.iso.org/lg/CDB-00138497-001" AppInfo="German (deu)">deu</item>
        <item ConceptLink="http://cdb.iso.org/lg/CDB-00138512-001" AppInfo="French (fra)">fra</item>
      </enumeration>
    </ValueScheme>
  </CMD_Element>
```

Value domains in CMDI 1.1: CMDI Attributes

- Simple value types (string, integer, etc)
  - Sub-element <Type>
- Restricted value types and controlled vocabularies
  - Excactly like CMDI elements

CMDI Attributes Examples

Using Type:

```
<Attribute>
  <Name>LanguageID</Name>
  <Type>string</Type>
</Attribute>
```

Using ValueScheme:

```
<Attribute>
  <Name>dcterms-type</Name>
  <ValueScheme>
    <enumeration>
      <item ConceptLink="" AppInfo="">DCMIType</item>
      <item ConceptLink="" AppInfo="">DDC</item>
          …
      <item ConceptLink="" AppInfo="">W3CDTF</item>
    </enumeration>
  </ValueScheme>
</Attribute>
```

Current relation between CMDI and concept registries linked to by ConceptLink

- Loose connection
- Linking a CMDI element E to an external concept C in (e.g.) IsoCat means:
  - E is defined by (the definition of) C
  - The value domain of C is not enforced on to E (but often used as guidance)
    - E may be open while C is constrained or closed (IsoCat lingo)
    - E and C may both be closed, but have different sets of values

CLAVAS and OpenSKOS

- OpenSKOS is a vocabulary service
  - Offers publication, management and use of SKOS-ified vocabulary data
  - RESTful API
    - Complies with Apache Lucene Query Parser Syntax
      - http://lucene.apache.org/core/2_9_4/queryparsersyntax.html
- CLAVAS – CLARIN-NL's application of OpenSKOS
  - 3 potentially useful vocabularies added so far
    - **Language codes**
      - ConceptScheme URI: http://openskos.meertens.knaw.nl/iso-639-3
      - OAI set: meertens:ISO-639-3 (>7776)
    - **List of organisations**
      - ConceptScheme URI: http://openskos.meertens.knaw.nl/Organisations
      - OAI set: meertens:VLO-orgs (2543)
    - All public **IsoCat data categories** (only simple datcats?)
      - ConceptScheme URI: many different conceptschemes
        - » corresp. to the closed datcats in which conceptual domains the simple datcats are included
      - OAI set: meertens:isocat (456)
        - » Also defined

Using OpenSKOS API for importing concepts into CMDI

- Harvest entire vocabulary:
  - Example: Language vocabulary in rdf format
    - https://openskos.meertens.knaw.nl/oai-pmh?verb=ListRecords&set=meertens:iso-639-3&metadataPrefix=oai_rdf
- Get part of vocabulary:
  - Example: Subset of language vocabulary: Only language whose prefLabel starts with «Norw»
    - https://openskos.meertens.knaw.nl/api/find-concepts?q=inScheme:http*iso-639-3 AND prefLabel:Norw*&format=rdf
  - Example: Subset of language vocabulary: Only language whose notation starts with «no»
    - https://openskos.meertens.knaw.nl/api/find-concepts?q=inScheme:%22http://openskos.meertens.knaw.nl/iso-639-3%22 AND notation:no*

SKOS

- Simple ontology in RDF for representing concepts and their interrelations
  - Typically used for hierarchical structures like thesauri, taxonomies, etc
- Concept is the main building block
  - May belong to one or more ConceptSchemes (vocabularies)
  - Interrelated by hierachical relations (broader, narrower), associative relations (related) and equivalence relations (use, used-for)
  - Designated by one or more labels (literal strings)
    - prefLabel (one per language)
    - altLabels and hiddenLabels
    - notation (typically used for codes)
  - Annotationlike properties like note, changeNote, skopeNote, definition a.o.
  - Mapping relations (relations to concepts in other schemes)
- Too generic to lend much semantics to entities in specific domains like actors (e.g. persons and organisations), places, etc
- Still, its simplicity, genericity and expansion are obvious strengths

Example (language)

```
<rdf:Description rdf:about="http://openskos.meertens.knaw.nl/iso-639-3/nob">
  <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
  <dcterms:source>
                http://www.sil.org/iso639-3/documentation.asp?id=nob
  </dcterms:source>
  <skos:notation>nob</skos:notation>
  <skos:inScheme rdf:resource="http://openskos.meertens.knaw.nl/iso-639-3"/>
  <skos:exactMatch rdf:resource="http://id.loc.gov/vocabulary/iso639-2/nob"/>
  <skos:exactMatch rdf:resource="http://id.loc.gov/vocabulary/iso639-1/nb"/>
  <skos:prefLabel xml:lang="en">Norwegian Bokmål</skos:prefLabel>
  <skos:broader rdf:resource="http://openskos.meertens.knaw.nl/iso-639-
3/nor"/>
</rdf:Description>
```

- Very few have anything else than prefLabels
- Language code in notation
- Links to other resources

Using OpenSKOS API for dynamic lookup and retrieval

- Autocomplete - with dedicated API OR using find-concepts
  - All prefLabels starting with «Norw»
  - https://openskos.meertens.knaw.nl/api/autocomplete/Norw?returnLabel=prefLabel&searchLabel=prefLabel
    - ["Norwegian","Norwegian Academy of Sciences, Jacob Dybwad","Norwegian Bokm\u00e5l","Norwegian Nynorsk","Norwegian Sign Language","Norwegian University Press","The National Library of Norway","Traveller Norwegian","\u00d8stsamisk museum Neiden, Norway"]
    - Not all are languages!
  - All prefLabels starting with «Norw» in one particular ConceptScheme
  - https://openskos.meertens.knaw.nl/api/find-concepts?q=inScheme:%22http://openskos.meertens.knaw.nl/iso-639-3%22 AND prefLabelAutocomplete:Norw

- Find the concept based on user selection:
  - «Norwegian»
    - https://openskos.meertens.knaw.nl/api/find-concepts?q=prefLabel:Norwegian
  - «Norwegian University Press»
    - https://openskos.meertens.knaw.nl/api/find-concepts?q=prefLabel:%22Norwegian%20University%20Press%22

CMDI 1.2 model: proposed solution

- Accommodates both <CMD_Element> and <Attribute>
- New element <Vocabulary> in <ValueScheme>
  - Parent to <enumeration>
    - If <enumeration> exists, the internal vocabulary (imported or locally specified)
    - If no <enumeration>, then external Vocabulary
  - Attributes for <Vocabulary>
    - @URI, @ValueProperty, @ValueLanguage

Proposed solution – element XML Example

```
<CMD_Element
        name="Language"
        CardinalityMax="1"
        CardinalityMin="1">
        <ValueScheme>
           <Vocabulary
              URI="http://openskos.org/api/languages"
              ValueProperty="skos:prefLabel"
              ValueLanguage="en">
              <enumeration>
                 <item ConceptLink="http://cdb.iso.org/lg/CDB-00138580-
001">Dutch</item>
                 <item ConceptLink="http://cdb.iso.org/lg/CDB-00138512-
001">French</item>
              </enumeration>
           </Vocabulary>
        </ValueScheme>
     </CMD_Element>
```

Proposed solution – changes in GCS

```
<xs:complexType name="ValueScheme_type">
    <xs:choice>
      <xs:element name="pattern" type="xs:string" maxOccurs="1">
        <xs:annotation>
          <xs:documentation>Specification of a regular expression the element should
            comply with.</xs:documentation>
        </xs:annotation>
      </xs:element>
      <xs:element name="Vocabulary" type="Vocabulary_type">
        <xs:annotation>
          <xs:documentation>Specification of an open or closed vocabulary</xs:documentation>
        </xs:annotation>
      </xs:element>
    </xs:choice>
  </xs:complexType>

  <xs:complexType name="Vocabulary_type">
     <xs:sequence>
      <xs:element name="enumeration" type="enumeration_type" minOccurs="0" maxOccurs="1">
        <xs:annotation>
          <xs:documentation>A list of the allowed values of a controlled
            vocabulary.</xs:documentation>
        </xs:annotation>
      </xs:element>
    </xs:sequence>
    <xs:attribute name="URI" type="xs:anyURI"/>
    <xs:attribute name="ValueProperty" type="xs:string"/> <!-- optionally selects a label -->
    <xs:attribute name="ValueLanguage" type="xs:string"/> <!-- optionally selects a language -->
  </xs:complexType>
```

Impact on centres

- Tools
  - Metadata editors (Arbil)
    - Facilitate external vocabulary lookup and autocompletion
  - Component Registry
    - Facilitate import of vocabularies and parts thereof
  - Discovery services?
    - Assistance for users through vocabularies
    - Probably a bit further down the line….
- Existing components
  - All components containing elements and attributes with closed value domains (enumeration) must be updated

Discussions

- Partial vocabulary import hampers component reuse?
- Other issues?