Metadata Tools Supporting Controlled Vocabulary Services

Daan Broeder, Freddy Offenga, Don Willems

Max-Planck Institute for Psycholinguistics
daan.broeder@mpi.nl

Abstract

Within the ISLE Metadata Initiative (IMDI) project a user-friendly editor to enter metadata descriptions and a browser operating on the linked metadata descriptions were developed. Both tools support the usage of Controlled Vocabulary (CV) repositories by means of the specification of an URL where the formal CV definition data is available.

1. Introduction

The use of metadata to describe available resources is an accepted way of making those resources locatable and accessible. This practice is no longer exclusive for the librarians domain who introduced it on a grand scale but now finds also entry in a variety of other domains such as the linguistic domain. Here recently two initiatives [1,2] were started to develop applicable metadata vocabularies for language resources. One essential part of a metadata vocabulary is to define the constraints of the different element values. Often such a constraint implies a choice from a set of permitted values and in that case we speak of the Controlled Vocabulary for the metadata element.

A basic requirement to achieve a high degree of uniformity, to allow the user to adapt his behavior over time and to facilitate searching is the support of such controlled vocabularies during metadata input and during search and browsing. It is obvious that the usage of controlled vocabularies also reduces the amount of encoding errors which increases the usability of the resulting metadata.

The issues regarding these CV's and their use by metadata tools is the subject of this article.

2. Metadata Vocabulary Interoperability

The availability of metadata on the Internet links together the producers and consumers of resources and creates "universal" metadata search spaces. The use of different metadata vocabularies can be considered as partitioning this universal space into a number of sub-spaces. Interoperability between such sub-spaces can only be achieved by carrying out mappings between the elements of the different MD vocabularies. Of course such mappings are lossy, i.e. relevant information will be lost or not carried over fully semantically correct. One key for the success in such mappings is the usage of well-defined CV's for associating values with certain metadata elements. Of course it makes sense in various respects to use the same codes for elements of the different sets that share the same semantics such as for languages in all the different sets. The existence of various language encoding systems such as provided by ISO [3] and SIL [4], however, indicate that even with respect to controlled vocabularies mapping schemes have to be applied.

3. CV taxonomy

In IMDI we distinguish between the use of CV's in several ways. First there is the CV where the metadata element may have as value one of the elements from the CV. Secondly a CV can be used as a CV list where the metadata element can have as value one or more elements from the CV. Both of these types may be used either as a mandatory rule in which case we call them closed CV's or as a "strong" advise and then we call them open CV's.

	Open	Closed
Single	Open CV	Closed CV
List	Open CV List	Closed CV List

The reason for introducing the open CV's is that at the moment the metadata concept is relatively new for our domain and we can not expect that the proposed CV's will be acceptable for all groups. The open CV allows research groups and individual users to provide their own values for some elements. We hope that after some time agreement can be reached about new entries for the CV's so that currently open CV's may become closed. This does demand a central authority that "harvests" the metadatadata descriptions and makes an inventory of the use of the open CV's. We do however seriously consider that open vocabularies will always be needed for a number of IMDI elements.

It may be clear that the use of the CV's is independend of the CV definition itself. So a CV may be used in one context as a closed single CV and as an open CV list in another.

Another mechanism used to distinguish between different CV's is the use of namespaces. IMDI supports the use of a namespace prefix when specifying a language id. It is possible to refer to well known CV's for language identification as ISO and Ethnologue. For instance:

ISO639-2:ger   German as specified by ISO639-2
RFC1766:en-US   US English specified by RFC1766
RFC1766:x-sil-dut Dutch as specified in the Ethnologue list.

At the moment Language identifiers are the only IMDI elements where this mechanism is used. The same namespace mechanism is also used by OLAC.

4. Implementation

4.1. Infrastructure

The CV's are available from central servers via the standard HTTP protocol this provides unique identifiers for the CV's and avoids problems with firewalls.

Having the CV definitions available via the Internet allows geographically separate research groups to share the CV's. For efficiency reasons and to support work

situations without Internet access all tools use a caching mechanism where CV's can be stored once and refered to when needed.

The important vocabularies that are defined by the IMDI standard should all be available from a central repository server and these definitions should be well maintained by a central authority. However IMDI tools should also be configurable in such a way that a user can link the free definable key/value pairs that are available at several levels within the IMDI session descriptions to specific project bound CV definitions or to provide subsets of the normal IMDI CV's such as for instance a small list of language names. These CV's can be available on local servers or on the local file system.

schema itself, loosing flexibility. A disadvantage of the IMDI method is however that we loose the possibility of having the XML parser check the validity of the values of metadata elements. However as stated above vocabularies are often not fixed so that the connection between metadata element and vocabulary can not be defined in the IMDI schema but only in an instantiation of that schema, this is not considered a big disadvantage.

This means that an XML parser on its own cannot determine if a specific value is permitted for a certain element. A separate validation process should check this. The IMDI Browser and Editor tools for instance provide such validation.
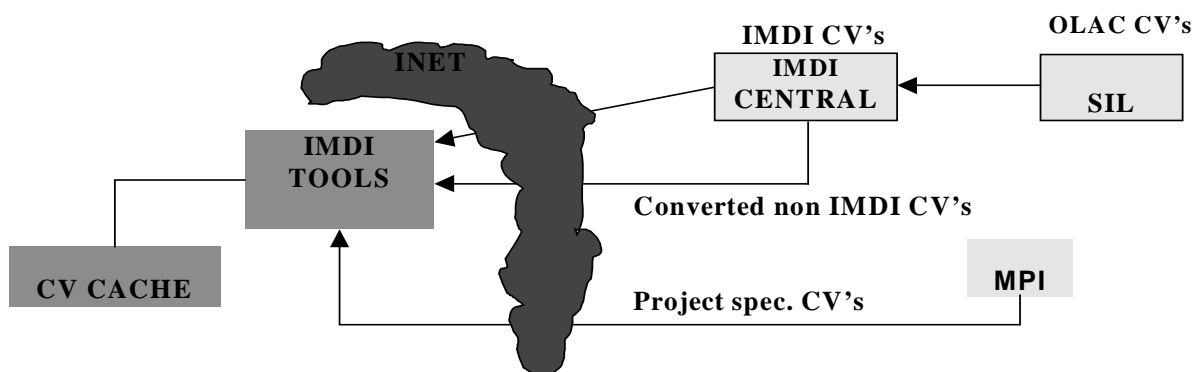

Figure 1. IMDI tools using CV's from different sources

Some vocabularies are more universal than just the IMDI world. Some of these pertain to the linguistic domain such as for instance the SIL language list, others like a reliable list of countries has much wider application. If such lists are available on the Internet it will probably not be in the IMDI format (although we will try to convince where possible) For accessing vocabulary servers that offer vocabularies in non-IMDI formats a bridge could be created in the form of an XSL converter on the central IMDI site, see figure 1 where a possible OLAC CV service is converted into an IMDI CV service.

4.2. Schema Implementation Issues

Although both the IMDI metadata standard and the format of a CV definition are defined as XML Schemas, the actual CV is not a Schema but an instantiation of a schema. From a metadata description (an instantiation of the IMDI Schema) there is a link to the Schema definition.

For instance for the "Continent" element:
```
<Continent
    Type="ClosedVocabulary"
    Link="http://www.mpi.nl/IMDI/Schema/Continents.xml">
    Europe
</Continent>
```

The "Continent" CV schema can be found in Appendix A.

Another possibility would be to define every vocabulary with its own XML-Schema but then we would be obliged to define all mappings between metadata elements and the corresponding vocabularies in the IMDI

4.3. Vocabulary Structure

We have chosen for a definition of vocabularies in the form of an XML file that is an instantiation of the XML-Schema shown in appendix A. A structural model is shown in Figure 2.
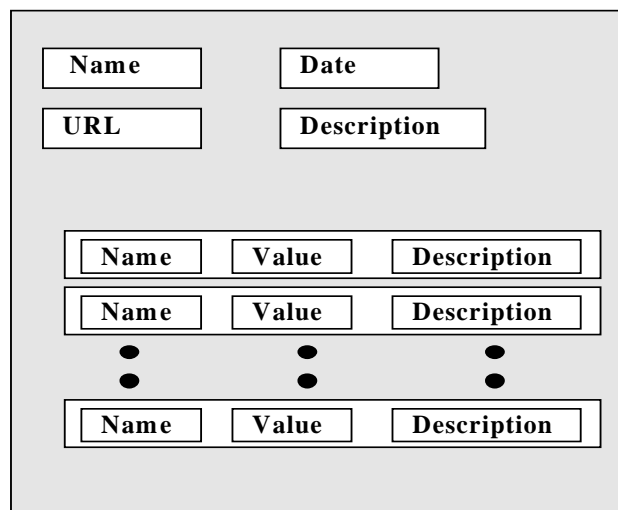

Figure 2. CV definition structure

Every CV definition has a "Name", "Description", "Date", and "URL". The "Date" and "URL" are necessary so that a tool may renew a CV if a new definition has become available. The CV elements are present as a number of elements with each a "Name", "Value" and

"Description" attribute. The "Name" attribute if present is a short form of "Value". Often "Name" will not be specified at all because the "Value" is in itself short and clear enough. The necessity of having both a "Name" and "Value" for a CV item becomes clear if we look at the case where we have a CV of CV's. The values of such a CV would be the different URL's of the constituent CV's clearly much to long and unwieldy to handle.

5. Tools

The IMDI project resulted in a number of tools for metadata exploitation:

    a)   The IMDI-BCBrowser. A viewer for the IMDI metadata descriptions that allows navigating the universe of linked IMDI metadata descriptions.
    b)   The IMDI-BCEditor that is used to create IMDI metadata descriptions.
    c)   The IMDI-BCSearchTool that allows the user to specify a query for specific resources in the IMDI universe.

All tools use a caching mechanism to store CV definitions that are copied from the Internet. The copy serves to avoid constant copying and to enable the tools to work in field conditions where no Internet access is possible.

At the moment the user himself is responsible to refresh the cached CV definitions when new ones have become available. It would of course be possible to automate this but it is perhaps not always desirable that an existing CV definition whereupon already a large part of a project is based suddenly is replaced.
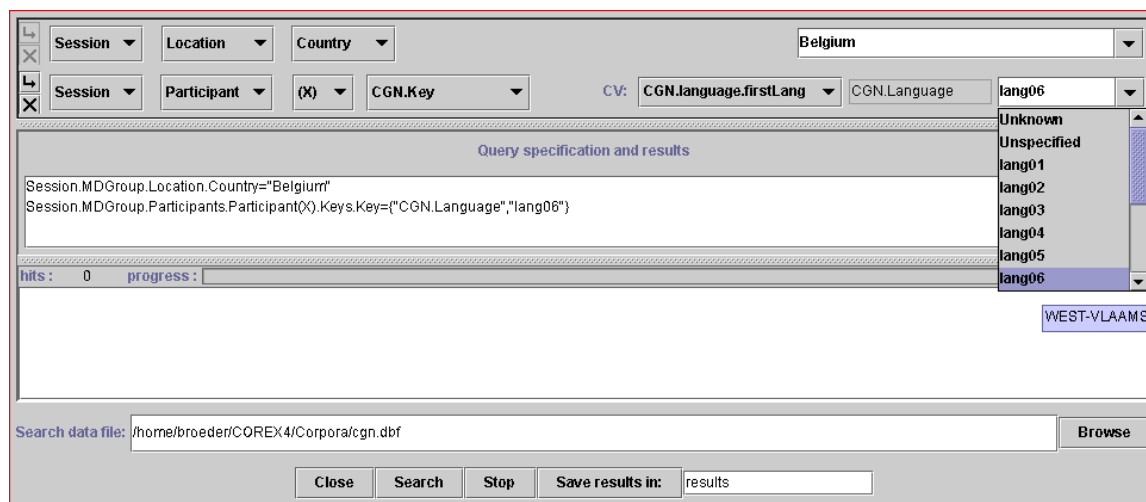
5.1. The IMDI-BCSearchTool

Figure 3. The IMDI-BCSearchTool

The IMDI-BCSearchTool is a tool meant to query repositories of IMDI tagged resources. The UI allows specifying queries following the IMDI vocabulary and structure. CV's are supported in two ways: The standard IMDI CV's are available as choices in a pull-down menu

for the IMDI metadata elements. See the country choice "Belgium" for the standard IMDI element "Session.Location.Country" in figure 3.
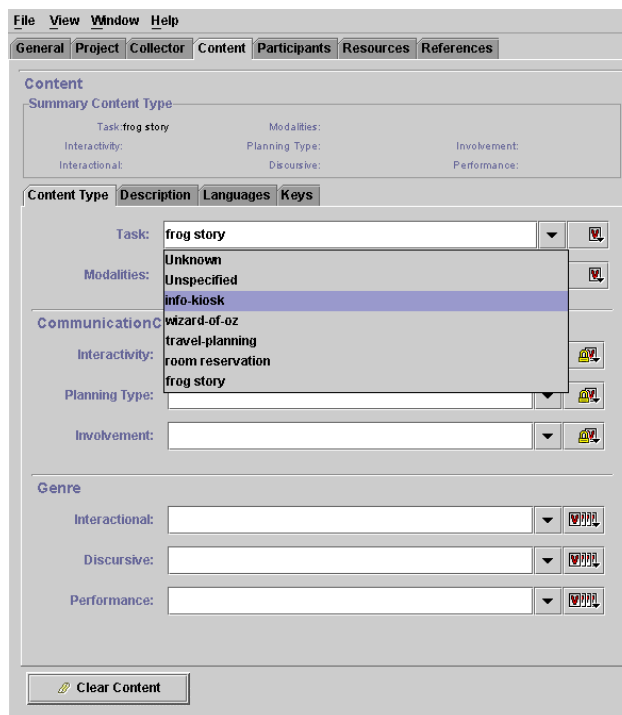

Figure 4. Part of the IMDI-BCEditor UI

Project specific key/value pairs are supported as CV's. For instance the Dutch Spoken Corpus Project specifies the mother tongue of a speaker with a special key at the "Session.Participant" level named: "CGN.Language.firstLanguage". The user may specify this project dependent key (the SearchTool can be configured as such) and automatically the value field becomes a pull-down menu with all possible choices for this field. Notice that the sometimes unintelligible values, in this case lang01,…lang06 become elucidated by a tool-tip window showing a more general definition in this case "WEST-VLAAMS" for lang06. These project specific cases are supported by having CV's of other CV's. The reader is justified asking why the more general definition is not available as a menu choice but in general the

general definitions are much larger then the project specific value and this would make the UI problematic.

5.2. The IMDI-BCEditor

This editor presents all the IMDI metadata elements in a structured GUI to the user. It supports the use of Controlled Vocabularies and user definable keyword/value pairs that the IMDI set allows for user or project specific purposes. In the user interface the CV's are offered as pull-down menus. Also it enforces constraints on the values for some metadata elements where applicable and practical.

5.3. The IMDI-BCBrowser

The IMDI BCBrowser is the central tool for exploiting the IMDI infrastructure. It allows navigation of the universe of linked IMDI metadata descriptions by clicking on corpus links. The browser keeps track of its position in a browsable corpus structure and shows the metadata and human readable descriptions associated with the subcorpus in focus.
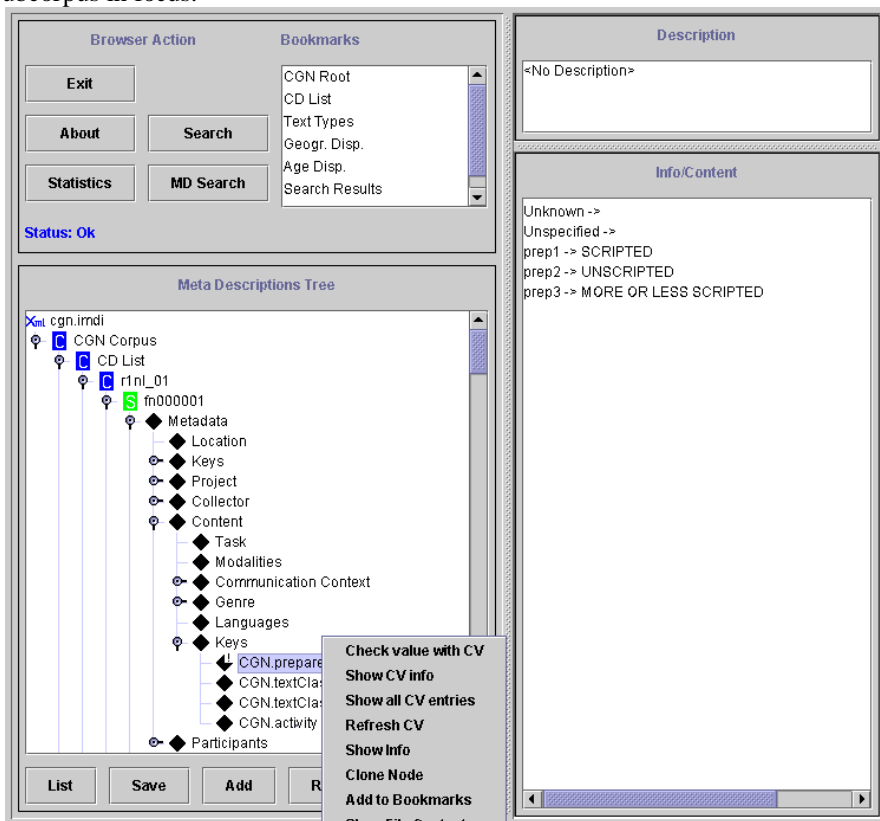

Figure 5. The IMDI-BCBrowser

The Browser supports CV's in a number of ways: It shows metadata items whose values are constrained by a CV in a special way, it checks if the value complies with the CV, and it shows descriptive information about the CV and the CV elements. See the popup menu in Figure 5 for a list of options.

6. Future Developments

We have sketched a scheme, where groups and individual users can define their own metadata elements and accompanying CV's in the form of key/value pairs. The question arises how can we keep the metadata space created by the IMDI set unified in the sense that all metadata descriptions conforming to its standard are searchable by the tools built for it. In essence this problem is the same as the one described under 2 where the mapping problems between different metadata sets is described. Its solution may be easier though because the use of different key/value pairs is embedded in the same IMDI metadata environment.

One way would be that a specific key/value pair that is not part of the IMDI core standard is ignored by all that are unaware of it. Only if a tool is made aware of a (locally defined) CV or key/value pair does it appear within its view and can be used to formulate search queries. This would only be acceptable if it would concern very specialist and relatively rarely used values.

Another way would require more of the key/value pair definition in such a way that the combination of key/value pairs would imply also the settings of one or more standard elements in the IMDI core set. This would perhaps not convey the precise meaning of the specific key/value pair (that is needed by the group that defined it) but it would make some of the semantics available to the greater community.

Using classical means we can only do this when the CV's that are used are accompanied by detailed information that allows the elements of a CV to be interpreted in the dimensions of other elements. Clearly a labor-intensive task especially for a small project that just needs to use one or two special keys.

Another possibility would be to leave such mapping questions to a central authority. This authority would regularly scan the IMDI-universe for metadata descriptions using new values in open CV's and non-standard key/value pairs. Often used new entries would be eligible for promotion to a set of "known" IMDI extensions that would be described together with "imprecise" mapping information to the IMDI core set and store them in a central repository. Ideally a mechanism should be available to process these key/value pairs and mapping information and automatically generate the values for the core IMDI elements. This would need further extension of the current CV format.

This authority could also take a more active role when for instance it would notice that key/value pair definitions were used that are not needed because their semantics can be captured in the core IMDI set. In such a case it could contact the metadata providers and try to convince them to use the regular elements.

Providing "imprecise" mapping information between different key/value pair sets and/or the core IMDI set could be easier when using techniques that are being developed within the framework of the Semantic Web [5]. There specification languages such as RDF [6] are being developed that allow concepts to be build on other concepts in such a way that automatic agents would be

capable of creating such vague mappings. This mechanism presupposes of course that the start of a reasonably developed ontology for the domain is already in place and can be referred to. Also these techniques can be used to link to data categories and terminology repositories outside the linguistic domain.

7. References

[1] ISLE Metadata Initiative http://www.mpi.nl/ISLE & Broeder, D.G., Brugman, H.,Russel, A., and Wittenburg, P., (2000), A browsable Corpus: accessing linguistic resources the easy way. In Proceedings LREC 2000 Workshop Athens.

[2] OLAC Open Language Archives Community http://www.language-archives.org/

[3] [ISO639-2]
Codes for the representation of names of languages - part 2: alpha-3 code, International Organization for Standardization (ISO), 1998. http://lcweb.loc.gov/standards/iso639-2/langhome.html

[4] Ethnologue language name index http://www.sil.org/ethnologue/names/

[5] Semantic Web: http://www.semanticweb.org

[6] RDF: http//www.w3.org/RDF

8. Appendix A. The CV Schema

This schema is part of the IMDI Schema that can be found at: http://www.mpi.nl/IMDI/Schema/IMDI.xsd

```
<xsd:complexType name="VocabularyDefType">
  <xsd:annotation>
    <xsd:documentation>
    The definition of a vocabulary. Attributes: Date of
    creattion, Link to origin. Contaills a Description be
    element to describe the domain of the vocabularyand
    a (unspecified) number of value enries
    </xsd:documentation>
  </xsd:annotation>
  <xsd:sequence>
    <xsd:element ref="imdi:Description"/>
    <xsd:element name="Entry"
maxOccurs="unbounded">
      <xsd:complexType>
        <xsd:simpleContent>
          <xsd:extension base="xsd:string">
            <xsd:attribute name="Name"
type="xsd:string"/>
            <xsd:attribute name="Value"
type="xsd:string"/>
          </xsd:extension>
        </xsd:simpleContent>
      </xsd:complexType>
    </xsd:element>
  </xsd:sequence>
  <xsd:attribute name="Name" type="xsd:string"
use="required"/>
  <xsd:attribute name="Date" type="xsd:date"
use="required"/>
  <xsd:attribute name="Origin" type="xsd:urlRef"
use="required"/>
</xsd:complexType>
```

an example CV is the (Linguistic) Continent CV

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-- edited with XML Spy v3.5 NT (http://www.xmlspy.com) by
Daan Broeder (Max-Planck Institute for Psycholinguistics) -->
<imdi:VocabularyDef
xmlns:imdi="http://www.mpi.nl/IMDI/Schema/IMDI.xsd"
xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance"
xsi:schemaLocation="http://www.mpi.nl/IMDI/Schema/IMDI.xs
d ./IMDI.xsd" Name="Continents" Date="2001-05-06"
Origin="httpts://www.mpi.nl/IMDI/Schema/Continents.xml">
  <imdi:Description>
    <Text Language="ISO639-2:eng ">List of linguistic
continents </Text>
    <Text Language=" ISO639-2:eng"
link="http://www.mpi.nl/IMDI/Documents/Continents.html" />
  </imdi:Description>
  <Entry Value="Africa"/>
  <Entry Value="Asia"/>
  <Entry Value="America-North"> Not a real continent
</Entry>
  <Entry Value="America-Middle">Not a real continent
</Entry>
  <Entry Value="America-South">Not a real continent
</Entry>
  <Entry Value="Europe"/>
  <Entry Value="Australia"/>
</imdi:VocabularyDef>
```