ISLE Metadata Initiative (IMDI)

PART 3 A

Vocabulary Taxonomy and Structure

Draft Proposal Version 1.1

December, 2001

IMDI[1] Technical Report

Max-Planck-Institute for Psycholinguistics
NL, Nijmegen

---

[1] For information about the ISLE Metadata Initiative, please, look at the following web-site: www.mpi.nl/ISLE

INDEX

1   Introduction

The term vocabulary as we use it in the IMDI documentation should not be confused with the term "metadata vocabulary". The last term refers to the total set of metadata elements defined for a specific domain or application. Whenever we use that concept we will write "metadata vocabulary" in full. With "vocabulary" we mean the set of string values that can be used to assign a value to a specific metadata element or attribute of an element.

1.1   Taxonomy of vocabularies

We distinguish four classes of controlled vocabularies:

- *Closed Controlled vocabulary*: The value of the metadata element is one and only one element from a finite set of values.
- *Closed Controlled vocabulary list:* The value of the metadata element is a list with a number of elements from a finite set of values.
- *Open Controlled vocabulary*: The value of the metadata element is one and only one element from a finite set of elements **or** is a user specified string.
- *Open Controlled vocabulary list*: The value of the metadata element is a list with a number of elements from a finite set of values **or/and** a number of user specified strings

The "Open CV" and "Open CV list" are there to advise users and not constrain them. This is a requirement for some metadata elements where some users were quite specific about. It is a matter of taste if we allow a metadata element to have a CV list as a value or allow the metadata element to be repeated. We chose the first approach for compactness sake.

If a vocabulary is a list, open or closed is dependent on the context, application or metadata schema itself, it is not something that is determined by the vocabulary itself. It is very well conceivable that a specific vocabulary is a closed CV for one metadata element and an open CV list for another. So we will specify in the IMDI XML-Schema what type of vocabulary a metadata element is connected to when applicable for the element.

2   Required functionality

For the IMDI metadata set "we have and still are refining" a number of vocabularies that need to be flexible and dynamic. That means that the definition of the vocabulary should be defined in a separate file and can be stored on a central vocabulary server. We have already developed a tentative XML schema for such a vocabulary definition and as a transmission protocol HTTP seems sufficient. Efficiency is important because sometimes "large" vocabularies have to be transferred (think of the set of "language identifiers"). Tools using these vocabularies would also need a caching mechanism for speed and offline work.

Vocabularies are downloadable. After they have been downloaded an application should be able to check if the vocabulary is still up to date. Therefore a vocabulary definition needs elements defining the creation date and an URL link to its origin.

2.1   Infrastructure

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. These CV's can be available on local servers or on the local file system.
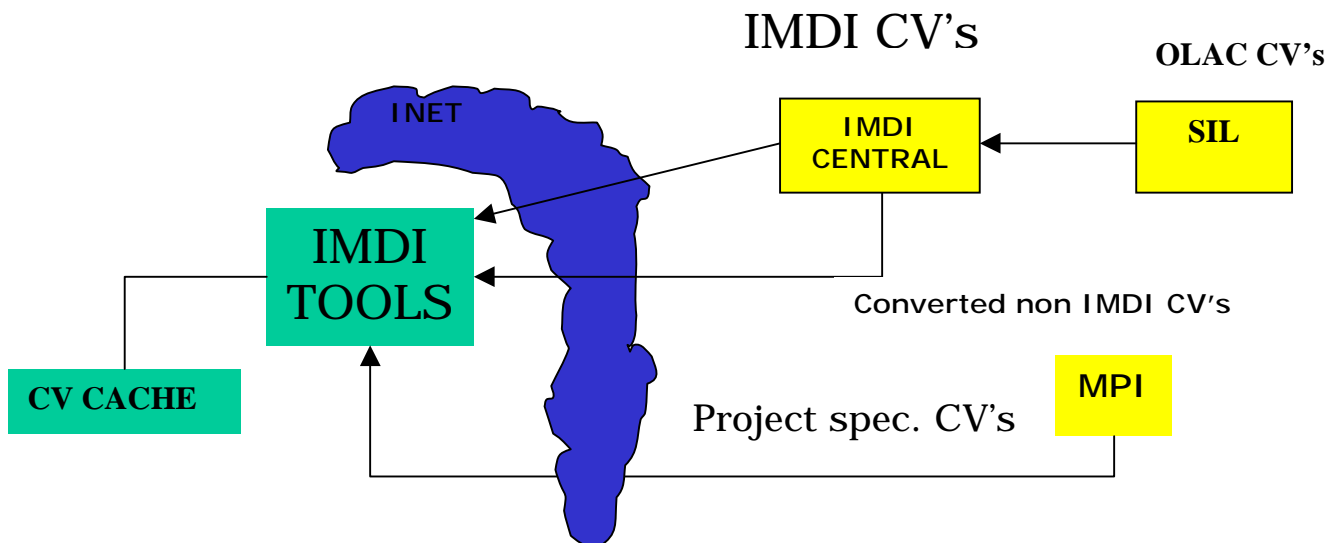

Figure 1 IMDI tools using CV's from different sources

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).

2.2   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 here:

```
<xsd:complexType name="VocabularyDefType">
   <xsd:annotation>
       <xsd:documentation>
   The definition of a vocabulary. Attributes: Date of creattion, Link to
   origin. Contails 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" Type maxOccurs="unbounded"/>
        <xsd:element name="Entry" maxOccurs="unbounded">
           <xsd:complexType>
               <xsd:simpleContent>
                   <xsd:extension base="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="Link" type="xsd:urlRef" use="required"/>
</xsd:complexType>
```

Another possibility is to define every vocabulary with its own XML-Schema but then we would obliged to define all mappings between metadata elements and the corresponding vocabularies in the IMDI schema itself, loosing flexibility. As a disadvantage we loose the possibility of having the XML parser check the validity of fixed mappings between metadata elements and corresponding vocabulary. 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.

2.3   Implementation of CV's in IMDI tools

At the moment all IMDI Tools, the IMDI-BCBrowser, IMDI-BCEditor and the IMDI-BCSearch tools implement the use of CV's although not in a completely user configurable fashion

Appendix A

Example of a continent vocabulary

```
<?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.xsd
D:\users\BROEDER\Documents\DOC\LAPP\ISLE\IMDI.xsd" Name="Continents" Date="2001-
05-06" Origin="httpts://www.mpi.nl/IMDI/Schema/Continents.xml">
    <imdi:Description>
        <Text Language="SIL:xxx">List of linguistic continents </Text>
        <Text Language="SIL:xxx"
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-South">Not a real continent </Entry>
    <Entry Value="Europe"/>
    <Entry Value="Australia"/>
</imdi:VocabularyDef>
```