Metadata Profile in the ISO Data Category Registry

Freddy Offenga, Daan Broeder, Peter Wittenburg, Julien Ducret, Laurent Romary

MPI for Psycholinguistic, LORIA
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
{freddy.offenga, daan.broeder, peter.wittenburg}@mpi.nl, {julien.ducret, laurent.romary}@loria.fr

Abstract

Metadata descriptions of language resources become an increasing necessity since the shear amount of language resources is increasing rapidly and especially since we are now creating infrastuctures to access these resources via the web through integrated domains of language resource archives. Yet, the metadata frameworks offered for the domain of language resources (IMDI and OLAC), although mature, are not as widely accepted as necessary. The lack of confidence in the stability and persistence of the concepts and formats introduced by these metadata sets seems to be one argument for people to not invest the time needed for metadata creation. The introduction of these concepts into an ISO standardization process may convince contributors to make use of the terminology. The availability of the ISO Data Category Registry that includes a metadata profile will also offer the opportunity for researchers to construct their own metadata set tailored to the needs of the project at hand, but nevertheless supporting interoperability.

1. Introduction

It is widely agreed that metadata sets such as proposed by IMDI [1] and OLAC [2] will play an increasingly crucial role in managing and allowing resource discovery for the growing amount of language resources that are created world wide. This is in particular true in the era of integrated services which we will enter during the coming years. Although together they can claim that about 85 institutions world-wide make use of the terminology they propose, they still miss a more complete acceptance. When speaking about the reasons for this a few major points can be mentioned: (1) The creation of useful metadata itself, is sometimes regarded reluctantly since it is a labour intensive process while it is information for others. (2) Also the creation of metadata is often foreseen in the original project proposals and so not included in the budget. (3) Resource creation projects usually suffer from high time pressure and metadata creation is often the last point on the lists. (4) There are sometimes doubts about the suitability of the available metadata sets and their long-term usability and survival. (5) Projects can not afford to sustain offering their (meta-) data and are not able to maintain their metadata and packaging to ensure interoperability.

To remove the objections relating to point four, we want to take more steps to improve the stability and the long-term survival of the concepts proposed by IMDI and OLAC. Problems and obstacles encountered with this work vary according to the nature of the metadata sets involved. Since OLAC is widely based on Dublin Core (DC) [3] it seems that only the additional concepts have to be anchored in an open domain-specific registry. The IMDI set, using its own linguistic domain based terminology and being in contrast to DC a structured set, requires more work.

The ISO Data Category Registry [4,5,6] offers the opportunity to register all metadata concepts that have proven their usefulness for the language resource domain during the last years. It will ensure that the investments that have to be done for metadata creation and management will not be lost after some years. This may motivate projects and initiatives to join the existing metadata domains.

Finally, the ISO Data Category Registry will allow every project in future to define its own metadata set, but remaining interoperable at the semantic level if registered concepts are re-used.

2. ISO 12620 Data Category Registry

The Data Category Registry currently being established within ISO TC37/SC4 has the potential to revolutionize the way in which we will use linguistic concepts and achieve interoperability at the level of linguistic encoding. The ISO DCR basically consists of a flat list of linguistic concepts covering a range of linguistic domains. It is organized in profiles covering concepts such as metadata, morpho-syntax, syntax etc. Each concept consists of a proper definition and may have a conceptual domain. Simple concepts that occur as values in the conceptual domain are atomic and don’t have a value range of their own. Each concept has language sections that specify the value ranges in the different languages. Where applicable the definition of concepts can refer to a broader generic concept if this helps clarification. New concepts that a linguist may find essential may be added to the DCR. In general, such new concepts will first be part of a private workspace, but they can be subject of an ISO-process that may result in its acceptance as part of the official DCR.

The model underlying the ISO DCR is shown in figure 1. It is basically a flat list of concept that are used within a certain domain – in our case the linguistic domain - and the way they are defined is based on the ISO 11179 [7] and ISO 12620 [8] models already being widely used. The model does not include relations to other concepts such as possible for example using RDF-S [9] or OWL [10] to describe the semantics of the relation. On purpose it restricts itself to what a community is able to agree

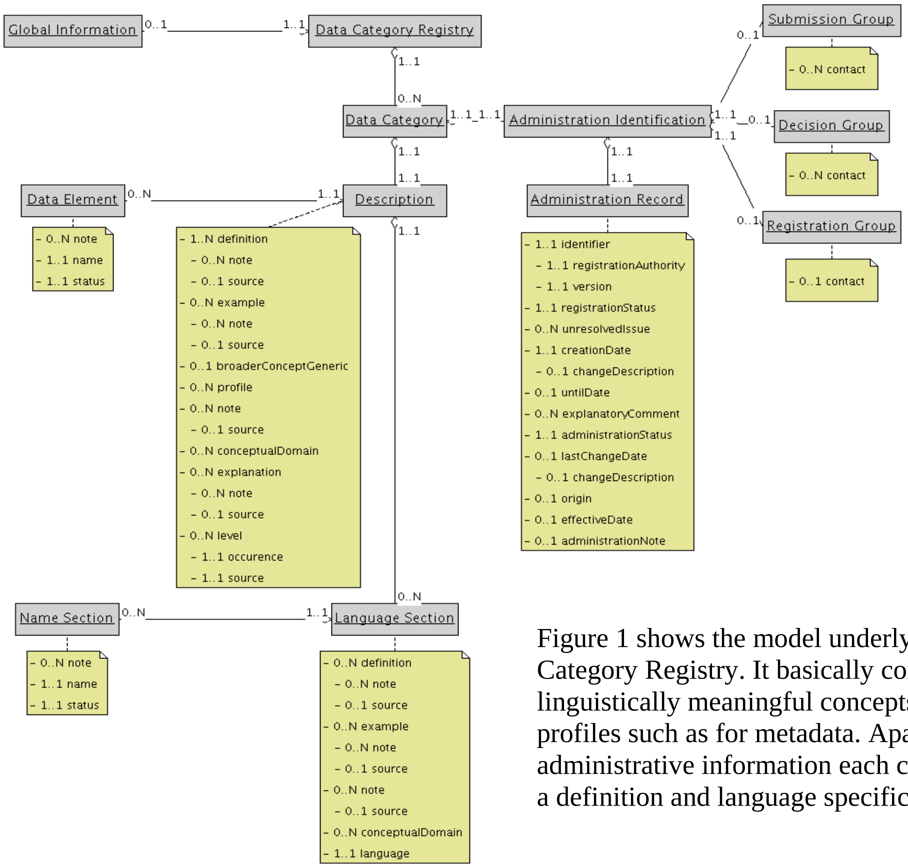

Figure 1 shows the model underlying the ISO Data Category Registry. It basically contains a flat list of linguistically meaningful concepts organized into profiles such as for metadata. Apart from the administrative information each concept consists of a definition and language specific information.

upon. So because it was only needed to express structure and not needing RDF or any existing vocabulary to express semantic relations, XML was chosen as representation language.

The DCR is available as an XML structure that is defined by a RelaxNG schema [11] and is in this way available in machine readable form. This achieves interoperability with tools that are aware of this. These tools can thus make use of existing concepts and incorporate them in schema definitions. Schemas created in this way and that define language resources such as metadata descriptions, annotations, lexica etc should include references that point to the DCR concepts in a unique way. A standard API to interface with the DCR was developed and is available.

3. IMDI Metadata Set

Different metadata sets have been introduced and are used to describe web-documents in general such as by Dublin Core and humanities resources such as by TEI [12]. Since Dublin Core lacks specific information about language resources such as the language the resource is about or the linguistic data type, the OLAC group extended the Dublin Core set by a few concepts. Nevertheless, the OLAC set is designed to locate language resources in large integrated metadata domains. In contrast to this the TEI header set was designed to describe linguistic content in great detail. However, TEI was designed for textual data and was therefore not useful per se for modern multimedia/multimodal resources, it lacked tool support and was too extensive for most projects. IMDI was designed to fill this gap: (1) It is a structured set reflecting the needs of the research community to specify language resources in enough detail for relevant research questions to be answered. (2) IMDI also has structuring capabilities that not only supports individual resource retrieval, but also captures tight relations between resources. (3) IMDI comes along with a comprehensive infrastructure to manage large language resource archives and allows metadata search to be combined with content queries.

The IMDI set is a result of contributions from different sub-domains such as field and corpus linguists, speech and language technologists and researchers working on multimodality. In addition, the IMDI set and infrastructure support project and sub-domain specific extensions called profiles and also individual researchers can make very specific extensions. Such a special profile was for instance designed by members of the sign-language community in Europe, since this community has very specific needs in addition to what the IMDI core set provides. Therefore, IMDI is a rather elaborated set offering fields that may not be relevant for everyone. However, in contrast to Dublin Core, for example, the terminology is tailored towards the needs of the language resource community. In the following overview the IMDI set is briefly characterized.

```
Resource Bundle (Name, Title, Date)
  Location (Continent, Country, Region, ...)
  Project (Name, Title, ID, ...)
  Content (Languages, Modalities, Genre, Task, ...)
  Actors (Type, Name, Role, Age, Sex, ...)
  Resources
    MediaFile (Link, Type, Format, Quality, ...)
    AnnotationUnit (Link, Annotator, Type, ...)
    Source (Format, Quality, ID, ...)
  References
```

Note that to satisfy the researchers needs, the set or packaging has to be structured, since otherwise one cannot distinguish the details between different participants, for example.

4. IMDI Concepts into DCR

To achieve the above mentioned goal with respect to anchoring metadata definitions in wider frameworks, it was one of the goals within the LIRICS project [13] to integrate the IMDI metadata descriptors into the ISO DCR. We will now discuss some of the problems we were confronted with when mapping the concept definitions on the underlying DCR model.

The DCR documents specify the following about a data category: it is “an elementary descriptor to specify and implement a linguistic annotation scheme in the wide sense”, i.e., it is any kind of information attached to a language resource. Translated for metadata this can be interpreted as “any descriptor element used to characterize a resource is a data category”. For Dublin Core there should be no problem since they are, for example, context independent. “DC:Type” which specifies the genre of the resource content is such a data category, for example, although it is only vaguely defined.

Metadata descriptions, in general, have the advantage that its concepts can be much more easily transferred to different languages. The concepts “age” and “date of birth”, for example, are universal except for some cultures where the exactness of the specification is not relevant. In contrast, for example, to morphosyntactic concepts where the value ranges dependent on languages metadata concepts can be more easily mapped to different languages. However, some are very weakly specified or still heavily debated such as /genre/ making it very difficult, if not impossible, to specify a suitable and widely agreed value set. However this is more caused by linguistic subdomain variation than language differences, but it makes us expect that there will be differences in interpretation of the field which will result in a wider semantic scope making resource retrieval more difficult.

4.1. Embedding Problem

In IMDI elements are embedded in structural contexts such as “IMDI:Content.Modalities” which are the modalities included in a recording and/or annotation. What is the data category: “Modalities” or “Content. Modalities”? It is obvious that the term “Modalities” is generic and that “Content. Modalities” may refine the generic meaning. The answer may be simple in this case, since there is just one “modality” element. What, however, in the case that a generic concept is re-used in different contexts: “Session.Name”, “Project.Name”, “Contact.Name” and “Actor.Name”. The embedding in the schema says something about the semantics, nevertheless they share the same generic concept “Name”. We can even refer to more deep embeddings such as “Content.Languages.Language. Language_Name”, “Participants.Participant. Language.Language_Name” and “Content. Description.Language_Name”. The first says something about the language a resource is about, the second something about a language an actor speaks and the third something about the language a description is written in. All three share the concept “Language_Name”.

In the first version of the IMDI mapping to the DCR model it was decided that the context should be part of the definition yielding:

<feat type=”identifier”>Content.Modalities</feat>

However, this led to a proliferation of specific concepts making re-usage impossible. So it was decided to have only the generic concepts in the DCR and to require that the schema defining the embedding takes care of refinements when necessary, i.e., this leads to the following specifications:

<feat type=”identifier”>Modalities</feat>
<feat type=”identifier”>Name</feat>
<feat type=”identifier”>Language_name</feat>

This also means that there is just one concept “Language_name” and that the distinction between the language a resource is in and a language a resource is about is not made at DCR level, but at schema level.

4.2. Different Semantic Scopes

Two other examples were discussed. For IMDI it was suggested by the researchers to include two elements that allow to specify the linguistic genre of a resource. The element “IMDI:Genre” is associated with a controlled vocabulary that includes simple data categories such as “Singing”, “Fiction”, “Narrative” etc. These could be further refined by specifications such as “religious singing”. To allow researchers to add such refinements an element called “IMDI:Subgenre” was introduced. Although “Subgenre” certainly is a refinement of “Genre” it was decided to have two categories, since the value ranges and therefore the semantic scope are different. While “Genre” is associated with an open vocabulary allowing users to add values, “Subgenre” is a free-text field. Experience over the years may show that this differentiation is not necessary.

4.3. Mapping OLAC and IMDI together

Where possible we should try to include data categories in the ISO DCR that are broader generic concepts to concepts that are used in IMDI or other metadata set. By re-using such a concept and when necessary re-fining it at schema level would automatically create interoperability. Both IMDI and OLAC use the concept “Role” and the definition largely overlaps, however, in the value range small differences can be found. The decision was taken that “ISO:Role” is a broader-generic concept for “IMDI:Role” and “OLAC:Role”. Given the same definition the controlled vocabularies have to be integrated. Such a merge, however, does not solve the interoperability problem completely, since many metadata descriptions have already been created using the existing value ranges. While IMDI uses a value called “Collector” OLAC uses a value called “Compiler”, both obviously meaning the same. One name will be included in the definition of “ISO:Role”. Since every value itself is a simple data category which has to be specified according to the same model, the other names could be entered as sub-community specific.

5. IMDI Working Languages

Until now IMDI is available for the following working languages: English, Dutch, German, Swedish, Italian and Greek. As was expected, no special problems for using the language sections in the ISO DCR occurred.

6. Conclusions

The well-known and mature IMDI metadata set has been mapped to the model underlying the ISO Data Category Registry. Some problems occurred while entering the definitions leading to a proliferation of concepts and context-dependent semantics that would make re-usage almost impossible. Since at the end interoperability is the goal to be achieved, this was changed. Context-independent concepts are now used in the DCR and all context-dependency has to be handled at schema level.

By including the major concepts of IMDI into the metadata profile into the DCR we ensure that the definitions can be maintained at a higher level and will become more persistent and stable. This will hopefully motivate projects and initiatives to make use of these concepts. Since the ISO DCR can be contacted via an API a practical mechanism is provided for re-using the concepts. Finally, this may help us to build truly interoperable domains.

7. References

[1] http://www.mpi.nl/IMDI
[2] http://www.language-archives.org/
[3] http://dublincore.org/
[4] http://syntax.inist.fr/
[5] http://www.tc37sc4.org
[6] http://syntax.inist.fr/page/home.RelaxNG.inc.php
[7] http://metadata-standards.org/11179/
[8] http://www.ttt.org/clsframe/datcats.html
[9] http://www.w3.org/RDF/
[10] http://www.w3.org/TR/owl-features/
[11] http://www.relaxng.org
[12] http://www.tei-c.org/
[13] http://lirics.loria.fr/