A Data Category Registry- and Component-based Metadata Framework

Daan Broeder, Marc Kemps-Snijders, Dieter Van Uytvanck, Menzo Windhouwer, Peter Withers, Peter Wittenburg, Claus Zinn

Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD, Nijmegen, The Netherlands
{firstname.lastname}@mpi.nl

Abstract

We describe our computer-supported framework to overcome the rule of metadata schism. It combines the use of controlled vocabularies, managed by a data category registry, with a component-based approach, where the categories can be combined to yield complex metadata structures. A metadata scheme devised in this way will thus be grounded in its use of categories. Schema designers will profit from existing prefabricated larger building blocks, motivating re-use at a larger scale. The common base of any two metadata schemes within this framework will solve, at least to a good extent, the semantic interoperability problem, and consequently, further promote systematic use of metadata for existing resources and tools to be shared.

1. Introduction

Metadata plays a crucial role in realizing the eScience vision. Without a proper description of research data, there is little hope to build environments that help researchers in finding, accessing, exploiting, and preserving the data they need to effectively conduct their studies.

The main lesson learned is that there is no single metadata scheme that satisfies the needs of all researchers of all disciplines, with their many different types of resources but also domain-specific vocabulary and research methods. On the other hand, it is unwise to create new metadata schemes, just because existing ones were ignored, or did not perfectly fit a community's requirements. This, in fact, lead to the many schemas existing today, so that the metadata universe can be described as rather fragmented, causing the interoperability problems that so many of us suffer.

This paper gives an up-to-date account of our approach for metadata management, extending (Broeder et al., 2008). It is based on a controlled vocabulary, centrally managed in a data category registry, to provide the basic building blocks, and a component registry for constructing larger components from existing ones, and for sharing them. The CLARIN research infrastructure (Váradi et al., 2008) will provide those registries, together with tool support, and aims at describing existing language resources within this new framework to facilitate their access, re-use, and interoperability.

2. Existing Metadata Solutions

The Library Sciences advocates the Dublin Core Metadata Element Set (DCMES), with its 15 metadata fields, to describe all kinds of objects. However, the strong emphasis on library-specific expert terminology hindered its wide acceptance in other disciplines. Also, the DCMES is an unstructured set of descriptors, which some felt make it unsuitable for the description of complex objects.

In the linguistic domain, metadata started being present in headers of annotation files such as CHAT (MacWhinney, 2000). In these first steps, however, the encoding and semantics of the metadata were all corpus specific and no attempts were made to cover a wider field of resources. The Text Encoding Initiative (TEI)[1] initiative has been successful in establishing a widely accepted system for text annotation, at least within the Humantities where a large amount of TEI resources exists. It also includes metadata fields to describe types of text resources, the TEI header.

The systematic use of DCMES for linguistic resources resulted in OLAC (Simons and Bird, 2003), which adds an (extensible) set of descriptors to the DCMES. Although OLAC started with the addition of only a single metadata element ("language identifier"), others followed over the years: codes for discourse type, participant roles, the linguistic field and data type were introduced. Now OLAC is the accepted metadata exchange format between language resource archives.

The IMDI metadata scheme (Broeder and Wittenburg, 2006) aims at addressing the need to describe resources in the linguistic domain, suitable for multimedia resources, using a domain specific terminology, and supporting complex resources. Although IMDI can be used to describe text corpora and lexica, its main strength and its primary use is the detailed description of bundles of tightly related resources of multimedia corpora (for instance, audio and video material stemming from DoBeS endangered language documentation projects). Metadata schemes, or *profiles*, were created as community-specific extensions such as for the Spoken Dutch Corpus and the Sign Language community.

Clearly, there is a tension between the need for sufficiently rich and domain-specific terminology to adequately describe resources and the desire for interoperability where terms have to be understood by humans (from varying disciplines) and machines alike. This has lead to a kind of rebound effect for metadata description sets: it started with a move from small sets of descriptors with broad semantics to large sets with highly specific descriptors, which was then followed by a backward move; Baker compared this to the linguistic theory of pidginization and creolization (Baker, 1998).

[1] http://www.tei-c.org

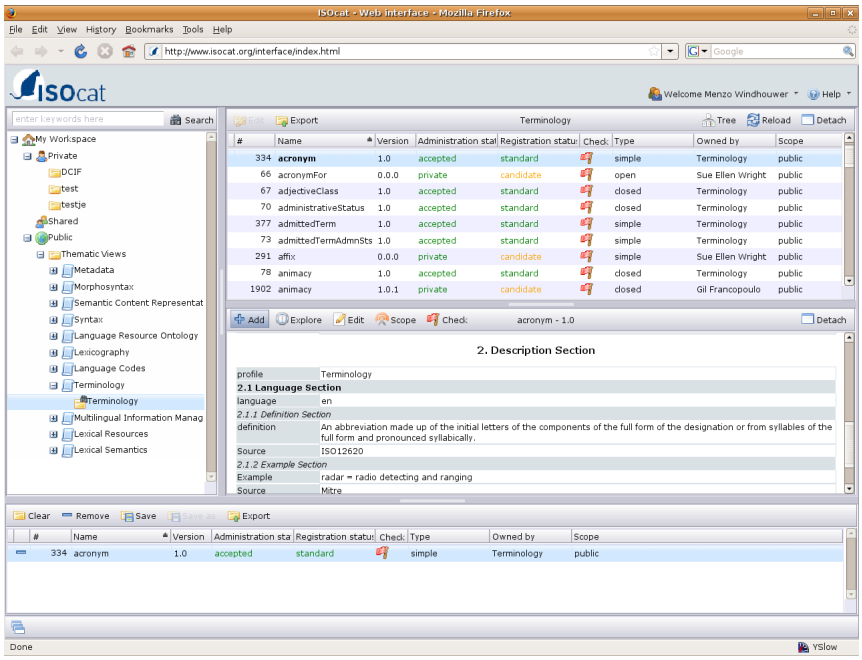


Figure 1: ISOcat reference implementation

Given the various metadata schemes, interoperability is facilitated by creating a semantic mapping from one to another. While the alignment path is certainly one option to address the interoperability issue, it rather attacks the symptons than the cause. Having a commonly agreed base set of controlled vocabulary, together with a component-based approach to build complex metadata blocks from simpler ones, is a better path towards achieving a common understanding and consensus across the various communities, and interoperability between resources and tools.

3. The ISO Data Category Registry

The main focus of the ISO group ISO TC37/SC4[2] is to instantiate a central registry of relevant linguistic data categories called the *ISO Data Category Registry* (DCR). The data model of and procedures around this registry are described in the latest revision of the ISO 12620 standard (ISO 12620, 2009). The previous 1999 version of this standard contained a hardcoded list of data categories commonly used by various standards produced by ISO TC37. This list is now being replaced by a registry which can be collectively maintained and used by the linguistic community at large. The coherence of the standardized core of the registry will be maintained by members of TC37, *i.e.*, data categories will undergo the normal ISO standardization process.

Data categories are defined in the standard as the "result of the specification of a given data field" (ISO 12620, 2009). The DCR data model allows very detailed specifications of these data categories with an administrative, descriptive, and linguistic section. The descriptive section contains at least an English name and definition for the data category, and optional translations of these in relevant other working languages. In the linguistic section the value domain of a data category is specified and can be restricted for specific object languages. To support the interchange of selections of data categories the standard describes an XML serialization of the data model, the Data Category Interchange Format.

Each data category is assigned a persistent identifier (PID), especially suited to be included in metadata and schemata of linguistic resources to foster semantic interoperability. Some schema languages, *e.g.*, TBX XCS and TEI ODD, have built-in support to embed these PIDs into the schema. However, more generic schema languages such as Relax NG and W3C XML Schema do not. But as these languages are XML-based they can easily be extended by annotating specific parts of the schema documents with attributes or elements from the DC Reference XML vocabulary (see Annex A in (ISO 12620, 2009)).

ISOcat[3], see Fig. 1, is a reference implementation of ISO 12620. It implements the data model and will support all procedures to use and maintain the DCR. Using its web-based interface, data categories can be created, edited, shared, exported and standardized; ISOcat also supports a web services interface.

Various efforts are undertaken to populate ISOcat with widely useful categories. In the RELISH project[4] in cooperation with the original authors concepts from the GOLD ontology[5] are incorporated into the registry. Also various tagsets, *e.g.*, STTS[6], are in the process of becoming available via ISOcat. In the Dutch CLARIN initiative a number of projects[7] are rewarded to curate their resources which

[2] http://www.tc37sc4.org/

[3] http://www.isocat.org/

[4] http://ems06.mpi.nl/relish/

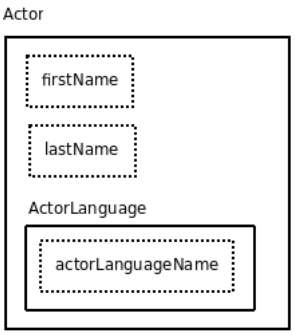
[5] http://linguistics-ontology.org/

[6] http://www.sfs.uni-tuebingen.de/Elwis/stts/stts.html

[7] http://www.clarin.nl/node/70

```
<CMD_Component name="Actor">
  <CMD_Element name="firstName" ValueScheme="string"
  ConceptLink="http://www.isocat.org/datcat/CMD-123">
  <CMD_Element name="lastName" ValueScheme="string"
  ConceptLink="http://www.isocat.org/datcat/CMD-124"/>
  <CMD_Component name="ActorLanguage" id="ActorLanguage"
  CardinalityMin="0" CardinalityMax="unbounded">
    <CMD_Element name="ActorLanguageName"
    ValueScheme="string"
    ConceptLink="http://www.isocat.org/datcat/DC-1766"/>
  </CMD_Component>
</CMD_Component>
```

(a) XML representation



(b) Schematic representation.

```
<Actor>
  <firstName>Foo</firstName>
  <lastName>Bar</lastName>
  <ActorLanguage>
    <ActorLanguageName>Kilivila</ActorLanguageName>
    <ActorLanguageName>French</ActorLanguageName>
  </ActorLanguage>
</Actor>
```

(c) Actor instance in XML

Figure 2: A metadata component description.

specifically involves relating the linguistic concepts at the data level and concepts used at the metadata level to relevant ISOcat data categories and extending the available set of data categories where necessary.

4. Metadata Component Model

The ISO DCR serves as a common ground; in our approach, it is complemented with a component registry to build a flexible metadata framework. Components serve as small reusable templates describing specific aspects of a resource which may be combined into higher level components to provide more complex reusable templates. Besides the aspect of reuse components also serve as guide lines for data structuring and may be extended depending upon a particular usage scenario.

4.1. Model

A *metadata component* describes different aspects or dimensions of a resource. Such a component is basically a collection of metadata fields. Each field refers via a URI to exactly one data category in the ISO DCR, thus indicating unambiguously how the content of the field in a metadata description should be interpreted. Components can have a recursive structure: next to the atomic fields they may contain other components. They are expressed as XML-files, an example component description can be found in Fig. 2(a) with the graphical representation in Fig. 2(b); here, fields are marked with dotted lines, components with a solid line. Any number of components can be combined with a header element into a *metadata profile*. A profile, also represented as an XML-file, provides a blueprint for the personalised metadata schema. It can be converted into a W3C XML-schema with an XSLT transformation; references to all occuring data categories will be maintained.

Finally, using the generated XML-schema to check for the formal correctnes of the data, the XML files containing the metadata descriptions can be created. A fragment of such an example description is shown in Fig. 2(c). As each metadata description will contain a link to its W3C schema, its validity can be checked and the data categories used in the description can be retrieved.

CLARIN will suggest a number of recommended components that will be made available in a component registry. But users can use and create their own components, given that its elements explicitly refer to concepts registered in ISOcat or other trusted registries. If a user wants to include an element which is not yet registered, she would need to register the new concept at least in the so-called ISOcat ”user space”. The ISOcat process will then decide whether the new category will be integrated in the official part of the registry. CLARIN will be strict and only accept categories that are registered in accepted registries, since otherwise no semantic interoperability can be established.

Fig. 3 describes the interaction between the data category registries, the component registry, the component editor and the metadata editor. The design aims at promoting the re-use of existing fields and components, but also gives users the opportunity to create new ones within the well-defined ISO DCR process. Naturally, we anticipate the creation of different metadata profiles for different communities, *e.g.*, sign-language researchers will need to describe a video signal differently than multimodality experts.

Clearly, the principles behind this model need to be enforced by the accompanying end user software: the component editor will check whether all elements used are indeed taken from ISOcat or another trusted registry, and it will interact with the component registry to store final component ensembles or profiles and make them re-usable.

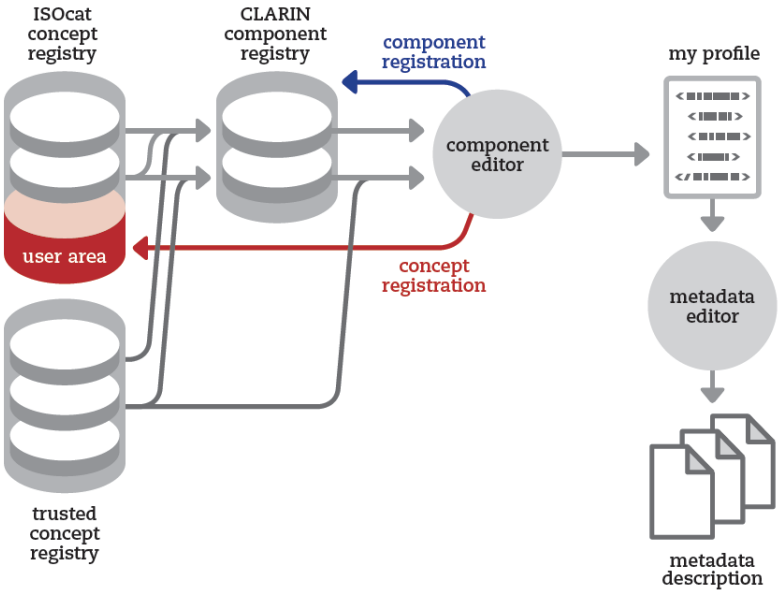


Figure 3: Interaction of registries and editors

4.2. Initial Set of Metadata Categories

A process was established within the CLARIN network to identify a first set of basic concepts to be included in the ISO DCR. A metadata Thematic Domain Group was established in the DCR and an initial list of 154 data categories with their definitions (and partly, with constraints) was created[8]; they describe the resoures' creation process, their content and the participants contributing to the content, formal characteristics of the resources, and the resources' identity and access. Taking together, they should provide a good basis for describing all major linguistic resource types (such as media resources, textual resources and corpora, annotations, and lexica) as well as tools and web services.

In the first selection of data categories, special attention was given to *profile matching*, an automatic process that identifies suitable processing components for a given resource (*e.g.*, a parser that can cope with a given resource's text format). Here, CLARIN aims to build metadata-based support for the construction of workflow chains where the output of an operation must be accepted by the input of sub-sequent operations. Categories describing resources at different granularities, ranging from the very general "mimetype" to the very specific "tag set", have been devised. These data categories have been structured in a so called *TechnicalMetadata* component describing the resource's relevant technical characteristics. Web services reuse this component to declare their requirements for input parameters to determine whether a resource satisfies the required technical characteristics for operating the web service. Similarly web services also use the *TechnicalMetadata* component to specify the characteristics for the resulting resources which is used to determine the web services that may be sued for subsequent processing.

Note that data category definitions shall be widely independent of the contexts in which they can occur. So far, we have thus chosen a high granularity semantics for those elements that we expect to be consumed by machine processing, searching or profile matching. For those categories aimed at human consumption, the semantic scope and granularity is not as critical as humans can use context quite efficiently to disambiguate between various meanings.

The ISO DCR model is providing a flat list of categories. No relations are included in the registry, except for more generic categories that are part of the definition *e.g.*, a *transitive verb* is a *verb*). Since relations often depend on their use in context, the ISO committee decided to separate them from the definitions. For this, the notion of a *relation registry* has been proposed (Kemps-Snijders et al., 2008), constituting a store for expressing relations between entries of data category registries, and thus imposing some structure. We anticipate that contents of the relation registry will support semantic search.

4.3. Initial Set of Components – CMDI

The CLARIN Metadata Infrastructure (CMDI) initiative (carried by *Institut für Deutsche Sprache*, Max Planck Institute for Psycholinguistics, *Språkbanken* – Department of Swedish Language at Göteborg University, German Research Center for Artificial Inteligence and the Austrian Academy of Sciences) will cooperate building the infrastructure to create a component specification language as a first step; Fig. 2 gives an example. A decomposition of the IMDI and OLAC metadata schemes has provided the CMDI with its first metadata components.

In September 2009, CLARIN-NL has initiated a project to develop metadata components for resources hosted at two major Language Resource centres in the Netherlands. Again, all resulting components will reference metadata fields from ISOcat.

At the time of writing, our component editor is a standard XML editor; schemas enforce the formal correctness of all

[8] http://www.clarin.eu/view_datcats

components created. The component registry is in its early stages and is based on the SVN versioning system. However, the implementation of a purpose-built component registry, a component editor, and a metadata editor has already started, and the aim is to deliver a usable infrastructure in the third quarter of 2010.

5. Conclusion

CLARIN is moving toward a DCR- and component-based metadata infrastructure with a critical mass of representative metadata descriptions for language resources and tools. The investments in this new component metadata infrastructure will pay off, given the large number of existing language resources and tools, and the crucial role of metadata to access and exploit them. The realization of workflow chains, supported by profile matching, will further push the need for detailed metadata, but our registries framework is flexible enough to cope with evolving requirements.

6. References

T. Baker. 1998. Languages for Dublin Core. *D-Lib Magazine*, 4:12.

D. Broeder and P. Wittenburg. 2006. The IMDI metadata framework, its current application and future direction. *International Journal of Metadata, Semantics and Ontologies*, 1(2):119–132.

D. Broeder, T. Declerck, E. Hinrichs, S. Piperidis, L. Romary, N. Calzolari, and P. Wittenburg. 2008. Foundation of a component-based flexible registry for language resources and technology. In *6th International Conference on Language Resources and Evaluation (LREC 2008)*.

ISO 12620. 2009. *Computer Applications in Terminology – Data Categories – Specification of Data Categories and Management of a Data Category Registry for Language Resources*. ISO, Geneva, Switzerland.

M. Kemps-Snijders, M.A. Windhouwer, and S.E. Wright. 2008. Putting data categories in their semantic context. In *IEEE e-Humanities Workshop*, Indianapolis, Indiana, USA, December.

B. MacWhinney. 2000. *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum Associates.

Gary Simons and Steven Bird. 2003. The open language archives community: An infrastructure for distributed archiving of language resources, June.

T. Váradi, S. Krauwer, P. Wittenburg, M. Wynne, and K. Koskenniemi. 2008. CLARIN: Common language resources and technology infrastructure. *Proceedings of LREC08*.