Standardizing a Component Metadata Infrastructure

Daan Broeder[1], Dieter van Uytvanck[1], Menzo Windhouwer[1], Maria Gavrilidou[2], Thorsten Trippel[3]

[1] MPI for Psycholinguistics, Nijmegen, The Netherlands, [2] ILSP, Athens, Greece, [3]University of Tübingen, Tübingen, Germany

E-mail: {daan.broeder, dieter.vanuytvanck, menzo.windhouwer}@mpi.nl, maria@ilsp.athena-innovation.gr, thorsten.trippel@uni-tuebingen.de

Abstract

This paper describes the status of the standardization efforts of a Component Metadata approach for describing Language Resources with metadata. Different linguistic and Language & Technology communities as CLARIN, META-SHARE and NaLiDa use this component approach and see its standardization of as a matter for cooperation that has the possibility to create a large interoperable domain of joint metadata. Starting with an overview of the component metadata approach together with the related semantic interoperability tools and services as the ISOcat data category registry and the relation registry we explain the standardization plan and efforts for component metadata within ISO TC37/SC4. Finally, we present information about uptake and plans of the use of component metadata within the three mentioned linguistic and L&T communities.

Keywords: metadata, standardization, infrastructure

1. Introduction

Since the conception and implementation of the first version of a component metadata infrastructure (CMDI) [1], [2] within the CLARIN project [3] [4], it was realized that component metadata could not only mean interoperability between kindred communities within the Language Resource domain, but also between domains that stand further apart. In fact, component metadata is a very good candidate to be used by the projects working on research infrastructures that cover many communities, since it allows them to use and develop their own schema relying on semantic interoperability by an external concept registry. In Europe there are currently two interdisciplinary infrastructure projects where this will be tested. First there is DASISH [5], a community cluster project combining linguistics, wider humanities and the social sciences. Secondly the EUDAT [6] project that needs to cater for the needs of disparate communities as for instance life-sciences, climate, High Energy Physics and geophysics. For each of these CMDI is a good candidate.

An important step in making the component metadata approach successfully accepted is to standardize the component metadata with its model and required infrastructure. Currently the standardization is approached within ISO TC37/SC4. ISO offers a suitable platform for involvement of international researchers working for example in META-SHARE [7] and CLARIN, the communities currently working with metadata components. Another reason to do so is that the means for solving semantic interoperability issues, i.e. the ISOcat data category registry is also under control of ISO TC37/SC4.

2. Component Metadata short overview

The component metadata infrastructure (CMDI) offers a flexible framework for metadata modelers and metadata creators to use an appropriate metadata schema for characterizing a resource. It aims at making the metadata modeling process easy by allowing reuse of different metadata components that bundle descriptions for certain resource characteristics. These components can then be used in different combinations to come to a suitable metadata profile for describing a specific resource type. So components are bundles of metadata elements*;* these are used to encode specific descriptive features of the LRs, and these components can be combined in so-called profiles to describe specific LR data-types. Profiles can be used either to describe singular resources or sets of related resources such as collections.

Metadata modelers are able to use their own terminology deemed appropriate for the task inside the components, including terminology related to their language. This flexible use of terminology inevitably also creates semantic interoperability problems that we try to solve by using a combination of a concept registry, - more specific the ISO data category registry (ISOcat) [8] - and a relation registry (RELcat) [8] that cooperate to make the semantics of the metadata terms used and possible relations between those metadata concepts (see figure 1) explicit. The figure shows that metadata modelers may use their own terminology for elements in the metadata components. However it is mandatory to make the semantics clear by linking the component elements to a data category entry in the ISOcat. This may be either one and the same entry in which semantic equivalence is evident or they may refer to two different entries in which case a relation stored in a relation registry (RELcat) may be used to establish another type of semantic relation.

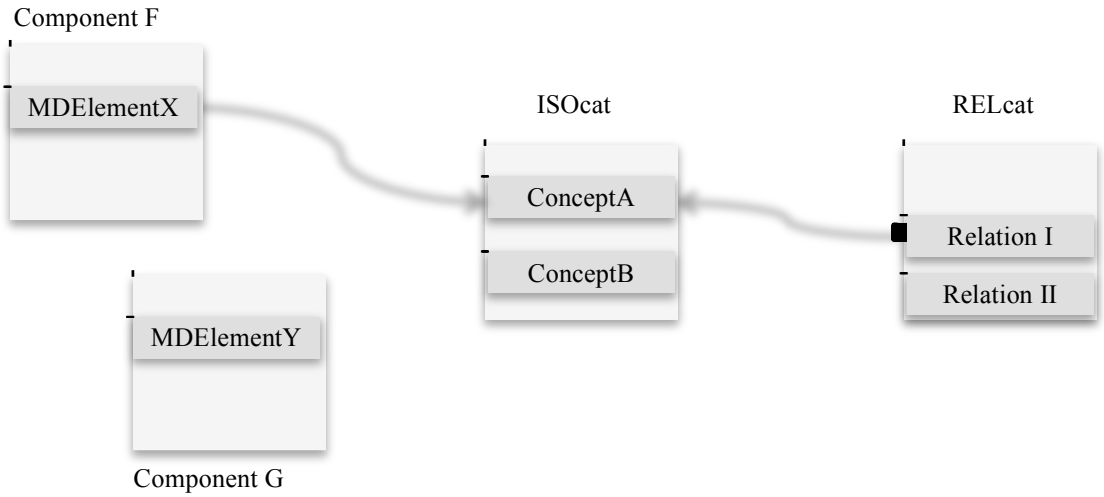

Figure 1 Semantic relations in CMDI

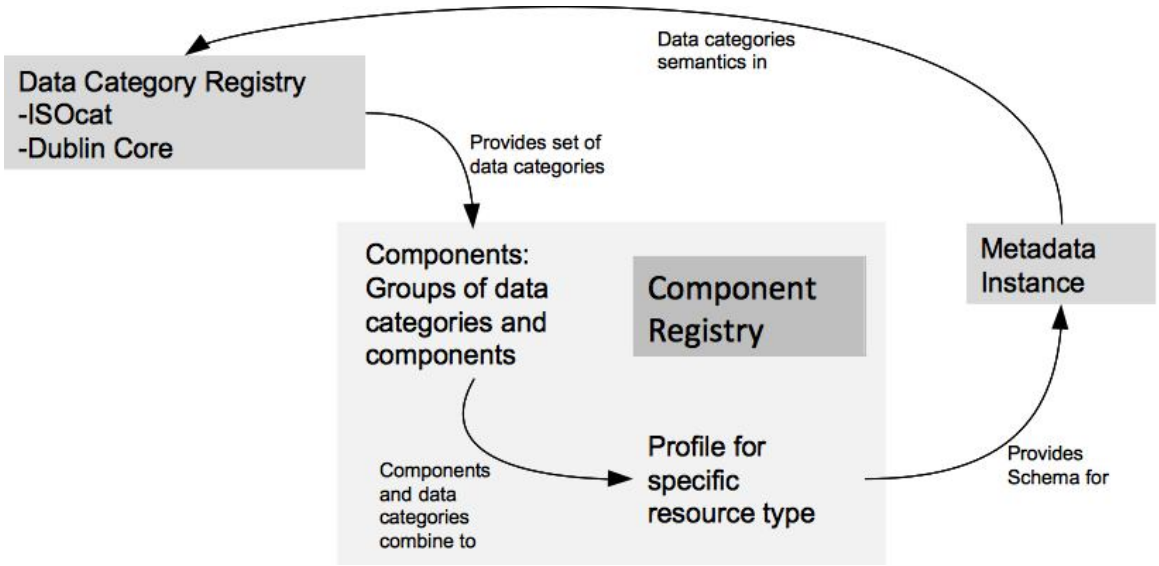

Figure 2 Interrelation of the DCR, Component registry and metadata instance

The interrelation of data category registry, component registry and metadata instance is shown in figure 2. Data categories, components and metadata instances exist independently of each other, but refer to each other as points of reference. The metadata instance for example contains data categories, which are defined in the data category registry; hence the interpretation of the data categories does not rely on the component registry. This is possible because the schema for the metadata instances refer to the data category registry.

3. The planned standardization process

We have chosen ISO TC37/SC4 for the standardization process because it is ISO's standardization working group on language resources management and metadata frameworks for language resources is appropriately represented there. Also some contributors to CMDI are already familiar with ISO standardization process and committee workings but most of all because it also allows for representation of all important groups within the Language Resource domain including CLARIN, META-SHARE and FlaReNet [9].

Within the ISO standardization process proposals for standards have to pass several cycles of drafts, reviews, discussions and modifications before they become international standards. The reviewers belong to the national standards committees, which expose any proposal to criticism from outside the originating group, preventing a closed group approach.

The strategy taken to divide work and allow proper input from is to divide the work between representatives of different communities and other interested parties. For the standardization work on CMDI, the work was split up in three parts:

1. A general model that specifies the system's terminology and all essential characteristics like the need for a recursive model allowing components to contain other components and the possibility for metadata schema instantiations to refer to other metadata instantiations as well as data resources.
2. A specification of one or more implementation languages that can be used to specify metadata components and specify the metadata schemas.
3. A (first) set of standardized components and

profiles for much used data types. This set can be updated with subsequent versions of the standard.

This was proposed at the ISO TC37/SC4 meeting November 2010 and was further endorsed at the meeting in Seoul June 2011. Currently the CMDI Part 1 was approved as a work item and a committee working draft will be circulated first quarter 2012. The project leaders of the different parts will include each other in their teams and work closely together.

4. Standardizing Metadata Concepts

An important element of the CMDI is the use of the data category registry ISOcat to help solve issues of semantic interoperability where metadata modelers use different terminology. ISOcat is positioned as a general registry for linguistic concept definitions, and it was natural for the component metadata initiatives in the LR domain such as those from CLARIN and META-SHARE to use the ISOcat to register metadata concept definitions.

In January 2009 experts met forming a group named the 'Athens Core' group that was a broad representation from the LR community. In two subsequent meetings ISOcat was seeded with appropriate metadata concepts taken from the IMDI, OLAC and Enabler metadata schemas; further elaboration of the registry, addition of new terms and improvement of the description of existing ones is an on-going procedure

However, the official ISO standardization process for ISOcat entries may prove to be too slow for helping the CLARIN and META-SHARE projects meet their requirement of having ISO approved concepts to build their metadata components. So a pragmatic approach is now considered where all Athens Core metadata concepts are considered valid to use in any CMDI standard, even if not officially sanctioned by ISO. Currently there is some discussion on the possibility to have separate LR community sub-ISO level standard approval registry and the Athens-Core metadata concepts could be part of that. It must become clear in 2012 if such an approach is acceptable for the different stakeholders and if they can collaborate in realizing such a structure.

5. Status of Component Metadata in the Communities

5.1 CLARIN

From the work within CLARIN with respect to a component metadata approach we already reported in [2]. Since then CLARIN has concentrated on making its CMDI tools more stable and work within several projects on creating CMDI schema for resources and services. Especially the creation of common metadata schema for web-services has been a challenge, trying to combine different approaches from different CLARIN partners, in such a way that these different web-service workflow infrastructures can make use of one another's web-services. Currently there are about 60 CMDI profiles and 170 metadata components. About 140000 CMDI metadata records are available from 5 different metadata providers via OAI-PMH harvesting.

For the LR metadata user, a new version of the VLO metadata facetted browser [10] [11] is available and shows all harvested CMDI metadata records. It is now also possible to use facetted browsing using ISOcat registered concepts.

5.2 NaLiDa

The German NaLiDa project [12] on sustainability of linguistic resources provides services to various linguists and linguistic research groups, aiming at a close collaboration with central university infrastructures such as the library and computer center to foster long time archiving of these resources in institutionalized contexts. For archiving primary resources they usually start by creating the metadata for the resources, using the CMDI framework, currently with ten self-developed profiles and 139 components, some of them derived from other components. For reference purposes 12 CMDI metadata records were manually created using almost all metadata categories available in the respective profile, manually corrected and with a high quality. About 180 records were the output of automatic processes, legacy data transformation not included. The metadata is used as an input for a search application [13], based on harvested metadata and local metadata, currently covering more than 10,000 records.

5.3 META-SHARE

META-SHARE is an open distributed facility for the exchange and sharing of LRs, i.e. a network of repositories of language data, tools and related web services. Uniform search and access to these resources is obtained through their descriptions. Within META-SHARE a schema for metadata components and profiles has become available [14], [15] with a total of about 80 components and more than 350 defined metadata elements currently, covering the LR types corpus, lexical/conceptual resource, tool/technology/service, language description and the media types text, audio, video and image.

The set of all the components and elements describing specific LR types and subtypes represent the profile of this type. Obviously, certain components include information common to all types of resources (e.g. identification, contact, licensing information etc.) and are, thus, used for all LRs, while others (e.g. components including information on the contents, annotation etc. of a resource) differ across types. This schema has already been used for the documentation of 1277 data-sets, supplied by the project partners. This provides a good start position for the CMDI Part 3 standardization work.

CLARIN, NaLiDa and META-SHARE have many participants that are active in each other's projects, which should give a good basis for convergence of ideas in the foreseen standardization work.

6. Conclusion and Future Initiatives

It maybe yet too early to come to final conclusions about the success of the component metadata approach especially since 'real' success should not only be measured in uptake by archives and projects that use it for their archiving and administration purposes, but also by outside users that can use the infrastructure to locate resources according to their needs. At the moment we will only state that at the metadata production side, things are coming along including a standardization initiative that can act as a uniting force with the LR domain.

There is no reason why the CMDI approach cannot also be applied to provide metadata interoperability within and between other domains that suffer from different metadata schemas and varying terminology. This relies only on the metadata concepts to be defined in (similar) concept registries as ISOcat and the metadata schemas to provide links to them. Of course some minimal semantic overlap between the metadata schemas of the disciplines must exist to make such interoperability useful.

We hope to gain experience in applying CMDI in other domains within the DASISH project and perhaps EUDAT that were already mentioned before and where obtaining semantic interoperability for metadata from different disciplines is one of the challenges.

7. References

[1] CMDI, http://www.clarin.eu/cmdi

[2] Broeder, D., Schonefeld, O., Trippel, T., Van Uytvanck, D., and Witt, A. (2011). A pragmatic approach to XML interoperability — the component metadata infrastructure (CMDI). In *Balisage: The Markup Conference 2011*, volume 7.

[3] Váradi, T., Wittenburg, P., Krauwer, S., Wynne, M., & Koskenniemi, K. (2008). CLARIN: Common language resources and technology infrastructure. *LREC 2008*

[4] CLARIN, http://www.clarin.eu/

[5] DASISH, http://dasish.eu/

[6] EUDAT, http://www.eudat.eu/

[7] META-SHARE, http://www.meta-net.eu/meta-share, http://www.meta-share.eu/

[8] Schuurman, I., Windhouwer, M. A., Explicit Semantics for Enriched Documents. What Do ISOcat, RELcat and SCHEMAcat Have To Offer? *SDH 2011*, 17-18 November 2011, Copenhagen, Denmark.

[9] FLaReNet, http://www.flarenet.eu/

[10] VLO, http://catalog.clarin.eu/ds/vlo/

[11] Van Uytvanck, D., Zinn, C., Broeder, D., Wittenburg, P., & Gardelini, M. (2010). Virtual language observatory: The portal to the language resources and technology universe. *LREC 2010*

[12] NALIDA, http://www.sfs.uni-tuebingen.de/nalida/en/

[13] NALIDA faceted browser, http://www.sfs.uni-tuebingen.de/nalida/en/catalogue.html

[14] Gavrilidou M., Labropoulou P., Piperidis S., Monachini M., Frontini F., Francopoulo G., Arranz V. and Mapelli V. (2011). A Metadata Schema for the Description of Language Resources (LRs), *IJCNLP 2011*.

[15] Labropoulou P., Desipri E., (eds.). (2012). Documentation and User Manual of the META-SHARE Metadata Model, http://www.meta-net.eu/meta-share/metadata-schema/