Originally published in: Arranz, Victoria/Broeder, Daan/Gaiffe, Bertrand/Gavrilidou, Maria/Monachini, Monica/Trippel, Thorsten (Eds.): Proceedings of the workshop describing language resources with metadata: towards flexibility and interoperability in the documentation of language resources. LREC 2012, May 22, 2012, Istanbul, Turkey. – Paris: European Language Resources Association, 2012. Pp. 1-4.

CMDI: a Component Metadata Infrastructure

Daan Broeder [1], Menzo Windhouwer[1], Dieter van Uytvanck[1], Twan Goosen[1], Thorsten Trippel[2]

[1]MPI for Psycholinguistics, Nijmegen, The Netherlands, [2]Eberhard-Karls-Universität Tübingen (Germany),
{daan.broeder|menzo.windhouwer|dieter.vanuytvanck|twan.goosen}@mpi.nl, thorsten.trippel@uni-tuebingen.de

Abstract

The paper's purpose is to give an overview of the work on the Component Metadata Infrastructure (CMDI) that was implemented in the CLARIN research infrastructure. It explains, the underlying schema, the accompanying tools and services. It also describes the status and impact of the CMDI developments done within the CLARIN project and past and future collaborations with other projects.

1 Introduction

Currently there is a fragmented world with respect to metadata for Language Resources (LR). However recently there have been initiatives that give some hope of creating interoperable schemas of high specificity that allow the creation comprehensive catalogues of LRs.

Before 2000 there were mainly the proprietary catalogues of the commercial companies and language resource centers as LDC and ELRA and the practice of inserting metadata in the transcriptions or annotation file headers as for example TEI and CHILDES formats support. Yet little attention was given to interoperability between archives and data centers using different metadata schema. Since 2000 we have seen the rise of new LR metadata schemas as IMDI, IMDI [2003], IMDI [2009] and OLAC but application and uptake of these schemas has been limited. Although OLAC is now used more or less as a standard for information exchange between LR archives, it is still delivering low specificity.

The experience in creating IMDI and trying to apply it to the variety of subdomains in linguistic research has helped to realize that a single metadata schema cannot succeed in conquering all sub fields of linguistics. The differences in needs, terminology and traditions will prevent uptake and acceptance of such a schema. Therefore, when there was a need to come to a comprehensive approach for metadata within the CLARIN infrastructure [Váradi et al., 2008], we chose to build an infrastructure permitting many different schemas to co-exist and supporting semantic interoperability by using a separate 'pragmatic reference system' for the semantics being implied. To support users with a low threshold for creating new schemas and reusing existing work at a conceptual level, an approach was chosen where small reusable snippets of metadata schema's can be created and recombined to form complete new schemas. This component based approach or Component Metadata Infrastructure [CMDI, Broeder et al.] is based on well-defined formal schemas and explicit semantics by using registries for the schema components, the final schema and the pragmatic ontology.

2 CMDI overview

CMDI is a flexible framework for metadata modelers and metadata creators to create and use appropriate metadata schemas for describing resources. It aims at making the metadata modeling process easy by allowing reuse of different snippets of metadata schemas or metadata components that bundle descriptions for certain resource characteristics. These components can be recombined to create a suitable metadata profile for describing a specific resource type. Components hence contain metadata elements or other components, forming profiles to be used either to describe singular resources or sets of related resources such as collections. Figure 1 illustrates the model.

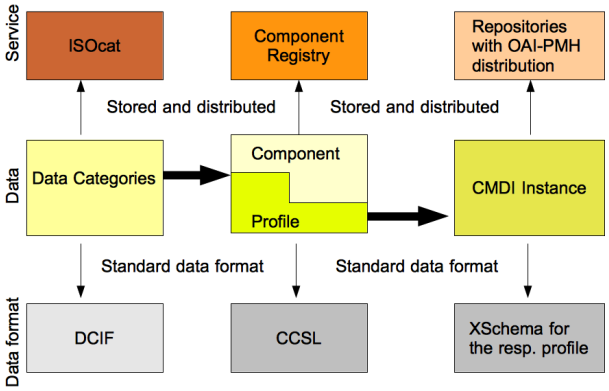

Figure 1: Model of the component metadata infrastructure

Each of the constituents of the model has a three layer structure, from the bottom: a data format, the data and a service storing and distributing the data.

Metadata modelers are able to use their own terminology deemed appropriate for the task in the components. This flexible use of terminology inevitably also creates semantic interoperability problems that we try to solve using a 'pragmatic' ontology, which is a combination of a concept registry — more specific the ISO data category registry [ISOcat] — and a relation registry [RELcat, Schuurman and Windhouwer, 2011]. The data in ISOcat is available in the Data Category Interchange Format (DCIF), which is a standard format as defined by ISO 12620 [2009].

Metadata registry and relation registy together provide the semantics of the metadata terms used and make possible relations between the metadata concepts explicit. Metadata modelers may also use their own terminology — or terms in their own language — for elements in the metadata components and remain interoperable by linking the component elements to the corresponding data category entry in ISO-

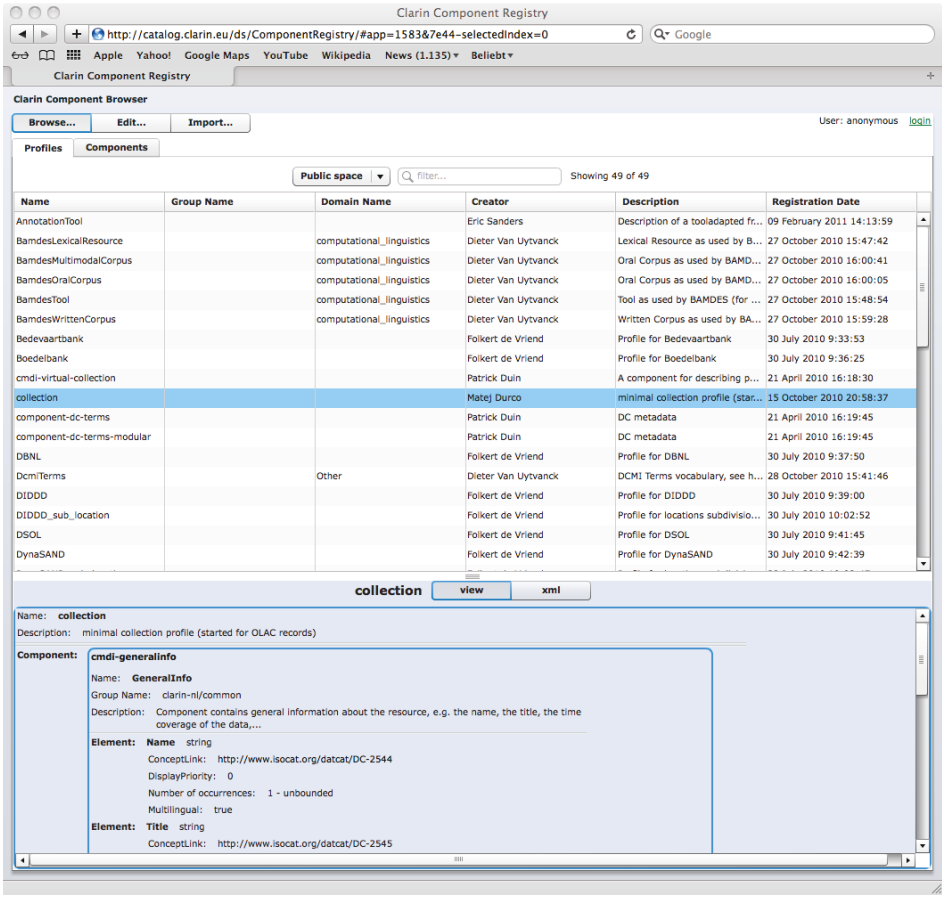

Figure 2: The CMDI Component Registry

cat. Different but similar elements can refer to the same entry if they are semantically equivalent or they can refer to different entries, for which their semantic similarity — their relation — is stored in the relation registry (RELcat).

The metadata components, the combined components and profiles are stored in the CMDI component registry, depicted in the center of Figure 1. They are defined in the CMDI-Component Specification Language (CCSL) and distributed using a REST-based API or a browser interface. Users can browse this registry and combine existing components in a new profile (see Figure 2). New components can be created using the component editor and stored in the component registry.

For creating actual instantiations of the metadata profiles, these are automatically transformed into XML schemas, also available from the component registry. They are used for validating the metadata instances, the metadata records that describe actual resources. These can be created in a variety of ways, for example by transforming legacy data. For direct creation we have developed ARBIL which is a versatile metadata editor. ARBIL allows users to manipulate and edit metadata of many metadata records by using table structures instead of the unformated XML-code.

Within the CLARIN infrastructure, CMDI is the metadata infrastructure of choice. The different CLARIN centers and others that act as LR providers share and distribute their metadata in CMDI format via OAI-PMH to be harvested by CMDI service providers. The right side of Figure 1 illustrates that. Such service providers may choose to harvest all or a sub-set of CMDI data-providers and aggregate the metadata in metadata catalogs. Within CLARIN we have currently the following catalogs: the CLARIN VLO [van Uytvanck et al., 2010] and the Meertens Institute CMDI Catalogue [CMDI MI Search Engine] and outside CLARIN there is the NaLiDa faceted browser [see NaLiDa FB], see Figure 3.

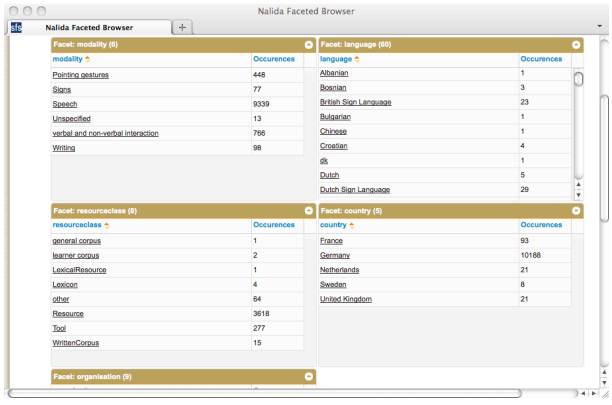

Figure 3: Faceted Browser for CMDI metadata

All the approaches mentioned above offer a faceted browser for structured access to the repositories, often combined

with a full text search of the metadata for example by using the Apache SOLR and Lucene combination [see SOLR]. This allows users to navigate in the harvested collection's metadata by defining criteria to be fulfilled by the searched resources. The possible criteria provided by a faceted browser are the facets, for this reason it is important to choose the facets appropriately to the intended use. As the facets may refer to different instantiations of similar data categories it seems appropriate to also use the terms and mappings from the ISOcat registry. The last requirement has been partly fulfilled in the VLO where facets are based on ISOcat data categories.

For more complex combinations of search terms, faceted browsing and searching has some limits, for example in the number of facets. To overcome these limits the Austrian CLARIN project is working on a prototype supporting complex CMDI metadata search queries. This work in progress aims at power-users that have a good grasp all the aspects of the CMDI infrastructure including the possibilities of varying precision and recall by varying the semantic mapping variables [Durco et al., 2012, submitted]. This prototype is the most complete implementation of the complete CMDI architecture that is shown in Figure 4.

The complete system for CMDI metadata creation and exploitation is depicted in Figure 4. At the (left) exploitation side CMDI metadata is harvested and put in a joint CLARIN metadata repository. There it is either consumed by simple but effective faceted browser tools as the VLO or by complex ones as the Austrian MD Search, making use of Semantic Mapping services provided by the pragmatic ontology using the combination of ISOcat and the Relation Registry.

3 Standardization efforts

An important step in making the component metadata approach successful and sustainable for long time archiving, is aiming a standardization of the framework. This also offers an opportunity to cooperate with like-minded projects such as META-SHARE [see Gavrilidou et al., 2011], which also wants to use a component metadata approach, to achieve interoperability. The standardization is running under the auspices of ISO TC37/SC4 offering an institutionalized platform for the involvement of relevant parties such as META-SHARE and CLARIN, the communities currently working with metadata components. This standardization bodies technical committee is also governing the means for solving semantic interoperability issues, the ISOcat data category registry, with ISO 12620:2009 being hosted by the sister subcommittee ISO TC37/SC3.

An important element of CMDI is the use of ISOcat to help solve issues of semantic interoperability where metadata modelers use different terminology. ISOcat is positioned as a general registry for linguistic data category definitions, and it was natural for the component metadata initiatives in the LR domain such as those from CLARIN and META-SHARE to use ISOcat to register metadata concept definitions. Currently, a group of experts informally termed 'Athens Core´ that is a broad representation from the LR community pushes the metadata concept ISO standardization process forward. More details on the CMDI related standardization processes are found in Broeder et al. [2012]).

4 Status of CMDI usage

Currently we know of the different national CLARIN projects, the German NaLiDa project and some smaller projects that have been using CMDI implementations or are planning to use it. We expect there to be some papers at the LREC 2012 'Describing Language Resources' workshop.

The VLO currently lists over 180000 resources, described by metadata, the component registry lists 49 different profiles and 218 components in the public section (as of February 2012), with about 15 commiters from various institutions. There are 62 registered users of the component registry. Registered users here means that they have created and modified components, read access does not require registration. Besides the public profiles and components there are currently 127 private profiles and 303 private components showing very active development going on.

5 Conclusion and future initiatives

It is too early to come to any final conclusions about the success of component metadata, also because its success cannot be measured only in acceptance by the metadata creators. It also depends if outside users can use CMDI to locate the resources they require, hence the success is depending on tools to work with CMDI.

At the moment on the metadata production side, things are coming along although some attention needs to be paid to the risk of insufficient reuse of existing CMDI components and profiles and proliferation of different profiles. At the metadata exploitation side there remain many challenges but we trust that there will be several solutions also because CLARIN centers are accepting CMDI tagged resources and will need to provide metadata exploitation solutions for their own users as well as for outside users.

We think that a communal standardization initiative of CLARIN and META-SHARE will lead to an acceptable implementation for all groups that are pledged to the use of metadata components and explicit semantics using ISOcat.

Another aspect of component metadata is that it is a very good candidate to be used by the projects working on research infrastructures catering for a variety of communities and disciplines. They have to deal with a large variety of data types and have to bridge differences in terminology used by different communities. One example is DASISH which is a community cluster project combining linguistics, wider humanities and the social sciences where CMDI could be successfully applied.

6 Acknowledgements

Work for this paper was conducted within the Clarin-NL project and in the NaLiDa project funded by the German Research Foundation (DFG) in the program for Scientific Library Services and Information Systems (LIS).

References

ARBIL. http://www.lat-mpi.eu/tools/ARBIL.

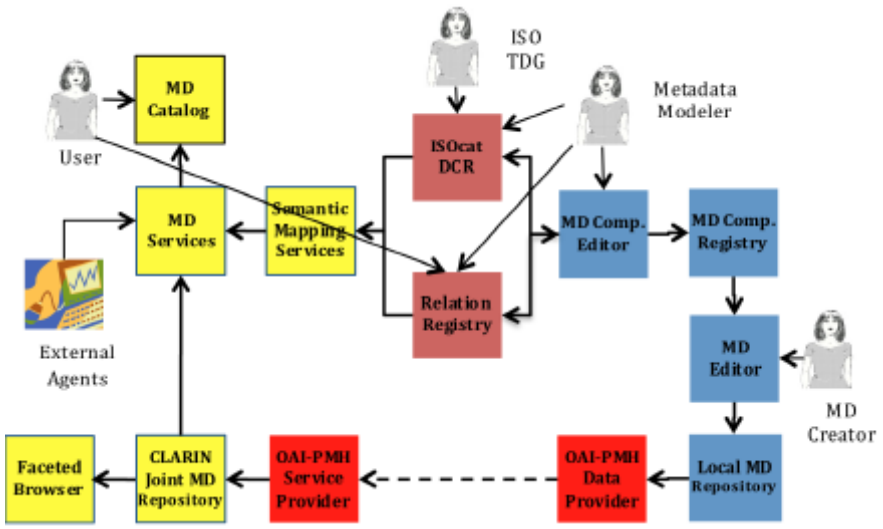

Figure 4: The CMDI Architecture

D. Broeder, O. Schonefeld, T. Trippel, D. van Uytvanck, and A. Witt. A pragmatic approach to xml interoperability - the component metadata infrastructure (CMDI). In *Balisage: The Markup Conference 2011*, volume 7.

D. Broeder, D. van Uytvanck, M. Gavrilidou, and T. Trippel. Standardizing a component metadata infrastructure. In *Proceedings of the 8th Conference on International Language Resources and Evaluation (LREC 2012)*, Istanbul, 2012.

CHILDES. http://childes.psy.cmu.edu.

CMDI. http://www.clarin.eu/cmdi.

CMDI MI Search Engine. CMDI Meertens Institute Search Engine. http://www.meertens.knaw.nl/cmdi.

DASISH. http://www.lat-mpi.eu/latnews/tag/dasish/, project website forthcoming.

M. Durco, D. Broeder, and M. Windhouwer. Semantic mapping - groundwork for query expansion and semantic search. 2012, submitted.

M. Gavrilidou, P. Labropoulou, S. Piperidis, M. Monachini, F. Frontini, G. Francopoulo, V. Arranz, and V. Mapelli. A metadata schema for the description of language resources (lrs). In *5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, 2011.

IMDI. http://www.mpi.nl/IMDI.

IMDI. Metadata elements for session descriptions, draft proposal version 3.0.4. `http://www.mpi.nl/IMDI/documents/Proposals/IMDI_MetaData_3.0.4.pdf`, October 2003.

IMDI. Metadata elements for catalogue descriptions, version 3.0.13. `http://www.mpi.nl/IMDI/documents/Proposals/IMDI_Catalogue_3.0.0.pdf`, August 2009.

ISO 12620. Terminology and other language and content resources - specification of data categories and management of a data category registry for language resource. Technical report, ISO, 2009.

ISOcat. http://www.isocat.org.

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

NaLiDa. http://www.sfs.uni-tuebingen.de/nalida/en/.

NaLiDa FB. NaLiDa faceted browser. http://www.sfs.uni-tuebingen.de/nalida/en/catalogue.html.

OLAC. http://www.language-archives.org/.

I. Schuurman and M. Windhouwer. Explicit semantics for enriched documents. what do isocat, relcat and schemacat have to offer? In *2nd Supporting Digital Humanities conference (SDH 2011)*, Copenhagen, November 2011.

SOLR. http://lucene.apache.org/solr/.

TEI. http://www.tei-c.org/.

D. van Uytvanck, C. Zinn, D. Broeder, P. Wittenburg, and M. Gardelini. Virtual language observatory: The portal to the language resources and technology universe. In *Proceedings of the 7th conference on International Language Resources and Evaluation*, Malta, 2010.

VLO. http://catalog.clarin.eu/ds/vlo/.

T. Váradi, P. Wittenburg, S. Krauwer, M. Wynne, and K. Koskenniemi. Clarin: Common language resources and technology infrastructure. In *Proeceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, 2008.