Creating & Testing CLARIN Metadata Components

Folkert de Vriend (1), Daan Broeder (2), Griet Depoorter (3), Laura van Eerten (3), Dieter van Uytvanck (2)

1) Meertens Institute
Joan Muyskenweg 25, Amsterdam, The Netherlands
2) Max Planck Institute for Psycholinguistics
Wundtlaan 1, Nijmegen, The Netherlands
3) Institute for Dutch Lexicology
Matthias de Vrieshof 2-3, Leiden, The Netherlands
Folkert.de.Vriend@meertens.knaw.nl, {Daan.Broeder, Dieter.vanUytvanck}@mpi.nl,
{Laura.vanEerten, Griet.Depoorter}@inl.nl

Abstract

The CLARIN Metadata Infrastructure (CMDI) that is being developed in CLARIN (Common Language Resources and Technology Infrastructure) is a computer-supported framework that combines a flexible component approach with the explicit declaration of semantics. The goal of the Dutch CLARIN project "Creating & Testing CLARIN Metadata Components" is to create metadata components and profiles for a wide variety of existing resources housed at two data centres according to the CMDI specifications. In doing so the principles of the framework are tested. The results of the project will be of benefit to other CLARIN-projects that are expected to adhere to the CMDI framework and its accompanying tools.

1. Introduction

Descriptive metadata is used to characterize data resources and tools, to facilitate discovery and management in large (virtual) infrastructures and repositories. One of the goals of the CLARIN (Common Language Resources and Technology Infrastructure) project is to create a joint metadata domain for all LRT resources (Váradi, Wittenburg, Krauwer, Wynne & Koskenniemi 2008). To achieve this purpose a metadata infrastructure is being developed that combines a flexible component approach with the explicit declaration of semantics. This framework is called CLARIN Metadata Infrastructure (CMDI) and is described in Broeder, Declerck, Hinrichs, Piperidis, Romary, Calzolari & Wittenburg (2008). The need for a flexible component based metadata framework resulted from the experience in the LRT community that fixed schema solutions hamper a broad usage due to the different needs and terminologies of subcommunities. This component metadata framework allows users, subcommunities and projects to design their own metadata schema as long as they make use of widely agreed upon concepts that are stored in the ISOcat registry and therefore guarantee interoperability.[1]

The goal of the Dutch CLARIN project "Creating & Testing CLARIN Metadata Components" is to create metadata components and profiles for a wide variety of existing resources housed at two data centres according to the CMDI specifications. In doing so the principles of the framework are tested. The results of the project will be of benefit to other CLARIN-projects that are expected to adhere to the CMDI framework and its accompanying tools. The project has three partners: The Max-Planck Institute for Psycholinguistics (MPI) carries out the coordination and management of the project.[2] The Institute for Dutch Lexicology (Instituut voor Nederlandse Lexicologie: INL) and the Meertens Institute (MI) are the two CLARIN-NL data centres that house the resources for which new CMDI metadata components and profiles are created and tested. Since INL and MI also aspire becoming an official CLARIN data centre for which adherence to the CMDI is a technical requirement, for these centers the project functions as a preparatory phase as well.

In section 2 we first summarize the basics of CMDI for creating and using metadata. In the rest of the paper the focus is on issues in using the CMDI principles for creating metadata for resources. In section 3 we first describe what resources were selected at the data centers. In section 4 we then discuss two aspects for which a resource needs to be analyzed before metadata can be created. In section 5 we go into detail about the actual creation of metadata components and profiles. Finally, in section 6 we draw some conclusions.

2. The CMDI infrastructure

Although the principles of CMDI have been described in Broeder et al. (2008) we summarize the basics of CMDI and its terminology in this section.

The CMDI design and construction was started by the European CLARIN (CLARIN EU) project to overcome the limitations of the existing metadata sets such as IMDI, OLAC and TEI.[3] Within the CLARIN EU project the MPI is responsible for guiding the implementation of the CMDI and therefore is very interested in participating in projects that will use or test the CMDI, like the project described in this paper.

The CMDI uses ensembles of metadata components that are called profiles to create xsd metadata schemas that can be used to describe resources or collections of resources. Every metadata component is a set of metadata elements that is supposed to describe a specific aspect of a resource, e.g. an "Actor" component specifies the biographical information of a person or a "Location" component specifies the place where an event occurred. Every metadata element is required to link to a recognized concept registry such as the ISOCat DCR[4] or the DCMI. Components and Profiles are stored in the CMDI component registry so others can reuse them.

Instantiated schemas describe actual resources and are called metadata descriptions or metadata records. CMDI should be flexible enough for any researcher to decide what metadata fits his or her needs best.

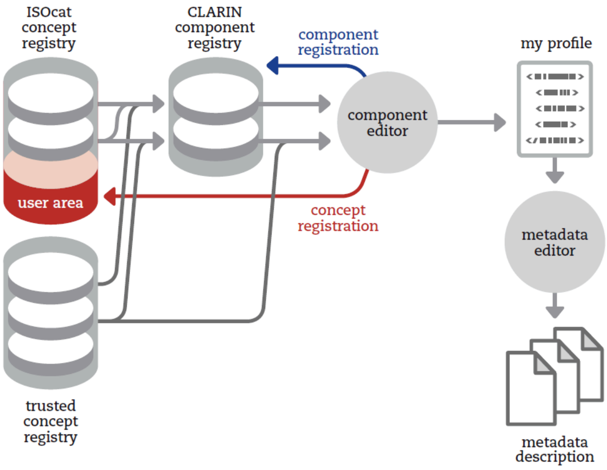

Figure 1: Creating a CMDI metadata description

In the CMDI infrastructure the metadata records are harvested with the OAI-PMH protocol and stored in a joint metadata repository. A CMDI service provider will then offer services like metadata search and browsing based on the repository's content to the world.

3. Data centers and resources

[1] www.clarin.eu/files/metadata-CLARIN-ShortGuide.pdf
[2] http://www.mpi.nl
[3] www.clarin.eu
[4] www.isocat.org

The MI data center studies and documents Dutch language and culture.[5] Its main fields of research are ethnology and variation linguistics. The research group Dutch Ethnology studies the dynamics and diversity in everyday cultural expressions. The research group Variation Linguistics studies variation in the Dutch language as manifested in dialects, sociolects and ethnic Dutch. Most of the MI resources are accessible online, some together with computer tools for resource based research (see for instance Barbiers, Cornips & Kunst 2007 or Meder 2010).

The INL data center collects and studies Dutch words, stores them in databases – along with various additional linguistic data – and uses them to make scholarly dictionaries. [6] The INL also hosts the Dutch-Flemish Human Language Technology Agency (HLT Agency), which manages, maintains and distributes Dutch digital LRs for research, education and commercial purposes (Beeken & van der Kamp 2004).[7] Many of the resources available through the HLT Agency were developed by third parties.

In the project metadata components were created for a subselection of the many resources housed at the two data centers.[8] A strong preference was given to those resources at MI and INL that were non multi-media or multi-modal type of resources. These types of resources are for instance lexical resources or text corpora. For the multi-media and multi-modal type of resources it was expected that the existing component set and profile that was derived from IMDI (ISLE Meta Data Initiative) would already be sufficient (Broeder & Wittenburg 2006).[9] Another reason to choose resources that are not typically described with IMDI metadata was that the main protagonists of CMDI are from the same group that were at the cradle of IMDI.

The complete list of resources that were selected at the two data centers can be found in table 1 and 2 in the appendix. As can be seen the resources vary greatly. For MI we selected lexical resources (with proper names), linguistic databases (with syntactical, morphological and phonological dialect variation) and ethnological databases (containing data about folktales, songs, probate inventories and pilgrimages). For INL we selected lexical resources (monolingual and bilingual lexica, historical and scientific dictionaries), corpora (spoken and written) and historical documents (bible texts). The tables also indicate whether a resource has the characteristics of typical IMDI resources. This is indicated with either a 1 for "IMDI like" or a 0 for "non-IMDI like".

4. Resource analysis

The primary goal of the CMDI is to enable the creation of adequate metadata components and profiles that have sufficient expressive power for the researcher to describe all relevant aspects of a resource. This can be a challenge when it is a new type of resource not covered by existing components yet. In this case a proper analysis of the resource and the existing associated information must be made.

4.1 Data, information and metadata

If for example we look at a resource containing speech recordings, such a resource can also contain textual transcriptions of the speech and annotations that assign POS tags to the transcriptions. Such transcriptions and annotations are examples of information that is created by interpreting the raw speech recording data and are essentially abstractions from the original data. For this reason in some systems or domains annotations are considered metadata. When more information is created at increasingly more abstract levels, information will start to touch upon, or overlap with, the metadata domain more and more. For the example resource containing speech recordings it might be decided to also add information about the subject that is being discussed in each recording, since a researcher is interested in how speaking styles differ depending on conversation subject. This information on conversation subject however could very well also function as key words at the metadata level of the resource. As such it can help make the recordings discoverable for researchers from other research disciplines that might be interested in the conversation subjects for completely different reasons, for instance because they research the telling of folk tales. Although the information on conversation subject was not intended for the metadata domain, it can be very useful metadata. Resources should be analysed for such information that can be potentially useful for metadata purposes.

In deciding what information can be used as metadata one should be led by functional criteria. Resource metadata should primarily guarantee that a resource, or subparts of it (see below), is discoverable in the CMDI infrastructure using multiple search criteria. A particular challenging issue is how we can foresee the uses of metadata by other domains than our own (linguistics). E.g. what metadata would be needed to ensure discovery of a resource by a historian?

One mistake easily made is to duplicate too much information of a resource in the metadata domain. By doing so one runs the risk of converting content to metadata descriptions. Here a functionally motivated balance must be found for how content search and metadata are dealt with.

[5] www.meertens.knaw.nl

[6] www.inl.nl

[7] The HLT Agency (www.inl.nl/tst-centrale), which is an initiative of and is funded by the Dutch Language Union, was set up to organize easy access and re-usage of LRs for the Dutch language developed with public funding.

[8] Although CMDI can also be used for creating metadata for tools and web services, in the project these were not taken into account.

[9] www.mpi.nl/IMDI/

4.2 Granularity

An analysis of the levels of granularity present in the data of a resource (if any) is also needed. A resource can be a complex resource that can be (recursively) subdivided into constituents. Think of a text corpus that can be divided into subcorpora that can again be divided into individual texts.

Especially for resources in a relational database it is not always clear what the level of granularity is. For instance, when a resource consists of recordings for 100 different locations on 10 different subjects (the syntactic phenomenon "double negation" for instance). What then is the most suitable granularity of these data? Are there 100 subcollections based on the 100 locations or are there 10 subcollections based on the 10 subjects? Neither is the "true" or "inherent" granularity for the data of this resource.

Here, again, in choosing one over the other one should be led by functional criteria. For example, should a subcollection be visible or citable? Or should one be able to transfer a certain subcollection to another repository? For the MI resources data granularity levels were assigned to the linguistic databases. For the INL resources data granularity levels were assigned to most text and speech corpora. For the other MI and INL resources no functionally motivated data granularity levels could be assigned during the project.

5. Creating metadata components and profiles

After the selected resources (see the Appendix) had been analysed metadata profiles could be created. At the start of the project, the CMDI framework offered an initial set of ready-made metadata components, some of which were (partly) derived from existing metadata sets like OLAC, TEI, IMDI and DC. This initial set was created by the CMDI project for interoperability with the huge installed base of metadata records found in the LRT world. CMDI also offers a so-called XML-Toolkit to create CMDI metadata. Using this toolkit components are created by using a standard XML editor in which schemas are used to enforce correctness and subsequently XSLT style sheets are used to create an actual CMDI metadata xsd schema. Generating instances for individual resources or subcollections needs to be done with an XML editor.[10] This method of creating metadata is very user-unfriendly since it requires knowledge of XML that most researchers do not have and the procedure is cumbersome in itself. The CLARIN project currently is working on a set of user-friendly tools for creating components and metadata descriptions that will greatly improve the usability of CMDI.

Newly created CMDI metadata components can be combined in metadata profiles that can then be used to create metadata records or descriptions for resources. Fig. 2 illustrates the hierarchical structure of corpus profiles for the JASMIN speech corpus (Cucchiarini, Driesen, Van hamme & Sanders 2008). A profile first contains very general metadata at the collection level. This is then followed by more specific metadata at the corpus and speech corpus levels. All of the metadata profiles that were created in the project consisted of such hierarchal structures.

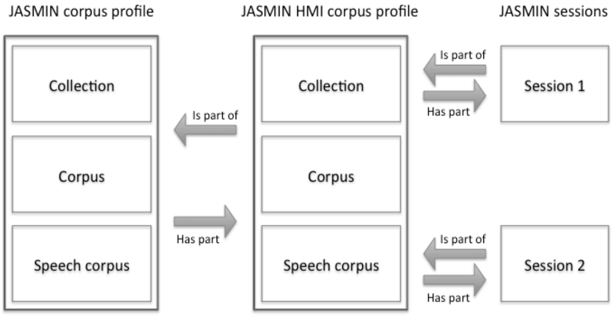

Figure 2: Hierarchy and granularity in JASMIN profiles

The different levels of data granularity of the resources are also reflected in the CMDI metadata. In the CMDI model it is possible to have a CLARIN metadata record (profile/schema instantiation) refer to other metadata records. The possibility to link metadata records to each other enables the creation of metadata instances on subcollection level and/or on the level of the individual resources. In Fig. 2 this is illustrated for the JASMIN profiles. The JASMIN resource contains two subcollections: one with read speech and one with speech for human machine interaction purposes called JASMIN HMI. The arrows depict how the metadata records for the whole JASMIN corpus and the JASMIN HMI subcollection refer to each other.

Some metadata components created in the project could be derived from the initial set of components that were derived from the existing metadata sets such as IMDI and OLAC. These components contained either very general metadata elements (e.g. Location, Language) or metadata elements that were specifically intended to describe multi-modal (IMDI type) resources. Often the existing metadata element sets were too limited or too detailed which then required adding or removing elements and/or components from the sets.

New components had to be created for the non IMDI type resources. One example is the component Headwordtype that is used for describing the headword type of a lexical resource at INL or MI. Another example is the Dimensions component that can be used for profiling a resource following the general research dimensions "time", "space" and "social". It describes for what research dimensions variation is present in the resource. For instance, it can describe that a resource contains social variation data for the social variables "religion", "age" and "gender".

[10] www.clarin.eu/toolkit

All newly introduced metadata elements had to be properly linked to existing concepts in the DCR (Data Category Registry). Each metadata element refers via a URI to exactly one data category in the ISOcat DCR, thus indicating unambiguously how the content of a metadata element should be interpreted. At the start of the project 217 data categories were available in the ISOcat DCR. Mapping onto existing ISOcat data categories, where possible, is strongly encouraged. Only if one is sure that the existing categories are not accurate enough new concepts should be added to the DCR. Examples of data categories that were newly added during the project were "Legal Owner" and "Pseudonym".

When trying to map the components onto existing data categories several issues emerged.

Sometimes the data concept definition given in the registry was too specific or too narrow. For the element "birth year" for example there was no concept "birth year" available in the DCR. There was however a concept called "birth date" which is related but not similar. We encountered the same issue for the concept "(overall) quality of the recordings of a speech corpus" for which only the related concept "quality of a recording" was available in the DCR. In these cases the decision was made to refer to the existing definitions, rather than to create new concept links. In other cases the definition of a term deviated too strongly from the definition that was envisioned when creating an element. For those concepts new data categories were created.

Another problem in relation to the DCR were double data categories (i.e. data categories with the same name and the same definition). In that case the favoured data category was the one that was already in the standardisation process. If the data categories had the same status, preference was given to the DC that belonged to the metadata thematic domain group.

6. Conclusions

In the project CMDI appeared flexible enough for creating semantic descriptions of the resources at MI and INL. We were able to create components for both IMDI and non-IMDI type of resources using CMDI. What metadata and data granularity levels to discern when making existing resources available through the CMDI infrastructure should be functionally motivated.

The CMDI approach strongly encourages reuse of existing components to avoid a proliferation of (similar) components in the component registry, but it is possible to adapt existing components to one's specific needs, e,g by adding an element. The profiles and components created during the project can be reused as well (e.g. the profile for a text corpus). Components containing very general metadata are specifically well suited for reuse (e.g. the component Language, which identifies a language). The least favoured option is to create components from scratch. But users of the component metadata infrastructure should always be aware that newly created elements have to be linked to the ISOcat data category registry.

One of the deliverables of the project is a document with best practices that will be made available through the CLARIN EU website. The components developed in this project are available at www.clarin.eu/cmd/components/clarin-nl/.

7. Acknowledgements

The authors would like to thank Jan Pieter Kunst (Meertens Institute) and Anna Aalstein (INL) for their valuable input during the project.

The project reported on in this paper was funded by CLARIN-NL (www.clarin.nl).

8. References

Barbiers, S., Cornips, L. & Kunst, J.P. (2007). The Syntactic Atlas of the Dutch Dialects: A corpus of elicited speech and text as an on-line dynamic atlas. In J.C. Beal & K.C. Corrigan & H. Moisl [red.] Creating and digitizing language corpora. Volume 1: Synchronic databases. Palgrave Macmillan, Hampshire, pp. 54-90.

Beeken, J. C. & van der Kamp, P. (2004). The Centre for Dutch Language and Speech Technology (TST Centre). In Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC), pp. 555-558.

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

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

Cucchiarini, C., Driesen, J., Van hamme, H. & Sanders, E. (2008). Recording Speech of Children, Non-Natives and Elderly People for HLT Applications: the JASMIN-CGN Corpus. In Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC).

Meder, T. (2010). From a Dutch Folktale Database towards an International Folktale Database. In: Fabula 51, Heft 1/2. Walter de Gruyter: Berlin – New York.

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

9. Appendix

Dynamic Syntactic Atlas of the Dutch	A linguistic database of elicited speech and text collected between 2000-2005 to chart the syntactic variation at the clausal level in 267 dialects of Dutch spoken in

dialects (1)	the Netherlands, Belgium and North-West France.
Dutch Database of Family Names (0)	A lexical resource containing an online dictionary and reference work for users interested in the origins, meanings and areas of distribution of surnames. At the moment the database contains 93466 names.
Dutch Database of First Names (0)	A lexical resource containing some 20.000 first names including explanations of the names and all sorts of other information on names.
Plant names in Dutch (0) dialect	A lexical resource containing the popular names of plants in the Dutch language area. The database contains more than 275.000 records and is the world largest collection with this type of information.
The Dutch Song Database (0)	An ethnological database with over 125.000 songs from the Middle Ages to the modern times. The sources are songbooks, song sheets (broadsides), song manuscripts and fieldwork recordings.
Goeman Taeldeman Van Reenen project (1)	A linguistic database of elicited speech and text collected between 1980 and 1995 to chart morphological (word-level) variation.
The Dutch Folktale Database (0)	An ethnological database that enables one to search for historical and contemporary fairy tales, legends, saints' lives, jokes, riddles and urban legends. Currently it contains 40224 stories.
Soundbites (1)	A linguistic database containing more than 1000 hours of sound recordings of dialect speakers in over 100 places in the Netherlands. These recordings were collected in the 1950s, 1960s and 1970s.
Diversity in Dutch DP Design (1)	A linguistic database with elicited speech and text collected between 2005-2009 to chart the syntactic variation at the level of nominal groups in the Netherlands, Belgium and North-West France.
Dutch Songs Online (0)	An ethnological database with song texts from the Digital Library of Dutch Literature (DBNL) merged with metadata from the Dutch Song Database.
Probate Inventories Database (0)	An ethnological database containing 2889 probate inventories from 10 places in the Netherlands dating from the 17th and 18th century.
Pilgrimage in the Netherlands (0)	An ethnological database containing data about 662 pilgrimage centres. The data are relevant for research into pilgrimage, devotions to saints, religious material culture and religion in general.

Table 1: Selected resources at MI

JASMIN speech corpus (1)	A speech corpus which contains ca. 115 hours of Dutch speech by young speakers, non-native speakers and elderly speakers living in Flanders and the Netherlands.
38 Million Words Corpus 1996 (0)	A text corpus, which consists of three main components: a varied component (1970-1995), a newspaper component (Meppeler Courant, 1992-1995) and a legal component (1814-1989).
PAROLE Corpus 2004 (0)	A text corpus, which contains modern Dutch texts (ca 20 million tokens), for the greater part originating from newspaper or magazine articles.
AUTONOMATA Spoken Names Corpus (1)	A speech corpus containing ca. 5000 read-aloud first names, surnames, street names, city names and control words. The corpus consists of a Dutch part and a Flemish part.
COREA Coreference Corpus (0)	A text corpus, which contains Dutch texts in which coreference relations were systematically marked. The corpus consists of newspaper, transcribed spoken language and lemmas from a medical encyclopedia.
Corpus of Old Dutch (0)	A (historical) text corpus, which contains all the Dutch appellative word material that originated from the period 475 - 1200.
5 Million Words Corpus 1994 (0)	A text corpus of ca. 5 million words derived from books, magazines, newspapers and TV broadcasts from the period 1970-1994.
Eindhoven Corpus (0)	A text corpus containing the VU version of the Eindhoven Corpus (also: Corpus Uit den Boogaart). It is a collection of Dutch written and (transcribed) spoken texts from the period 1960 - 1976.
Statenvertaling 1637 (0)	A historical document, which consists of the digitized texts of the bible Statenvertaling 1637.
Dutch Electronic Lexicon of Multiword Expressions (0)	A monolingual lexical resource that contains more than 5000 Dutch multiword expressions (MWEs). MWEs with the same syntactic pattern are grouped in the same equivalence class.
Woordenboek der Nederlandsche Taal (0)	A historical, scientific and descriptive dictionary of Dutch as used in 1500-1976.
e-Lex (0)	A monolingual lexical resource of Dutch consisting of a one-word lexicon and a multi-word lexicon. The one-word lexicon contains ca. 220.000 entries and more than 600.000 word forms, annotated with morphological, syntactical and phonological information.
OMBI Dutch-Arabic (0)	A bilingual lexical resource, which contains ca. 35.000 entries. Dutch is the source language and Arabic the target language..
Reference Lexicon for Belgian-Dutch (0)	A monolingual lexical resource of Belgian-Dutch containing ca. 4000 entries (lemmas). The resource contains only those words, which have a specific meaning in Belgian-Dutch or appear only in Belgian-Dutch.

Reference Lexicon for Dutch (0)	A corpus-based monolingual lexical resource of Dutch containing ca. 45.000 entries (lemmas). Apart from examples, each word has been provided with detailed linguistic information.
Dictionary of Old Dutch (0)	A scientific and historical dictionary, which contains over 2200 official documents from the thirteenth century.
Wordlist of the Dutch Language 2005, source file (0)	A monolingual lexical resource with more than 100.000 entries annotated with linguistic information. The entries are high-frequency words of Standard Dutch and/or words that may cause spelling problems.
Dutch PAROLE lexicon (0)	A monolingual lexical resource, which consists of about 20.200 entries, distributed over 13 parts of speech (POS). The entries have been described along the dimensions of morphosyntax and syntax.

Table 2: Selected resources at INL