Best Practice Guide for using CLARIN metadata components

Creating & Testing CLARIN Metadata Components

CLARIN-NL-09-001

Terminology

Metadata element or **metadata attribute**: A single item of metadata that should describe a single aspect of a resource. Usually it is also specified with a value scheme.
Value scheme: A formal specification of the range the values of an element may have e.g. a controlled vocabulary or a regular expression.
CMDI Metadata component: A bundle of metadata elements that together describe related aspects of a resource, e.g. speaker metadata.
CMDI Metadata profile: A specification of a set of metadata components that together should give an adequate description of a resource.
Metadata set: A specification of an enumeration of metadata elements.
Metadata schema: A formal specification for the format of metadata descriptions.
Metadata record or **metadata description**: a description of an actual resource with metadata. This is usually an instantiation of a metadata schema.
Resource: digital object with a specific identity that can be addressed via an identifier (e.g. a URI or a PID). N.B. A resource can also be a collection of objects that is then represented by the collections metadata.
Language Resource (LR): digital resource that provides information about languages, such as lexicographical, terminological, morpho-syntactical, corpus-related, or semantic resources or digital resources used to study linguistic phenomena like texts and multi-media/multi-modal recordings.
Complex resource: a resource consisting of multiple constituent parts, each of which can be accessed individually. A collection can be considered as a complex resource.
Collection: grouping of any number of resources that need to be referenced as a whole.

1. Introduction

CLARIN has concluded that the available metadata sets are insufficient to cover the whole domain of Language Resources (LR) because, either the sets have too much or too little specificity and also because the terminology used in the different metadata sets does not always fit with the terminology of a specific sub-domain or

community. Therefore CLARIN decided to create a metadata infrastructure based on components that would be flexible enough to allow users to create their own metadata sets by aggregating different metadata components.

CLARIN metadata components are bundles of metadata elements that describe one particular aspect of a resource or collection, for instance the biographical information about a speaker or the contact information of a publisher.

Components can contain other components and can be aggregated in profiles that then are used to form a metadata schema specifically targeted to describe (language) resources of a specific type. Researchers can either use existing or create new components to make profiles that may then be reused by others. Reuse of existing profiles and components is of course encouraged but it is the choice of the researchers to do so.

Both metadata components and profiles are stored in the CMDI component registry where others can reuse them. Existing profiles can either be reused entirely or they can be edited by adding (new) components or by deleting or adapting components. It is also possible to create completely new profiles with existing and/or new components.

The Dutch CLARIN project "Creating and using CLARIN metadata components" was the first to actually test the use of components and to try to create metadata descriptions for resources available in two Dutch language resource centers: the Institute for Dutch Lexicology (INL) and the Meertens Institute.

This "Best Practice Guide" is the result of this project. It will however in the future be extended with new experiences gained by new projects that will make use of the CMDI.

The project could draw from a prepared set of components derived from existing metadata sets for the linguistic field as IMDI and OLAC but the intention was to have the project deliver new components and profiles that would be usable for a broader spectrum of researchers (and projects) needing to create CLARIN metadata for their resources.

The project management monitored the progress and tried, to their best knowledge, to limit unnecessary variety if thought appropriate. No one will benefit from a proliferation of new components where existing ones will do. Some extra suggestions were proposed for adoption by the project management and were discussed carefully.

1. A set of combinable components of increasing specificity to describe corpora of different types e.g. collection, corpus, text corpus,… This allows to maximize possible reuse.
2. A procedure for modeling the granularity of collections, corpora by using separate metadata records for different levels of granularity. Relations

between the different metadata records are then expressed by using “partOf” and “hasPart” references in the different metadata records (see 6.1).

2. The CMDI Tools

When the CLARIN NL metadata project started, there were no special tools for creating or using the CLARIN metadata components. This means that to use the system a user had to be XML-savvy enough to edit XML files by hand (or use an XML editor like <oXygen/>) and apply an XSLT style-sheet to transform a CMDI profile specification into an XSD schema. Although currently much more user-friendly tools are available, the CMDI XML Toolkit can still be useful e.g. when a network connection is lacking. All relevant information on working with the CMDI XML-Toolkit can be found at: http://www.clarin.eu/toolkit

The current recommended way of creating CMDI components and profiles is to use the CMDI component registry and editor:
http://www.clarin.eu/cmdi/componentregistry
For editing metadata descriptions we recommend the use of ARBIL:
http://www.clarin.eu/cmdi/arbil

The component registry is a web application that works with any browser for which the flash plug-in has been installed while ARBIL is a locally installed application for which installation packages for different platforms (Windows, Mac, Linux) are offered. There is also a web-start version available that should work on all platforms that have Java installed.

3. What is a resource?

For many researchers that need to create CMDI metadata or convert existing information into CMDI it will be clear what is the resource and what is metadata. This is especially so if the existing information system already supports a clear distinction between data and metadata. However this is not always the case, e.g. a database with extensive information on Carolingian Sagas contains, from the point of view of the CLARIN metadata user both metadata that can be used for searching and data. For such information system a separation must be made between the “general” metadata that is interesting to a wide community and the remaining information that from the point of view of the CLARIN metadata user is considered a resource rather than metadata. There is a separate chapter 6.2 “Issues with the distinction between data and metadata” to help guide some decisions.

4. Available legacy metadata

If the metadata of a resource can be described using one of the linguistic "legacy" metadata sets (IMDI, OLAC, DC) existing profiles (or parts thereof) are to be used. This saves considerable time and there must be good reasons to not follow this practice.

If it is clear that the metadata set used is not sufficient to describe the resource adequately a new profile has to be created. However there must be enough time available for such a metadata curation round. Also wishes to harmonize the legacy metadata with that of another data set may be a reason to leave the "legacy" set based profile and adopt another.

5 Creating metadata

5.1 Reusing components

5.1.1 Finding existing Components and Profiles in the Registry

The CMDI approach favors reuse of existing components to avoid a proliferation of (similar) components in the component registry:

- Components that can be included more than once within the same profile and which apply to different kinds of resources, e.g. the "Language" component (which simply identifies a language). The context of that component defines the exact meaning of it, e.g. within the component "Actor-Language" "Language" identifies a language used by a speaker. And when used in the component "DocumentationLanguage" the component identifies a language in which documentation about the resource can be found.
- Components that can only be used once within a profile, but which apply to all kinds of resources, e.g. the component "Access" that stipulates how, where and under which conditions the resource can be accessed.
- Another kind of components of which reuse is strongly recommended are the standard components used to identify a certain kind of resource. e.g. the components "Collection", "Corpus", "Speech corpus", "Text corpus", "Lexical resource" and "Database".

Not all components have the same degree of reusability: some components can only be used for one specific type of resource, e.g. a component that contains technical information about a speech corpus.

5.1.2. Components strongly recommended for reuse

The project spent some time trying to find a good solution for components implementing metadata items that are widely needed when describing language resources. The reuse of these components is very strongly recommended.

a. "cmdi-language": using a language name and iso-639-3 language codes
b. "cmdi-location": Component for describing a certain location (address, region, country, continent with ISO vocabularies where appropriate)
c. "cmdi-annotationformat": list of frequently used annotation formats. Mostly mime-type codes but some "legacy" IMDI codes
d. "cmdi-mimetype": list of frequently used mime types

It was already mentioned that the CLARIN NL project came to a proposal to express the metadata distribution over multiple components that can together be used in different combinations to describe a collection or corpus of resources. Reuse of these components is recommended where possible.

a. Collection
b. Corpus
c. TextCorpus
d. SpeechCorpus

5.2. Creating new components

5.2.1 When to create new components

The components and profiles available in the component registry come from a number of sources:

1. Decomposition of "legacy" metadata sets IMDI, OLAC, DC provides a basic set of components
2. The CLARIN NL project "creating and using metadata components" has provided a number of components and profiles to describe resources at the INL and Meertens Institute, but did also try a coherent approach for some standard metadata fields used in linguistics as "Language" etc. Also standardization was attempted for "Collection", "Corpus", "TextCorpus" and "SpeechCorpus".
3. Other CLARIN NL and CLARIN EU activities have created components based on (1) and (2) and added new ones.

Although all components are intended to be reusable in the first place, it is impossible to describe all kinds of resources with the final CMDI components and profiles. Therefore new (preferably reusable) components and profiles can and should be created.

Motivations for new profiles/components:

- When the resource for which a metadata description is being composed deviates from the resources described till now, creating a complete new profile can be the right thing to do. The meaning of a component also depends on its context, so a change of this context can already necessitate a new component. Always try to allow a broad interpretation and reuse as many available components as possible. When this is not possible (e.g. when needing a profile with the components Collection/corpus/sign language corpus), a new component following a similar design pattern as with existing components should be created. From the example "sign-language corpus" should follow the same design pattern as "speech corpus" and "text corpus".
- The CMDI approach favors complete metadata description if slightly possible. So ideally whenever a metadata item is not available it has to be created in a (copy of a) suitable component. Using a copy of an existing component can be seen as using rudimentary inheritance.

5.2.2 How to create new components

When creating new components, the following needs to be taken into account.

- Reusability: when a (collection of) metadata element(s) could be used for different kinds of resources, or when the elements could be used more than once within the same profile, that (collection of) metadata element(s) is to be included in a separate, reusable component. Apart from reusability this also facilitates the creation of components: one can simply refer to the component in question instead of again adding all elements. The CMDI approach wishes to stimulate reuse in order to standardize metadata profiles as much as possible.

- Cardinality: cardinality defines how often an element/component can or must occur. Only one element is mandatory for each resource, i.e. the name. All other elements are optional. But when certain components are chosen, some of the in that component included elements (or components) can be obligatory to use. E.g. when one wishes to incorporate the component "Actor Language" - which gives information about a language spoken by a speaker - in his profile, the included "Language" component has to be filled in. Otherwise it would be impossible to identify the speaker language about which information is given.

- Linking to the Data Category Registry: each metadata element has to be linked to a data category in a DC registry (e.g. the ISOcat Datacatagory Registry www.isocat.org) but sometimes the definition given in the registry is too specific or too narrow. The element "BirthYear" for example: the concept "BirthYear" is not available in the ISOCat, but the concept "BirthDate" was. The same is the case for (overall) "quality of the recordings of a speech corpus" for which only the related concept "Quality of a recording" is available in the DC. In these cases the decision can be made to refer to the existing definitions, rather than to create new concept links, because it is assumed the DC definition can't be misunderstood. In other cases the definition of a term can deviate too strongly from the intended definition when creating an element, in that case new data categories should be created for those concepts. E.g. when creating a component "Contact" the idea is to design a very general component which can be used for contact data for persons and/or organisations who are/were involved in the distribution and the creation of the resource. But the definitions of the DC's organisation and email were too specific for this purpose. The solution was to create two new DCs, namely "Organisation" and "Person" (with the very general definitions "the name of an organisation / a person). Which person or organisation is meant is being made clear by the context (e.g element "Organisation" in the component "Contact" in the component "Creator" indicates that an organisation which was involved in the creation of the resource is intended). Another problem in relation to the DC registry can be double DC's (i.e. with the same name and the same definition). In that case the favored DC was the one that was already in the standardisation process. If the DCs had the same status, a random choice had to be made.

6. Special Issues when creating metadata

6.1 Issues with Granularity

When a description of a resource like a corpus is distributed over several metadata records e.g. corpus, sub-corpus, text file), decisions have to be made whether to duplicate metadata within the descriptions of the different metadata records. Although duplication takes care that the individual descriptions are self-sufficient, it can lead to consistency problems. The CMDI infrastructure does provide relations (links) between related metadata records e.g. between a collection and a sub collection, or between a text corpus and a single text. These relations are specified by "isPartof" and "hasPart" links that are embedded in the metadata records. Duplicated metadata will then be redundant, since searching for metadata on e.g. text level also gives access to the metadata on corpus level. There are a few

exceptions to this automatic percolation of information down from the collection metadata to individual resource metadata. For instance when describing the included languages in a multilingual text corpus (by means of the component "SubjectLanguages"). On corpus level one will indicate that e.g. the languages English, Dutch and French are included, but the individual texts most probably will only contain one language. Hence, the component "SubjectLanguages" needs to be repeated on text level. In those cases, the "lower level" text metadata overrules the "higher level" corpus metadata.

6.2. Issues with the distinction between Metadata & Data

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 for which no metadata is available yet. In those cases a resource first needs to be properly analyzed for data residing next to the raw data of a resource (audio, video, etc) that could potentially be used as metadata. The following two types of such data are rather common:

1. Very general data about the raw data that are used for data management purposes, for instance the ID of a recording.
2. Data containing interpretations of the raw data, like the orthographic transcriptions of speech in recordings or a specialized (scientific) classification of recordings. An example of the latter would be a classification based on syntactic phenomena present in recordings (agreement, double negation, etc).

We also need to distinguish between three main types of metadata, each one specific for a certain purpose: descriptive metadata, structural metadata and administrative metadata.

- Descriptive metadata describes a resource for purposes such as discovery and identification. It can include elements such as "title", "abstract", "author", and keywords.
- Structural metadata indicates how compound objects are put together, for example, how pages are ordered to form chapters.
- Administrative metadata provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it.

(NISO "Understanding Metadata". NISO Press.
http://www.niso.org/publications/press/UnderstandingMetadata.pdf).

CMDI metadata primarily used as descriptive metadata, i.e. for discovery and identification of a resource (or parts of it). So when analyzing the aforementioned

two types of data for their usefulness as metadata, this is the purpose that should guide you in deciding over what data to use as metadata. For instance, data about the location(s) of recordings in a resource could be valuable CMDI metadata for that resource, since it is plausible that a researcher searching for resources to use for his or her research is interested only in data that are bound to a specific geographic region. Data that is very specific for the corpus or database, such as for instance extensive "recording protocols" in speech corpora, are better stored as special information resources.

In principle no intrinsic distinction between the data and metadata of a resource exists. Some data can be used as metadata depending on the context that it functions in, like resource discovery or data administration. We have to look at the context wherein the metadata will be used and be pragmatic with respect to the benefits and costs. But always remember that it is expensive to create metadata long after the resource creation.

Another guiding principle in deciding what data to use as metadata is how useful the data are for researchers from other disciplines than the research discipline the resource originated from. The CMDI infrastructure encourages reuse of resources by researchers from any sub discipline in humanities or social sciences. Therefore the most valuable metadata is that that is useful to any researcher when browsing or searching for resources. Therefore it is advisable to focus more on the generic characteristics of a resource like "location of the recording" than on specific characteristics like "syntactic attribute set used for adjectives".

Once a resource has been analyzed for data that can double as metadata one needs to see what other metadata is still needed. This will have to be created from scratch.