Using TEI, CMDI and ISOcat in CLARIN-DK

Dorte Haltrup Hansen[1], Lene Offersgaard[2], Sussi Olsen[3]

University of Copenhagen, Centre for Language Technology
Njalsgade 140, DK-2300 Copenhagen
E-mail: [1]dorteh@hum.ku.dk, [2]leneo@hum.ku.dk, [3]saolsen@hum.ku.dk

Abstract

This paper presents the challenges and issues encountered in the conversion of TEI header metadata into the CMDI format. The work is carried out in the Danish research infrastructure, CLARIN-DK, in order to enable the exchange of language resources nationally as well as internationally, in particular with other partners of CLARIN ERIC. The paper describes the task of converting an existing TEI specification applied to all the text resources deposited in DK-CLARIN. During the task we have tried to reuse and share CMDI profiles and components in the CLARIN Component Registry, as well as linking the CMDI components and elements to the relevant data categories in the ISOcat Data Category Registry. The conversion of the existing metadata into the CMDI format turned out not to be a trivial task and the experience and insights gained from this work have resulted in a proposal for a work flow for future use. We also present a core TEI header metadata set.

Keywords: CMDI, TEI, metadata

1. Background

CLARIN-DK[1] is a Danish infrastructure of language resources. It is part of DigHumLab[2], a national distributed research infrastructure that integrates and promotes digital resources, tools, communities, and opportunities to Danish researchers in the humanities and social sciences At the same time CLARIN-DK is the Danish national consortium of CLARIN ERIC[3], the Common Language Resources and Technology Infrastructure, which aims at providing easy and sustainable access for scholars in the humanities and social sciences to digital language data.

CLARIN-DK is rooted in an earlier Danish infrastructure project (The Danish CLARIN project[4], 2008-2011) with focus on the detection, collection and creation of resources for researchers from the humanities. The project was a broad collaboration between 8 educational and cultural institutions, which resulted in a large amount of varied, heterogeneous language resources, written, spoken and visual, that were made available for the researchers in a digital repository.

The aim of CLARIN-DK is firstly to transform the resources of the old repository into a research infrastructure for Danish researchers from the humanities, adapting the existing resources to the new structure, and secondly to facilitate the import of new resources and to give access to a variety of language processing tools and services.

2. Motivation

By adapting the existing resources to the new infrastructure (Offersgaard et al., 2013), we want to facilitate the processing, storing, and sharing of data, sharing not only within the Danish research community but also internationally with the other CLARIN partners. The exchange of resources inside CLARIN ERIC takes place by providing the resources with harvestable CMDI metadata which is visible through the Virtual Language Observatory, VLO[5]. The text resources of the CLARIN-DK had originally metadata following the TEI standard. A common TEI text header (Asmussen, 2012) was defined and used for all text based resources of the repository. Therefore, the main goal of the conversion is to maintain all the information stored in the TEI headers of the existing resources while converting it into the CMDI format. Opposed to other projects (see e.g. Hedeland & Wörner, 2012) we have decided to convert all metadata and not only a subset of them to CMDI. In this way we simplify the complexity of the repository by working with only one metadata format for texts. Furthermore, we can share the richness of the metadata with others through e.g. CLARIN federated content search.

This document presents a tested workflow for this conversion task and gives a list of issues and challenges to consider for an optimal conversion process. It also

[1] http://info.clarin.dk
[2] http://www.dighumlab.dk
[3] http://www.clarin.eu
[4] http://dkclarin.ku.dk/english/
[5] http://www.clarin.eu/vlo/

provides some suggestions for the developers of CMDI and ISOcat that we hope can be taken into consideration.

3. TEI, CMDI and ISOcat

Before going into details, we will briefly introduce the standards TEI, CMDI and ISOcat that are in play in the conversion of metadata for the text resources in CLARIN-DK.

TEI P5[6] is a standard for representation of texts in digital form. It specifies syntax and semantics for metadata and text in a very flexible way, allowing for a wealth of more or less fine-grained information. This flexibility gives an enormous expressive power which on the other hand can make it difficult to agree upon one common set of metadata across different projects.

CMDI, the Common Metadata Initiative, was initiated in CLARIN in order to develop a standard for flexible structuring of metadata for language based resources. CMDI (Broeder et al., 2012) provides a framework for creation and use of metadata structuring schemes in which smaller metadata parts, components, can be grouped together into a resource description, a profile, which is expressed in an XML-file and scheme. The creation of the CMDI components and profiles is done through the CLARIN Component Registry[7] and to assure semantic interoperability the component elements must be linked to data categories in the Data Category Registry [8]. The possibility of reuse of the CMDI components in different profiles makes the framework suitable for describing various resource types.

The Data Category Registry (DCR) is an implementation of the ISO 12620 (ISOcat) that provides a framework for defining persistent concept definitions. Besides being used for metadata category definition in CMDI, it is also used for defining linguistic concepts ranging from morphosyntax and terminology to sign language and audio.

4. Workflow

The conversion of metadata from one format to another can be a “simple” mapping task. Transforming TEI metadata into CMDI is of another nature especially when no existing CMDI-TEI profile fits the needs. As a starting point, CMDI is only a shell that defines and limits the structural format of metadata. In the Component Registry you can either reuse already existing CMDI components and profiles or you can define new ones, and therefore either map to existing structures and content, or model the structure of your metadata set from scratch.

In CLARIN-DK the repository already contained app. 40.000 text resources compliant to a TEI specification agreed upon by two university institutions and two cultural institutions. The conversion task from TEI to CMDI was therefore set in a strict framework.

Based on our experience the following steps for a general workflow of the conversion of existing metadata to CMDI are proposed:

1. Analysis of own (TEI) metadata specification in order to structure the metadata into attributes, elements and components in accordance with CMDI.
2. Analysis of existing public CMDI profiles and components in the CLARIN Component Registry. These must of course cover the same type of resources as the ones in focus, in this case the TEI header.
3. Use of the SMC browser[9] to inspect the structure of potential profiles and components.
4. Decision of whether an existing profile can accommodate all needs, or a new profile must be created, either based on existing components or by creating new ones
5. Creation of ISOcat references for the elements and components needed, if not already defined.
6. Creation of needed components starting with the innermost nested ones.
7. Definition of the profile and download of its XML-schema.
8. Transformation of the original (TEI) metadata into the new (CMDI-TEI) format and validation of that with the created XML-schema.
9. Publication of the profile and components in the CLARIN Component Registry to allow others to reuse it.

To fertilise and ease the reuse of created components and profiles, the attributes, elements and components should be created with as loose bounds (0 – unbounded) as are allowed by the standard (TEI).

We suggest prioritizing the reuse of existing profiles and components. This will save the users for the tedious task of defining components and definitions, but also make it easier to understand and exchange metadata from different metadata providers.

Unfortunately, we had to conclude that no existing components completely fitted the structure in our TEI header specification. An already existing CMDI-TEI header profile contained 33 elements in 15 components whereas our TEI header structure needed 85 elements in 50 components. Not only were components needed at a higher level, e.g. the *teiHeader/encodingDesc* describing the relations to the source of the electronic text, but also at a deeply nested level our TEI specification differed from the existing CMDI-TEI header profile. This made it very difficult to just add new components to the existing profile.

In total we created 36 new components with 133 new ISOcat references for the CMDI-TEI header. Of these 5 were modifications of already existing components

[6] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/

[7] http://catalog.clarin.eu/ds/ComponentRegistry

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

[9] http://clarin.oeaw.ac.at/smc-browser/

which were provided with extra elements whereas the rest were created from scratch.

5. Issues and Challenges

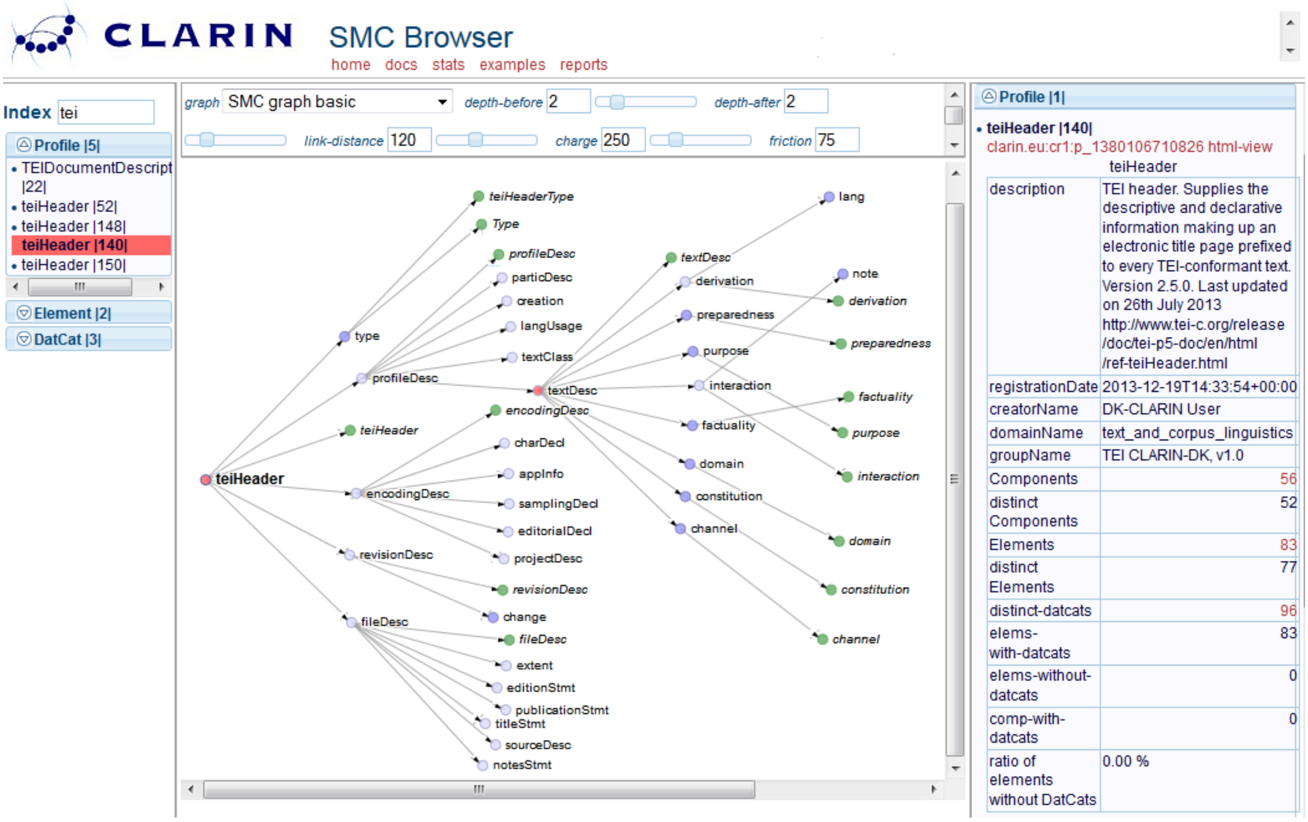
5.1 CMDI issues

When creating a CMDI profile from an existing metadata scheme, challenges arise both when mapping the existing structure onto the CMDI structure, when trying to reuse existing CMDI profiles and components and when trying to create a common usable CMDI profile. These challenges are discussed in following sections.

5.1.1 Re-use of components

CLARIN-DK cooperates with other research communities on a shared CMDI-TEI header profile but it is not as straightforward as expected. A CMDI-TEI profile along with various components has, as mentioned previously, already been defined and published in the CLARIN Component Registry but since public profiles (and public components) cannot be changed, the existing public profiles cannot be modified in order to account for our specific requirements. If extra elements or components are required, a new version of the profile including these features should be made, but if a slightly different structure is needed, then the profile isn’t backward compatible any longer and it is not a new version of the profile but an entirely new profile. Currently CMDI does not have a facility to handle versioning.

The work on a shared profile must therefore take place before profiles and components are published and still are private to anyone but the user who created them. To accommodate the discovery and analysis of both published and unpublished profiles and components that might be reusable, the recently created SMC browser is a very useful tool (Ďurčo, 2013).



Index tei

Profile |5|
- TEIDocumentDescript |22|
- teiHeader |52|
- teiHeader |148|
- teiHeader |140|
- teiHeader |150|

Element |2|

DatCat |3|

graph SMC graph basic depth-before 2 depth-after 2

link-distance 120 charge 250 friction 75

Profile |1|

- teiHeader |140|
  clarin.eu:cr1:p_1380106710826 html-view

teiHeader

description	TEI header. Supplies the descriptive and declarative information making up an electronic title page prefixed to every TEI-conformant text. Version 2.5.0. Last updated on 26th July 2013 http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-teiHeader.html
registrationDate	2013-12-19T14:33:54+00:00
creatorName	DK-CLARIN User
domainName	text_and_corpus_linguistics
groupName	TEI CLARIN-DK, v1.0
Components	56
distinct Components	52
Elements	83
distinct Elements	77
distinct-datcats	96
elems-with-datcats	83
elems-without-datcats	0
comp-with-datcats	0
ratio of elements without DatCats	0.00 %

Figure 1: The defined CMDI-TEI header viewed through the SMC browser

In the SMC browser, the user can visualize a selected CMDI profile as a tree. In the visualization the profile is shown as the root node, the used components are shown as intermediate nodes and elements or data categories are shown as leaf nodes. Figure 1 shows our CMDI-TEI header profile in the SMC browser. To ease the readability of the figure only the component *textDesc* is extended into details. The green bullets refer to ISOcat references, blue refer to elements, and grey refer to components. The user interface is easily configurable with a number of parameters[10], and gives info and statistics about the current profile.

The SMC browser also facilitates that the user can visualize and compare different profiles at the same time, where overlapping nodes of the profiles show shared components or elements. Figure 2 shows the overlap and differences between two TEI header profiles.

[10] http://clarin.oeaw.ac.at/smc-browser/docs/userdocs.html

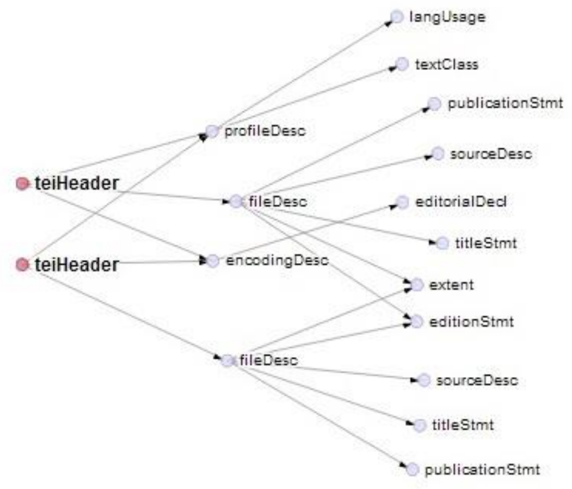


Figure 2: Overlap of two different CMDI-TEI headers viewed through the SMC browser

We provided the researchers[11] developing the browser with the CMDI identifier of our unpublished profile after which they included the profile into the browser. We were then able to explore both the overlap of components between different profiles and the completeness of ISOcat references of the profile.

5.1.2 Simple values in CMDI

Problem with structural differences may occur when converting existing metadata into CMDI. In TEI elements can contain attributes as well as other elements, but the structure in CMDI must be divided into components and elements where only elements can contain string values and only components can contain other elements and components. The difference in the two standards might introduce new levels in the structures, here illustrated by the TEI element *name* which in the TEI standard can contain other elements such as *date* and *email*:

<name> Center for Language Technology

<date when=”2011”/>

<email>cst@hum.ku.dk</email>

</name>

In CMDI this must be modelled as:

<cmd:name>

<cmd:name> Center for Language Technology**</cmd:name >**

<cmd:date when=”2011”/>

<cmd:email>cst@hum.ku.dk</cmd:email>

</cmd:name>

which introduces a *name element* in the *name component* resulting in a new structure.

5.1.3 Fixed structure in CMDI

The defined structure in the CMDI component requires that all defined elements must be structurally placed before the rest of the content of the component. This can result in another ordering of child nodes than stated in the examples in the TEI header standard specification.

Furthermore, the TEI standard has a bracket-nested alternating syntax for expressing allowed sub-structures. This cannot be modelled in CMDI as the CLARIN Component Registry only allows for optional elements in a specified order, but does not give the option to specify a number of alternating elements in the same position. This can be illustrated by part of the declaration for the TEI element *notesStmt*:

(model.noteLike | relatedItem)+

which means that one of the two elements *model.noteLike* or *relatedItem* must be present in the structure. The arity (0-n or 1-n) on each of the elements in CMDI does not give the same result.

For all the above mentioned reasons a valid CMDI-TEI header cannot necessarily be converted backward to a valid TEI header.

5.1.4 Attributes in CMDI

TEI operates with global attributes and attributes specific to a particular module of which some have been grouped in attribute classes. The eight global attributes (including xml:id and xml:lang) are optional and can be applied to all modules. CMDI does not facilitate the use of user defined global attributes nor allow for using the xml namespace as defined in the global TEI attributes .

The locally specified attributes in TEI can be either mandatory or optional with either a fixed list of legal values or free values. E.g. the TEI element *application* has two mandatory attributes *@ident* and *@version* with no specified values plus a list of optional attributes. The element *title* has a list of optional attributes, of which one, *@level*, has a list of fixed values.

In the current version of the CMDI standard there are no means to express the requiredness of attributes. This is of course a problem when having to create CMDI metadata from an existing (TEI) metadata scheme where some attributes are defined as mandatory as e.g. *date@when*:

<cmd:date when=”2011”/>

where the actual value of the element *date* is inside the attribute *when*. The former mentioned attributes *application@ident* and *application@version* which are defined as mandatory from the TEI standard raise the same problem. As we see it, the only way of dealing with this problem is to validate the resources with a stricter scheme.

A new version of CMDI (CMDI 1.2) will, however, soon be available and to our knowledge one of the changes will be to insert requiredness on attributes, thus making it easier to transfer the TEI structure to CMDI.

[11] Our contact person is Matej Ďurčo from the CLARIN Center Vienna

5.2 Defining ISOcat categories

As a consequence of the collaborative work on the CMDI-TEI header, CLARIN-DK joined the ISOcat TEI header group to have access to the existing unpublished ISOcat definitions and share the ones we defined with the rest of the group. Since TEI is very well-defined and since ISOcat does not require structural information, creating a common set of persistent TEI definitions is fairly easy. The independence of structure gives the freedom to use the same DC in different CMDI profiles or even in different metadata sets outside CMDI. All components, elements and attributes in the defined TEI header profile have links to definitions in ISOcat. The reason for making these “new” definitions instead of just pointing at the TEI standard is that IDs in ISOcat are persistent while the TEI standard might change over time.

Working in the DCR is generally very simple but it can still be difficult to find the right DCs to use. Even though the DCs are ordered in thematic groups it is not clear which ones are the most reliable and which ones fit best to the user’s point of departure. An example is the category *language* defined 4 times as *language* and more than 30 times in some sort of combination such as *language name.* For the TEI header we decided to use the exact element names and definitions as stated in the TEI standard which made the retrieval of useable DC less complex. For other not so well-defined tasks it can however be difficult to choose the best category. In the longer term it could be nice to have a facility showing how widely a DC is used and accepted for example by linking to projects, resources or schemes (e.g. CMDI profiles) outside DCR. This might give an indication of the quality of the DCs.

5.3 Common TEI header profile

A general TEI header profile would be preferable, but it doesn’t seem to be feasible as researchers have different points of departure when creating language resources - including different needs regarding metadata. Furthermore the TEI header can in principle contain an infinite number of nestings and since it allows for the same elements to occur in different places in the structure, it is in practice impossible to generate one universal TEI header profile to be used for all purposes. After a discussion of defining a general TEI header in CMDI among the CLARIN partners, CLARIN-DK took up the challenge of merging the metadata from our TEI header with the TEI header of the Austrian CLARIN Centre, into a new CMDI-TEI header covering the specifications from both institutions. More information on TEI to CMDI conversion can be found in (Mörth & Ďurčo, 2013).

5.4 A Core TEI metadata set

In this section we discuss the challenges to define a broad acceptable interoperable core of metadata for text resources.

As point of departure, inspection of the TEI standard shows that it only requires three pieces of mandatory information:

```
<teiHeader>
  <fileDesc>
      <titleStmt>
          <title><!--title of the resource--></title>
      </titleStmt>
      <publicationStmt>
          <p>Information about distribution</p>
      </publicationStmt>
      <sourceDesc>
          <p> Information about source from which it
          derives</p>
      </sourceDesc>
  </fileDesc>
</teiHeader>
```

Evidently, the TEI standard itself does not suggest a metadata core sufficient for interoperability.

Another very used metadata standard for text resources is the OLAC standard[12] with the following core elements: *contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, type*. The OLAC standard inspired the choice of mandatory metadata elements in CLARIN-DK, but the entire CLARIN-DK metadata set is expressed in TEI since it includes a broad diversity of metadata for special text characteristics. This richness of metadata is needed to describe many different types of text resources from different sources, including old manuscripts, modern text and texts from specific subject domains.

The mandatory metadata in the CLARIN-DK CMDI-TEI header (expressed as xpaths) are:
fileDesc/titleStmt/title (*title*),
fileDesc/titleStmt/respStmt/name (*responsible*),
titleStmt/respStmt/name/note (*capture method*),
titleStmt/respStmt/name/date@when (*capture date*),
fileDesc/extent/num (*size*),
fileDesc/publicationStmt/distributor (*dataprovider*),
fileDesc/publicationStmt/availability@status (*rights*),
fileDesc/publicationStm/availability/ab@type (*rights*),
fileDesc/noteStmt/note (*description*),
fileDesc/sourceDesc/biblStruct/idno@type (*filename*),
fileDesc/sourceDesc/biblStruct/analytic/author/name (*creator*),
fileDesc/sourceDesc/biblStruct/analytic/respStmt/resp (*publisher*),
fileDesc/sourceDesc/biblStruct/monogr/title (source *title*),
profileDesc/creation/date (*creationDate*),
profileDesc/language/language (*language*)

The experience gained from creating metadata for different types of text resources is that the diversity of sources and types of texts calls for very different kind of metadata. Furthermore, the researchers creating the metadata might have very different points of view on the importance of the different metadata. We therefore suggest that the research community explores the possibilities of reusing CMDI components and elements when defining CMDI profiles, and participates in the discussion on a larger set of obligatory metadata.

[12] Open Language Archives Community metadata standard, http://www.language-archives.org/OLAC/metadata.html

VLO [13] (Van Uytvanck, 2012) is a source for the investigation of the diversity of metadata for text resources. It harvests metadata from a large number of repositories with different types of resources using the OAI-PMH protocol and enables facetted browsing using only a few metadata. Selecting the value “CLARIN-DK” for the facet “NATIONALPROJECT” everyone is able to inspect CMDI TEI metadata from the CLARIN-DK repository. VLO, however, uses other labels for metadata e.g. *collection, continent, country, dataprovider, format, genre, keywords, modality, nationalproject, organization, resourceclass, subject*. A broadly acceptable interoperable core of metadata for text resources could also here ease the search and retrieval of resources.

6. Conclusion

In this paper we have outlined a workflow for the conversion of existing metadata (in this case TEI) to the metadata framework standard CMDI. The workflow was used for developing a common TEI header profile covering TEI resources of two CLARIN repositories. The profile is now public and available in the CLARIN Component Registry with ISOcat references on all elements. We encountered some structural challenges converting existing TEI metadata, that it was not possible to transfer the power of expression of the TEI standard directly to CMDI syntax. More important we experienced that it is very difficult to create and share general profiles, components and data categories since the focus and use will differ from project to project. We believe, however, that creating sharable components, profiles and working together in focused groups (e.g. with a common interest in creating and sharing a profile for a TEI header) is a very fruitful way to interoperable repositories and infrastructures. Along this path, we suggest a core TEI metadata set as a point of departure for the discussion on a larger set of obligatory metadata.

It is our hope that developers of language resources in future will use the already existing CMDI profiles and components as their metadata starting point even before creating new resources.

7. Acknowledgements

We want to thank Menzo Windhouwer from CLARIN-NL and Matej Ďurčo from the CLARIN Center Vienna for cooperation on the common TEI header profile in CMDI.

The TEI header definitions in ISOcat were made as cooperation between Menzo Windhouwer, Matej Ďurčo and the CLARIN-DK-UCPH group (consisting of Lene Offersgaard, Sussi Olsen and Dorte Haltrup Hansen). The definition of a common CMDI-TEI header was made in cooperation with Matej Ďurčo from the CLARIN Center Vienna.

8. References

Asmussen, J. (2012): *Text metadata - What the header of a text item looks like*. Technical Report, DK-CLARIN WP 2.1, Copenhagen.

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

Ďurčo, Matej (2013). SMC Browser. CLARIN Annual Meeting, Prague, Czech Republic. https://www.clarin.eu/sites/default/files/SMC_Browser.pdf

Hedeland, H., Wörner, K. (2012): Experiences and Problems creating a CMDI profile from an existing Metadata Schema. In the proceedings of the workshop *Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR*. LREC 2012, Istanbul, Turkey.

Mörth, K., Ďurčo M. (2013): *CMDI & TEI, TEI & CMDI*. Presentation at CLARIN and TEI workshop, Rome,2013-09-30. http://www.clarin.eu/sites/default/files/DurcoMoerth_0.pdf

Offersgaard, L., Jongejan, B., Seaton, M., Haltrup Hansen, D. (2013). CLARIN-DK – status and challenges. *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013, NEALT Proceedings Series 16*

Van Uytvanck, D., Stehouwer, H., & Lampen, L. (2012). Semantic metadata mapping in practice: The Virtual Language Observatory. In *Proceedings of the 8th Conference on International Language Resources and Evaluation.* LREC 2012, Istanbul, Turkey.

[13] http://catalog.clarin.eu/vlo