Facilitating Metadata Interoperability in CLARIN-DK

Lene Offersgaard, Dorte Haltrup Hansen
University of Copenhagen, Centre for Language Technology, NFI
Njalsgade 140, DK-2300 Copenhagen, Denmark
E-mail: leneo@hum.ku.dk, dorteh@hum.ku.dk

Abstract

The issue for CLARIN archives at the metadata level is to facilitate the user's possibility to describe their data, even with their own standard, and at the same time make these metadata meaningful for a variety of users with a variety of resource types, and ensure that the metadata are useful for search across all resources both at the national and at the European level. We see that different people from different research communities fill in the metadata in different ways even though the metadata was defined and documented. This has impacted when the metadata are harvested and displayed in different environments. A loss of information is at stake.

In this paper we view the challenges of ensuring metadata interoperability through examples of propagation of metadata values from the CLARIN-DK archive to the VLO. We see that the CLARIN Community in many ways support interoperability, but argue that agreeing upon standards, making clear definitions of the semantics of the metadata and their content is inevitable for the interoperability to work successfully. The key points are clear and freely available definitions, accessible documentation and easily usable facilities and guidelines for the metadata creators.

Keywords: metadata, interoperability, CLARIN, VLO

1. Introduction

The goal for the CLARIN-DK infrastructure is to enable researchers to share and reuse language-based resources and to facilitate services that ease the research in these resources. CLARIN-DK therefore sees it as essential to provide a framework for easy storage, easy metadata assignment and easy retrieval. Compared to traditional archives that handle homogeneous materials, CLARIN-DK handles a wide range of resource types from different research areas. It currently consists of an archive with language-based resources in the form of single texts, text corpora, video, audio, lexica and web services for corpus search and processing of textual data.

CLARIN-DK is both an integral part of the Danish research infrastructure for Digital Humanities, DIGHUMLAB, and member of the pan-European CLARIN ERIC. As part of the CLARIN infrastructure, CLARIN-DK can be described as a bottom-up receiver of heterogeneous material from different research areas in Denmark and propagating metadata description of these to the European CLARIN level, giving researchers the options to find new data and re-use them in a top–down approach.

To enable the discovery of resources, use of standards for resource formats and metadata are crucial. In some sense everybody can agree upon the necessity of using standards in archives and research infrastructures but heterogeneous material with heterogeneous metadata can be difficult to facilitate in fixed metadata sets, and it can be difficult even to agree upon the metadata values to be used.

In this paper we will view the quality aspect of metadata (henceforth MD) in the light of interoperability for both bottom-up and top-down approaches, and show how ensuring interoperability can be implemented and how it has more challenging aspects than sharing the use of standards.

2. The use of CLARIN-DK

From user surveys and user projects CLARIN-DK has experienced that tools and workflows are often created for very specific purposes and areas, and therefore are difficult to share without a big adaptation effort. While empirical data in form of text, video or audio files can be reused and object to new research questions (see Henriksen et al., 2014).

In CLARIN-DK we aim to provide a service that all researchers working with language-based data can use, and furthermore letting the researchers deposit their data themselves. We recognize that the users have different needs regarding the use of the CLARIN infrastructure. For some the preservation of their lifelong work is essential, for others the storage of empirical data after the end of a project or to share data with research colleagues is important. Some have a need to retrieve a large amount of data for training of statistical models, while others search for specific text genre to find evidence for linguistic variation.

Depositing data in CLARIN-DK should not be an overwhelming job preventing the users from doing so. The primary data might be in a format used in a specific project or in a specific academic field or they might be in plain Word or PDF document formats. If CLARIN-DK demands data in a specific format e.g. TEI-compliant XML, some users might refrain from depositing the data at all. But on the other hand fixed standards and interoperability is a necessity if we want to share, reuse or further develop e.g. corpora and annotations and make them potable across different hardware and software platforms (see a discussion in Simons, 2014).

The issue for CLARIN-DK at this level is to facilitate the user's possibility to describe their data with MD, even with their own MD standard, and at the same time make these MD meaningful for a variety of users with a variety of resource types and by this ensure that the MD are useful for search across all resources.

3. The semantics of MD, a TEI case

“At a superficial glance, the major problems of metadata harmonization seem to relate to formats … The problem instead lies on another level, in the interpretation or semantics of the metadata expressions” (Nilsson, 2010, pp 29)

The quality of MD is strongly connected to the availability of clear definitions of the semantics of the MD elements. The TEI[1] standard and the CMDI[2] format (stored in the Component Registry[3]) are frameworks that lend themselves to this task. In TEI clear definitions are provided for each MD element, but the syntax allows for options. In CMDI the syntax of a MD set for a given resource type can be defined with reference to the semantics of each element in CCR [4] (CLARIN Component Registry).

The CLARIN-DK group together with researchers from other cultural institutions in Denmark selected a common subpart of TEI to describe MD for single texts. This set covered text types from annual reports, newspaper articles, press releases and web blogs to the very first printed books in Denmark.

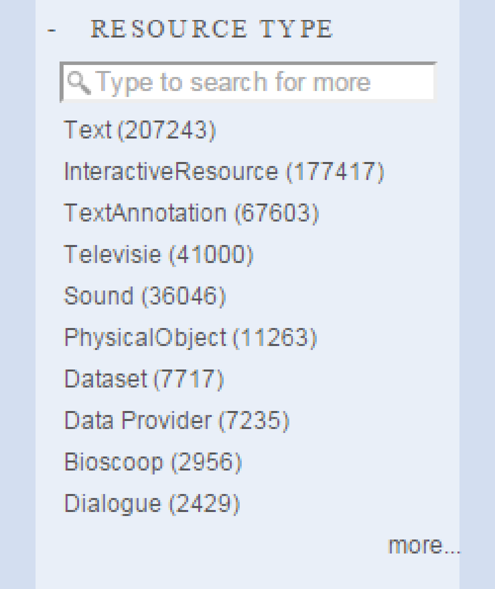
In this first version of the scheme all elements were mandatory (Asmussen, 2012). This led to users just filling in e.g. “n/a” for open text values (Offersgaard et al., 2013). We therefore decided to make a less strict scheme when doing the implementation of the TEI-header scheme in CMDI. The current scheme has only 15 mandatory elements (Hansen et al., 2014) as opposed to the 101 elements in the former scheme. In addition, we extended the scheme (and the corresponding CMDI profile) with a new module for historical manuscripts. This work was made in collaboration with first CLARIN Center Vienna (Mörth and Ďurčo, 2013) and later CLARINO for extra extensions to broaden the user group for the schema. The result is a CMDI TEI-header profile[5] with strictly defined syntax and semantics for single texts that can be used by the CLARIN community. It gives room for those who wants to give an extensive description of their single text data and allows those who only need to make a brief description to do so. For the conversion from the well-known TEI standard to a CMDI compliant MD format a local XSLT transformation was developed. This ensures that the researchers can use the TEI standard to express their MD as usual, but now with the option for easy conversion to CMDI, which is the required MD format in the CLARIN infrastructure.

Even though the syntax of this TEI header was agreed upon by several scholars and MD curators, and the definitions of the MD elements were given by TEI and documented in the CMDI Component Registry with links to relevant definitions in CCR, the semantics of the elements are not always clear and unambiguous since the syntax plays an important role in the semantics. This will be exemplified in the next section.

4. Aggregation of MD at the European level, VLO

We see that different people from different research communities fill in the MD in different ways even though the MD was defined and documented. This has impacted when the MD are harvested and displayed in different environments. A loss of information is at stake.

An example from the VLO[6] is the values of the element “subject” for text resources. All CLARIN-DK text resources currently use the DK5 taxonomy, which is the structuring principle in all public libraries in Denmark. The different classes in the taxonomy have Danish names e.g. *landbrug (dk5-631)* meaning agriculture. This way of viewing the world goes somehow hand in hand with resources using a term like *modernism (art)* referring to a class in the Library of Congress Subject Heading taxonomy. But other harvested values like *child language development* or *morphology* show a different views on “subject” category, as these are research specific subjects. We believe that different perspectives on “subject” will inevitably be present in the VLO since resources from various fields are aggregated, but MD creators should be encouraged to use standardised taxonomies.



- RESOURCE TYPE

Type to search for more

Text (207243)
InteractiveResource (177417)
TextAnnotation (67603)
Televisie (41000)
Sound (36046)
PhysicalObject (11263)
Dataset (7717)
Data Provider (7235)
Bioscoop (2956)
Dialogue (2429)

more...

Figure 1: VLO values of the search facet “resource type”

[1] The Text Encoding Initiative (TEI): http://www.tei-c.org/index.xml

[2] CMDI Component Metadata Infrastructure

[3] CLARIN Component Registry: https://catalog.clarin.eu/ds/ComponentRegistry/#

[4] CLARIN Concept Registry: https://www.clarin.eu/ccr

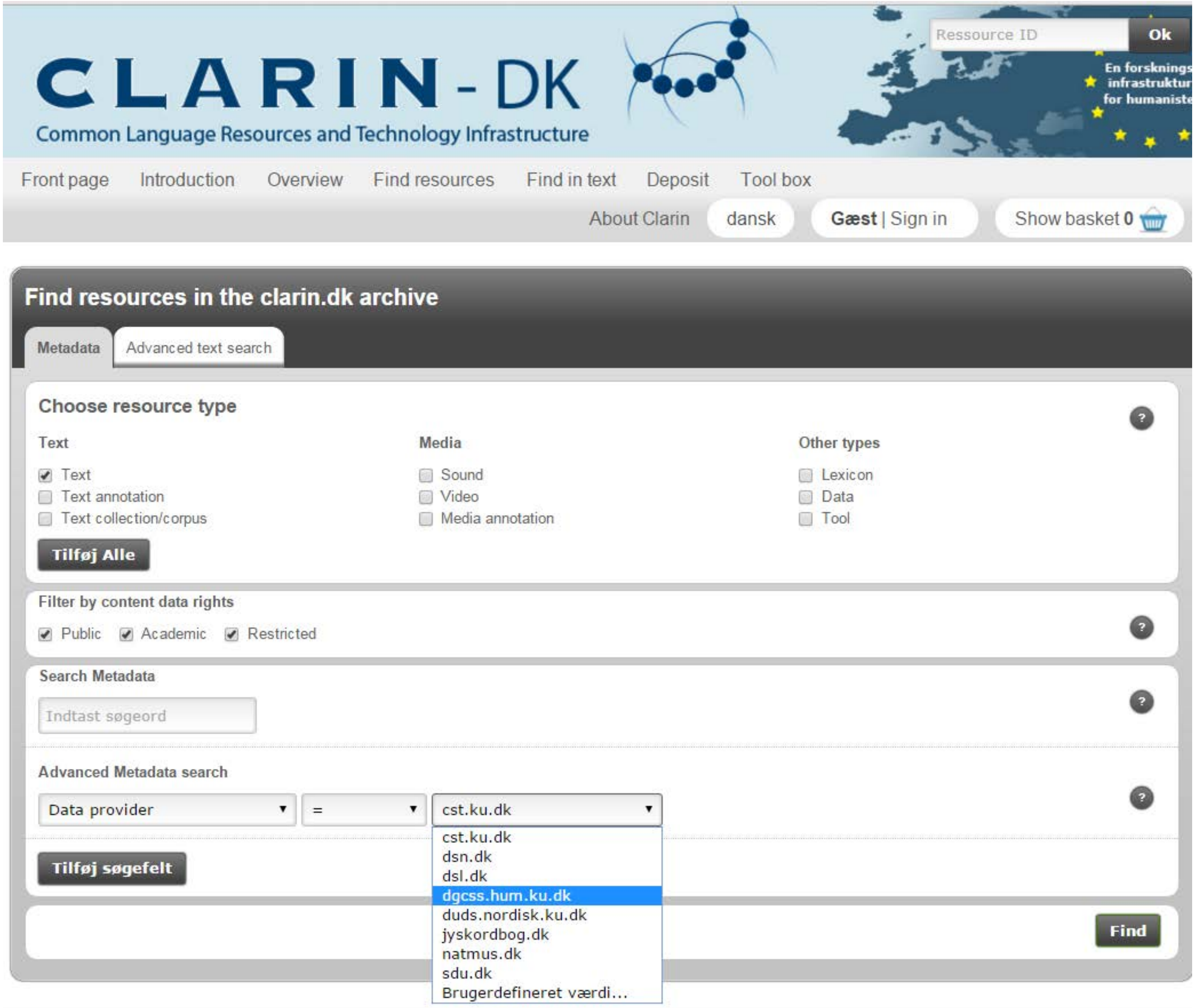
[5] https://catalog.clarin.eu/ds/ComponentRegistry/#/?registrySpace%20=published&itemId=clarin.eu:cr1:p_1380106710826&_k=vdldot

[6] CLARIN virtual language observatory: http://catalog.clarin.eu/vlo

"Resource Type" is another problematic element. In the VLO values like *Text* and *Sound* are shown as well as *Data Provider*, *Televisie* and *Bioscoop*, again showing different views on the term "Resource Type" from the MD creators. Ideally only standardised MIME types would appear in "Resource Type" and other values mapped to e.g. "genre" or "subgenre".

The values of "Data Provider" and "Organisation" are other cases with very high diversity. "Organisation" in the VLO is defined as: "*The organisation currently responsible for the resource or tool*" and has many values such as: "*Australian National University*", "*Wikisource*", "*The Danish eHealth Portal, Copenhagen, Denmark*". In CLARIN-DK single texts encoded in TEI map the element *distributor* (in the TEI structure: fileDescr.publicationStmt.distributor.name) to the search facet *Data Provider* at the Danish search interface (see Figure 2). In TEI the distributor element is defined as: <***distributor***> *supplies the name of a person or other agency responsible for the distribution of a text.* This element could be mapped to the VLO organisation facet but so could the element *publisher* (in the TEI structure: fileDescr.sourceDesc.biblStruct.monogr.imprint.publisher), defined in TEI as: <***publisher***> *provides the name of the organization responsible for the publication or distribution of a bibliographic item.*

This example shows that there can be different perspectives on the MD semantics depending on which actor fill in the MD, and it shows that it can be difficult to fill in the right values when creating the MD, especially if you do not have access to the definitions at the aggregating site. Furthermore, it shows that definitions of MD elements independent of the syntax might not be enough, since the internal structuring of the MD can contribute to the semantics. Finally, it shows that the right mapping of elements at the various levels is crucial.



Find resources in the clarin.dk archive

Choose resource type

Text
☑ Text
□ Text annotation
□ Text collection/corpus

Media
□ Sound
□ Video
□ Media annotation

Other types
□ Lexicon
□ Data
□ Tool

Filter by content data rights
☑ Public ☑ Academic ☑ Restricted

Search Metadata
Indtast søgeord

Advanced Metadata search
Data provider ▾ = ▾ cst.ku.dk ▾

cst.ku.dk
dsn.dk
dsl.dk
dgcss.hum.ku.dk
duds.nordisk.ku.dk
jyskordbog.dk
natmus.dk
sdu.dk
Brugerdefineret værdi…

Figure 2: CLARIN-DK values of the search facet "Data provider"

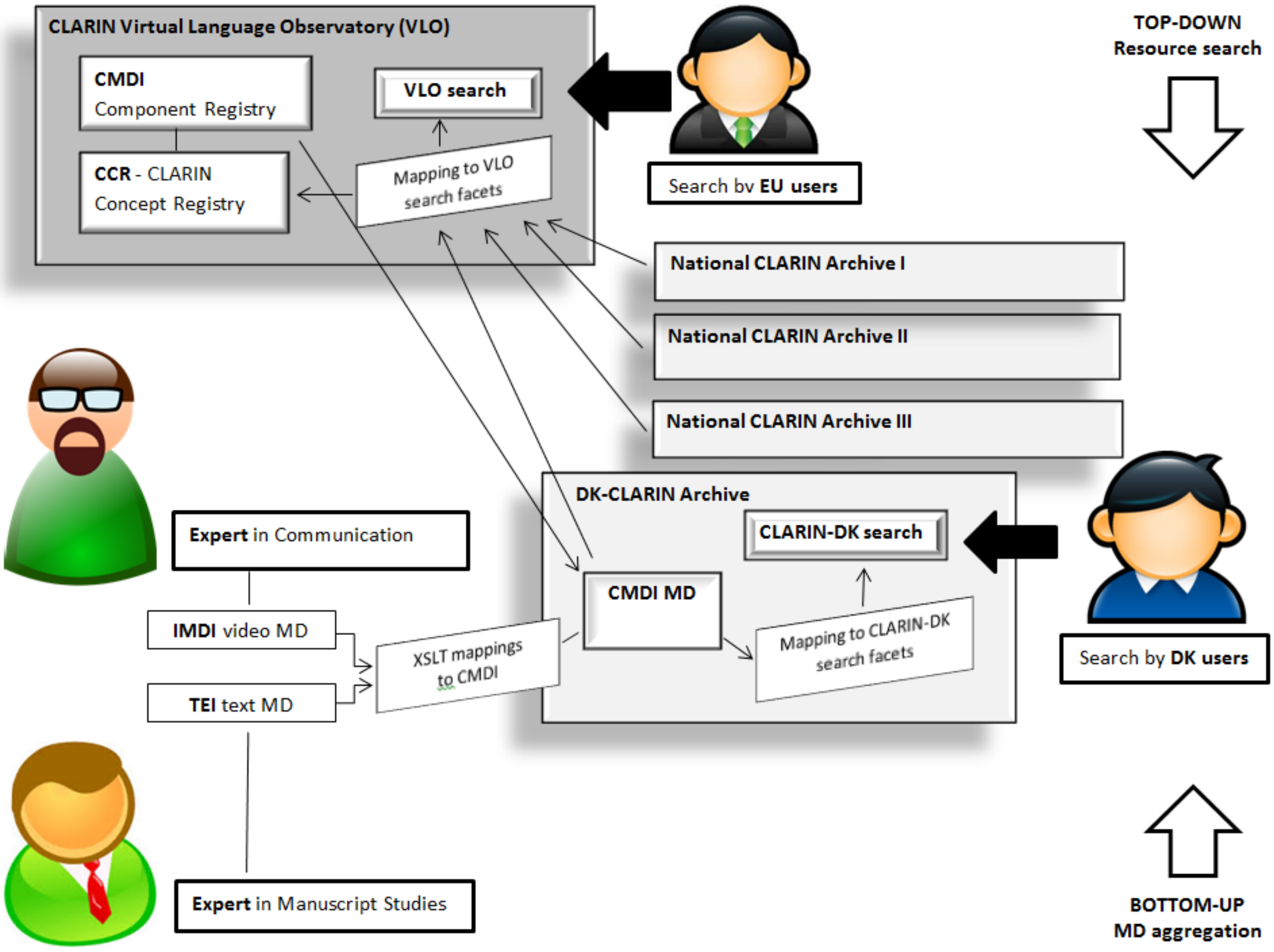

CLARIN Virtual Language Observatory (VLO)
CMDI Component Registry
VLO search
CCR - CLARIN Concept Registry
Mapping to VLO search facets
Search bv EU users
TOP-DOWN Resource search
National CLARIN Archive I
National CLARIN Archive II
National CLARIN Archive III
DK-CLARIN Archive
CLARIN-DK search
CMDI MD
Mapping to CLARIN-DK search facets
Search by DK users
Expert in Communication
IMDI video MD
TEI text MD
XSLT mappings to CMDI
Expert in Manuscript Studies
BOTTOM-UP MD aggregation

Figure 3: Agents with different roles interacting with the repositories.

Figure 3 shows the bottom-up aggregation of MD and the top-down search to access the MD. A mapping to CMDI MD is done for all new MD that is provided in formats known in the research communities like TEI and IMDI. The figure also shows the two mappings done to convert and present MD for the users searching either the Danish repository or the VLO.

An important issue for improving the MD interoperability is therefore to specify the mapping of the harvested MD carefully; and to discuss these mappings with both archives and data provider institutions. If mappings are documented and easily viewable, it will be easier for data providers to understand the importance and complexity of agreeing on semantics for MD. When harvesting of MD is enabled, a mapping of MD to search facets is imposed by the harvester. Currently the CLARIN infrastructure is working on smoothing this procedure.

5. Quality of MD in the light of interoperability

Ensuring MD interoperability casts a special light on quality criteria. In this section we look into how quality criteria for MD can support interoperability and what challenges have to be handled.

When discussing quality criteria for MD Bruce and Hillmann's seven quality criteria are often referenced: *completeness*, *provenance*, *accuracy*, *conformance to expectations*, *logical consistency and coherence*, *timeliness* and *accessibility* (Bruce and Hillman, 2004). Bruce and Hillmann found that the major issues for improving the MD quality was the development of standards and documentation for MD. As mentioned above a number of initiatives have been taken by the CLARIN Community in the direction of handling interoperability of MD: The development of the VLO harvesting that maps MD from a number of archives to a central facetted search interface, the development of CMDI framework and specification for CMDI schemas, the CMDI Component Registry and most recently the CCR (taking over from the deprecated ISOcat Data Category Registry) offers tools for documentation and supports use of standards.

Some investigations of MD quality give the *completeness* parameter, meaning that all MD are filled in, large focus (Palavitsines et al., 2014). It is our viewpoint that if MD and resources are to be shared between communities, then a fairly expressive and flexible MD scheme has to be allowed. Some data providers will not be able to describe their resources with information types that primarily are meant to serve other specific parts of the research community. As mentioned earlier, experiences from CLARIN-DK show that when many MD elements are obligatory, some users just fill in e.g. "n/a" for open text values. As we find a flexible MD scheme a necessity, completeness can in our view only be applied to the obligatory elements and can be used to test to which degree the optional elements have been filled, but measuring if all elements have been filled cannot be used as a quality measure when aggregating MD from different research groups.

The criterion of *timeliness* refers to the need of MD to be synchronized with the data objects. As CLARIN data centres provide persistent identifiers (PIDs) for both metadata and data content, this quality criterion has focus already, and testing for compliance can be done automatic by testing that the PID's resolve.

The criterion of *accessibility* is addressed as a key issue in the CLARIN community as a goal is to be able to share and reuse language-based resources. The whole framework established in the CLARIN Community with aggregation of MD to the VLO, CMDI etc., focus on MD and data being searchable, accessible and usable.

The criterion of *provenance* assess if it is clear who is responsible for creating, extracting, or transforming of the MD, how MD was created and what transformations was done on the data since the creation. Providing information on provenance is mainly left to the national data providers, as the VLO only facets information about the data centre (collection) providing the MD and the national project (national CLARIN consortium) where the data was harvested from.

In our view the two criteria *conformance to expectations* and *logical consistency and coherence* cover the main issues with regard to interoperability for MD. Bruce and Hillman test the criteria for *conformance to expectations* by asking: "Are controlled vocabularies aligned with user characteristics and understanding of objects?" We think that CLARIN should meet the users understanding of the objects by providing tools for documentation of not only the MD schemes, but also for the vocabularies content, to enable knowledge sharing for MD. The CCR, which is currently under further development, holds definitions of elements and values of MD, but there are no plans to include information on semantics e.g. inhered by the structure of the elements. The structure is currently only documented in the schemas defined in the CMDI Component Registry. The mappings done by the VLO (see Figure 3) are mostly based on CCR-identifiers, but there is an option to specify a mapping based on an XPath, which allows for mapping based on elements in a specific context in the schema.

For data archive administrators the VLO offers an option to test the mapping done for a specific schema[7] when aggregating MD into the VLO. For the archive administrators this is a usable feature, but it is too advanced for the researcher creating MD. To make it easier for the researcher creating interoperable MD, extensive help and information of similar kind should be available when creating MD for resources. This could include an interface with links to CCR, VLO facets, national facets and the used schema, as this would make it more obvious for the MD creator where and how the MD will show up through harvesting and how other researchers can use the MD for search.

Furthermore, future research group discussions about logical consistency and coherence of MD might also lead to a change in the use of MD elements and values. Such development in the MD use, and new interesting MD elements, e.g. the geolocation information for creation of text (as in Twitter messaging known as a tweet) should be welcomed as an obvious development.

We see it as a responsibility of the national infrastructures to ease the knowledge sharing on MD, to inspire research communities to agree on conformance and to support the researchers with knowledge on MD production and helpdesk assistance. The national infrastructure can be seen as the "man in the middle" offering mediation.

6. Conclusion

Interoperability of MD from different providers is not only a matter of harmonization of MD syntax (format), and not even of MD semantics, different research fields will have different views on the matter described. An example of this is the values of the MD element *subject*. This being said, agreeing upon standards, making clear definitions of the semantics of the MD and their content is inevitable for the interoperability to work successfully. The key points are clear and freely available definitions, accessible documentation and easily usable facilities and guidelines for the MD creators.

In our view the agreement on standards used and the semantics of the elements and their values, should emerge from user groups and research communities as e.g. for the TEI header or the use of IMDI. Only when the MD are clearly defined, can a careful mapping take place in e.g. the CLARIN infrastructure and in the aggregating facilities.

We believe that user groups and research communities should define what kind of MD is needed for a certain field. But as well the definition of MD cannot be solely a top-down process, it cannot only be bottom-up either. To

[7] Use "Check profile" at https://vlo.clarin.eu/mapping. XML details of the mapping can be found at https://github.com/clarin-eric/VLO/blob/master/vlo-commons/src/main/resources/facetConcepts.xml

ensure a proper interoperability, it needs to go both ways. The CLARIN infrastructures must function as mediators and facilitate both the sharing of resources and ensure the interoperability by helping to define MD for new resources, hosting the data and schemes used, and of course define a set of common top-level MD that can be extracted from the various resources MD and aggregated by other institutions. Lastly, the infrastructure should give support in terms of man power to structure data and to do knowledge sharing in network communities.

We can conclude by repeating (Nilsson, 2009):

"At a superficial glance, the major problems of metadata harmonization seem to relate to formats … The problem instead lies on another level, in the interpretation or semantics of the metadata expressions". And further add that the problem relates not only to the syntax or to the semantics but to the interpretation of the semantics, the perspective from which you view the data.

7. References

Asmussen, J. (2012). Text metadata – What the header of a text item looks like. *Technical Report, DK-CLARIN WP2.1*, Copenhagen.

Bruce, T. R., Diane I. Hillmann, D. I. (2004). The continuum of metadata quality: Defining, expressing, exploiting in *Metadata in Practice*, ALA Editions.

Hansen, D. H., Offersgaard, L., Olsen, S. (2014). Using TEI, CMDI and ISOcat in CLARIN-DK. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, (LREC 2014), Reykjavik, Iceland.

Henriksen, L., Hansen, D. H., Maegaard, B., Pedersen, B. S., Povlsen, C. (2014). Encompassing a spectrum of LT users in the CLARIN-DK Infrastructure. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, (LREC 2014), Reykjavik, Iceland.

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

Nilsson, M. (2010). *From Interoperability to Harmonization in Metadata Standardization*. Diss. Doctoral thesis, Stockholm, Sweden.

Offersgaard, L., Jongejan, B., Seaton, M., Hansen, D. H., (2013). CLARIN-DK – status and challenges. In Proceedings of *The workshop on Nordic language research infrastructure*, (NODALIDA 2013), NEALT Proceedings Series 20.

Palavitsines, N., Manouselis, N., Sanchez-Alonso, S. (2014). Metadata Quality in Digital Repositories: Empirical Results from the Cross-Domain Transfer of a Quality Assurance Process. *Journal of ASIS&T*. In Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.23045

Simons, G. F. (2014). The Role of Metadata in the Infrastructure for Archival Interoperation. In *Language and Linguistics Compass*, Vol. 8, Issue 11.