Reusing CMDI components for a textCorpusProfile
- towards a generic textCorpusProfile

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

Keywords: metadata, text corpus profile, CMDI

1. Background

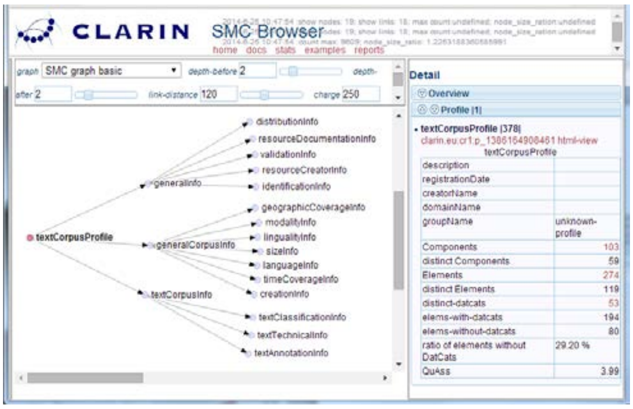
The CLARIN-DK repository, *clarin.dk*[1], contains app. 150.000 resources of various kinds: text documents, pictures, audio and video files, lexica, tools and stand-off annotations (Offersgaard et. al, 2013). Most of the resources are single texts. Due to shortage of resources in the preparatory phase, none of the resources were organized in collections although compiled as such, no metadata for corpora were created. This deficiency is being addressed now by adding metadata for the current 14 specialized corpora.

One part of the implementation of the corpus concept in *clarin.dk* was the definition of a metadata scheme covering different kind of collections. The creation of the textCorpusProfile in CMDI was the first step in this process, reusing as many components, elements and data categories from Component Registry as possible. The reuse approach is in line with the objectives for CMDI (Broeder et al., 2010).This paper will address perspectives of reuse of CMDI components, present a generic textCorpusProfile, and suggest a minimal set of obligatory metadata.

2. Finding CMDI profiles for collections

When creating metadata schemes for new resources, in this case text corpora, it is helpful to be able to reuse already existing CMDI profiles. Different profiles for corpora are available in the Component Registry, e.g. from META-SHARE[2], CLARIN-NL, and CLARIN-D, but it is difficult to find out which of the 199 profiles are the right candidates to compare because the naming is very different. The META-SHARE profile is for example named *resourceInfo*. Furthermore, it is increasingly difficult to get an overview of similarities and differences between existing profiles, their components, elements and ISOcat Data Categories (DCs), not to mention their structure and granularity. For this purpose the SMC Browser[3] that visualizes the graph structure of profiles, is a useful tool.

The SMC Browser (Durco et al., 2014) functions as a viewer with the possibility to inspect more profiles at the same time seeing mutual overlap. The overlap is, however, only among components and elements with the same ID. All kinds of even small changes to a component will result in a new ID and the component will therefore appear as being a completely new with no link to its originating component. This loss of linking to the originating component makes it very difficult to discover similarities between components, even when only small changes has been made.



2014-6-26 10:47:54 show nodes: 19; show links: 18; max count undefined; node_size_ration undefined
2014-6-26 10:47:54 count max: 9609; node_size_ratio: 1.226318836056981

graph SMC graph basic depth-before 2 depth-after 2 link-distance 120 charge 250

- textCorpusProfile [378]
clarin.eu:cr1:p_1386164908461 html-view
textCorpusProfile

description	
registrationDate	
creatorName	
domainName	
groupName	unknown-profile
Components	103
distinct Components	59
Elements	274
distinct Elements	119
distinct-datcats	53
elems-with-datcats	194
elems-without-datcats	80
ratio of elements without DatCats	29.20 %
QuAss	3.99

Figure 1: The 2 top levels of the created *textCorpusProfile* viewed in the SMC browser[4]

Before choosing a profile, one has to decide on the level of granularity of the metadata. On the one hand filling in the metadata description of resources can be tedious work and some users might like it to be as easily done as possible, only specifying the most needed information. On the other hand the metadata scheme should be fine-grained enough to enable the expression of e.g. subject domain, time coverage, annotation tool and other characteristics for text corpora. Moreover, detailed and structured metadata enables sharing and search across text corpora.

3. The textCorpusProfile

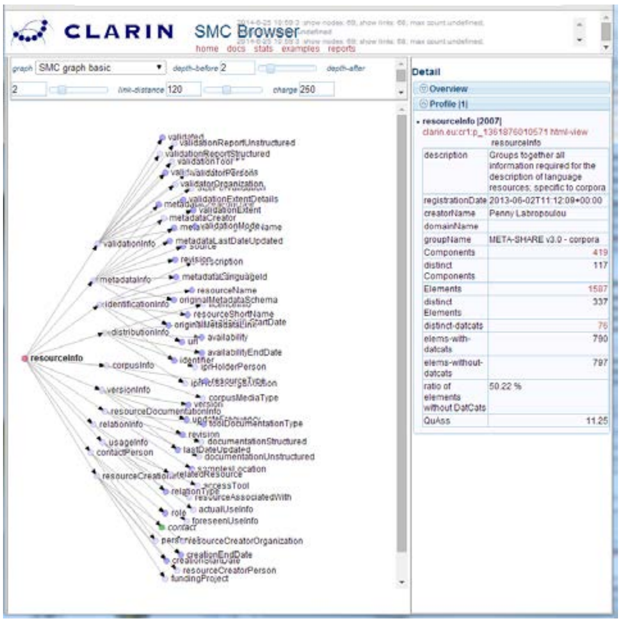
Our goal was to get a profile, with the possibility of specifying a wealth of detailed metadata while leaving as many of the metadata elements optional as possible. With this strategy the profile should cover the details of the

[1] https://clarin.dk/clarindk/forside.jsp
[2] http://www.meta-share.org/
[3] http://clarin.oeaw.ac.at/exist/apps/smc-browser/index.html

[4] The profile is currently private but can be inspected in the SMC Browser by searching for textCorpusProfile in the *Index* search field in the left column, and then clicking on the one with 378 elements.

current text corpora, and allow for expressing information from future text corpora.

We inspected different profiles in the Component Registry covering: collections, corpora, written corpora and text corpora. After comparing coverage, granularity and information types, we chose to take our point of departure in the META-SHARE profile *resourceInfo v3.0 - corpora*. This profile is widely used in both the META-SHARE platform, currently containing 1032 corpora of which 503 are text corpora, and in the VLO where 163 resources use this profile[5]. The other profiles for corpora were only used for a maximum of 10 resources. The *resourceInfo* profile is extensive, covering all sorts of corpora but it has very tight restrictions in the form of a long list of mandatory fields, which makes the profile less flexible to use, see Figure 2.



graph SMC graph basic depth-before 2 depth-after 2 link-distance 120 charge 250

Detail

Overview

Profile |1|

- resourceInfo |2007|
clarin.eu:cr1:p_1361876010571 html-view

	resourceInfo
description	Groups together all information required for the description of language resources; specific to corpora
registrationDate	2013-06-02T11:12:09+00:00
creatorName	Penny Labropoulou
domainName	
groupName	META-SHARE v3.0 - corpora
Components	419
distinct Components	117
Elements	1587
distinct Elements	337
distinct-datcats	76
elems-with-datcats	790
elems-without-datcats	797
ratio of elements without DatCats	50.22 %
QuAss	11.25

Figure 2: The 2 top levels of the original *resourceInfo* profile

During the META-NORD project our experience was that the META-SHARE scheme had both a too wide and a too deep structure which made it difficult to find the right metadata element even though you knew that the element was there.

Therefore we have restructured the profile to have a clear distinction between i) general information, ii) corpus specific information and iii) text corpus specific information, see Figure 1. In this way the different components are reusable for different collection profiles. Working with CMDI profiles and the Component Registry we have experienced that it is much easier to modify a profile by deleting elements and components than creating a new one. We therefore promote having a large well-structured textCorpusProfile that can easily be modified to a more simple profile by deleting components. By structuring the profile in generalInfo, generalCorpusInfo and textCorpusInfo it should be easy to add information in the right places or to simplify the profile.

We have also simplified the profile by leaving out all components dealing with non-text corpora to have a profile for text corpora only. We have left out elements and components at a general level like *metadataInfo* and *usageInfo* because they focus on META-SHARE use, and we have replaced the component *resourceDocumentation* with the slimmer NaLiDa *Documentation*, omitting a wide range of bibliographic information. Finally we have added elements in a few places e.g. information on annotation tool, as this information was not suited to be added in any other metadata elements.

In general we have reused many deeply nested elements and their component location. We have reused many component names but deleted elements and given looser bounds in form of optionality on elements. The surface structure and the coverage for only text corpora is however very different from the original *resourceInfo* profile, see Figure 1 and 2. The changes resulted in a new leaner textCorpusProfile containing 103 components and 274 elements compared to the 419 components with 1587 elements in the *resourceInfo v3.0 - corpora*. Of the 274 elements 255 are overlapping *resourceInfo* elements, sharing Data Categories with definitions in ISOcat.

One could argue that creating yet another textCorpusProfile is not necessary, but we did it as the META-SHARE profile had a large number of obligatory elements, was very deep in structure, and it was in our view difficult to decide where to add information.

Although looser bounds on elements results in more flexible profiles, a core set of mandatory metadata is a necessity.

4. A core set of Metadata elements

We agree in the objectives of the CMDI architecture, that the flexibility to create and share profiles is the right way to collaborate about metadata, but interoperability could be better if CMDI (or CLARIN ERIC) could require a small core set of common metadata for all profiles that are mandatory to fill in. This core set should not be implemented as an obligatory component to be included in all profiles, as we find it very important that researchers from different communities can use standards that have emerged from or are normally used in the research communities. These communities can now create a CMDI profile that reflects the already used metadata. Although there are some difficulties in the current version of CMDI (Hansen et. al, 2014), the upcoming version of CMDI solves the most important of these. As examples of existing profiles mapping widely used metadata

[5] Instances of profiles in the VLO is found by searching on the profile ID, found in the Component Registry when choosing the *xml* tab and selecting the content of the ID-element.

standards, can be mentioned IMDI for which three CMDI profiles exists[6], and TEIP5 for texts with four CMDI profiles[7].

Nor should the core set be implemented as mandatory ISOcat DCs in conceptLinks of the profile since the same DCs can be used many places in the structure, e.g. can *title* be used as the *resource title* or the *source title*, and *date* can be the *creation date* or the *publication date*. The core mandatory metadata need to be uniquely identifiable in the profiles but currently there does not seem to be a proper place where this can be stated. Instead of an obligatory component or ISOcat DCs, the core metadata set could be stated as a list of XPaths that for each element in the set of core metadata specifies where to find the corresponding information in the metadata profile. It is, however, an open question where this list of XPaths should be stored to be overt for all.

Currently the VLO[8] partly performs this mapping by requiring a mapping configuration for each profile that VLO is harvesting metadata from. Creating the mapping in the Component Registry would enable users to have access to it when investigating profiles, and would also make it available for other aggregating search facilities.

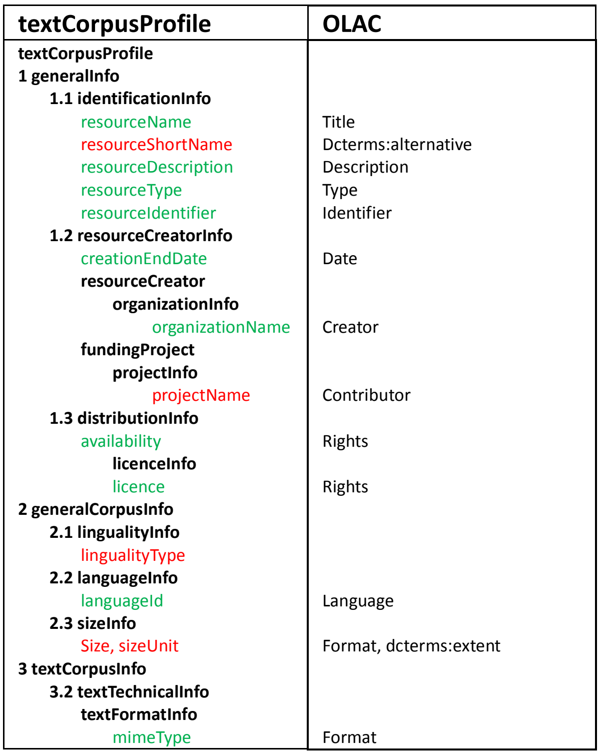
We think that the OLAC standard[9] is well suited as inspiration for such a core set of metadata, it has the following core elements: *contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, type*. But we do not find all the OLAC elements optimal as obligatory elements, and therefore suggest the following for a common metadata set:

- title,
- description
- identifier
- type
- creator
- date
- rights
- language
- format

The following OLAC elements are left out: *subject, coverage, publisher, relation*, and *source*. We suggest excluding *publisher, relation* and *source*, as these metadata elements will not be suited for all resources. *Subject* and *coverage* are also left out as we expect users to have multiple views on how to express this information. These elements could be added as optional fields.

5. The minimal textCorpusProfile

The developed textCorpusProfile has already the suggested metadata core set included, see Table 1.



textCorpusProfile	OLAC
textCorpusProfile	
1 generalInfo	
1.1 identificationInfo	
resourceName	Title
resourceShortName	Dcterms:alternative
resourceDescription	Description
resourceType	Type
resourceIdentifier	Identifier
1.2 resourceCreatorInfo	
creationEndDate	Date
resourceCreator	
organizationInfo	
organizationName	Creator
fundingProject	
projectInfo	
projectName	Contributor
1.3 distributionInfo	
availability	Rights
licenceInfo	
licence	Rights
2 generalCorpusInfo	
2.1 lingualityInfo	
lingualityType	
2.2 languageInfo	
languageId	Language
2.3 sizeInfo	
Size, sizeUnit	Format, dcterms:extent
3 textCorpusInfo	
3.2 textTechnicalInfo	
textFormatInfo	
mimeType	Format

Table 1: Minimal set of obligatory metadata for the textCorpusProfile. Green indicating elements from the core metadata set and red additional mandatory elements.

Besides the nine obligatory elements from the core metadata set, we have added a few more obligatory metadata elements, that we find important for a general collection profile. *Size* is included as we find it important for users when deciding to use a collection. *ResourceShortName* might be more useful when citing the collection than the official title, than might be very long, *lingualityType* specifies if the resource is mono-, bi- or multilingual and *projectName* declares the sponsor, or *contributer* in OLAC terms.

One of the goals for the textCorpusProfile was to design a profile where the obligatory metadata are easily filled in, and as can be seen in Table 1 this is the case with only 15 obligatory elements.

6. Conclusion

The Component Registry now contains 199 public CMDI profiles, a number that is increasing. New users are therefore facing larger and larger challenge when having to choose the right profile to use.

We present a generic textCorpusProfile. As point of departure we have chosen the META-SHARE profile,

[6] Profiles: imdi-session (clarin.eu:cr1:p_1271859438204), imdi-corpus (clarin.eu:cr1:p_1274880881885), and SpeechCorpusProfile (clarin.eu:cr1:p_1302702320401)
[7] Three profiles named teiHeader: clarin.eu:cr1:p_1380106710826, clarin.eu:cr1:p_1380106710826, and clarin.eu:cr1:p_1282306194508. OneTEIDocumentDescription (clarin.eu:cr1:p_1337778924992)
[8] http://catalog.clarin.eu/vlo/
[9] http://www.language-archives.org/OLAC/metadata.html

restructured the information, and loosened the number of obligatory elements. With the created textCorpusProfile we believe to have an extensive metadata set for text corpora of various types, with the possibility to express a minimum of only15 mandatory elements. The profile is available in the Component Registry[10] and will soon be used for the text corpora harvested from *clarin.dk*. At the same time we have a general and systematic point of departure for the creation of i) other types of corpus profiles, e.g. for multimedia and ii) other kinds of collection and resource profiles.

Furthermore, we suggest CLARIN to agree on a small core set of metadata which should be obligatory for all profiles, and we suggest that 9 obligatory elements from OLAC should be present in the core set.

We have experienced that the reuse of profiles and components often includes changes, to make them fit specific needs, but seen that sharing of ISOcat references is a clear benefit that will ease interoperability.

7. References

Broeder, D., Kemps-Snijders, M., et al. (2010). A data category registry- and component-based metadata framework. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC ’10),* Valletta, Malta. ELRA

Durco, M.& Windhouwer, M (2014) *The CMD Cloud*. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Reykjavik, Iceland. p.687-690. ELRA.

Hansen, D. H., Offersgaard, L. & Olsen S (2014) Using TEI, CMDI and ISOcat in CLARIN-DK. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Reykjavik, Iceland. p. 613-618. ELRA

Offersgaard, L., Jongejan, B. & Hansen, D. H. (2013). CLARIN-DK – status and challenges. In: *Proceedings of the workshop on Nordic language research infrastructure* at NODALIDA 2013. Linköping University Electronic Press, p. 21-32. (NEALT Proceedings Series; Nr. 20).

[10] http://catalog.clarin.eu/ds/ComponentRegistry/rest/registry/profiles/clarin.eu:cr1:p_1386164908461/xsd.
Note that the profile is currently not public.