Lightweight Semantics in the CLARIN Infrastructure

Menzo Windhouwer, The Language Archive – DANS
Matej Durco, Institute for Corpus Linguistics and Text Technology

One of the aims of the European CLARIN infrastructure [1] is to allow scholars to easily find and integrate data from a wide range of sources. This brings not only the problem of a broad diversity of formats and data structures, but also of terminology and semantics. To enable semantic interoperability the infrastructure uses lightweight semantic annotations, i.e., metadata and potentially resources refer to concepts in a concept registry to make their semantics explicit.

This approach has been fully realized in the CLARIN joint metadata domain. The metadata uses a component-based framework, CMDI (Component Metadata Infrastructure) [2], where resource specific metadata profiles can be assembled from reusable and adaptable components. The components themselves and the metadata elements they group use concept links to refer to metadata concepts from Dublin Core [3] or the ISOcat Data Category Registry [4]. Both these registries contain basically a flat or very shallow list of concepts, i.e., there is no rich set of ontological relationships between the concepts. Still this creates a semantic layer on top of the metadata profiles that helps to overcome different modeling choices made by the metadata modelers. The SMC (Semantic Mapping Component) browser [5] helps the modelers to get insight in how components are semantically related.

Metadata records of many different kinds are hosted by tens of CLARIN centers. For the central CLARIN catalogue [6] these metadata records are harvested and mapped to a set of common facets. This mapping process uses the semantic layer, i.e., in a profile it finds metadata elements that use a concept related to the facet and selects all the instances of these elements as facet values. Further refinements of this mapping process, e.g., taking more of the context of the elements into account, are currently investigated.

Additional registries are currently being developed. The first is a relation registry [7], which stores at least equivalence and near sameness relations between concepts. The second is a schema registry [8], which stores semantically annotated schema’s for the content of linguistic sources.

5. [http://clarin.oeaw.ac.at/smc-browser/](http://clarin.oeaw.ac.at/smc-browser/)