Standards in CLARIN

A position paper of the CLARIN Standards Committee

We, the CLARIN Standards Committee (CSC), seek the cooperation of the CLARIN boards and committees in solving some problems known to exist with respect to interoperability and sustainability. Problems that in our opinion are currently blocking further efficient progress with respect to integration of data and services in the CLARIN domain.

We see a need for action to achieve:

- limiting development efforts to a limited number of conversion tools allowing to move between formats, and for recommending proper archiving formats
- guiding users in choosing an appropriate format to encode linguistic information for specific purposes

As a start, we propose establishing a procedure for defining a number of CLARIN supported formats, preferably defining as few of such formats for each data type[1]. Follow-up steps will be among others the creation of documentation guiding users to use such formats when creating 'new' data for specific purposes[2].

However, all such actions are of little value, if the list of CLARIN supported formats remains a list only and is not part of a broader CLARIN policy framework. CSC feels that explicit support and guidance from the CLARIN boards and committees is needed for CLARIN standards to have sufficient impact with researchers and the CLARIN resource centers that are creating and providing resources and services. Therefore, we look for policies supporting standards from the CLARIN boards and committees (especially the CLARIN Center Committee) that can effectuate a sufficient impact.

Supporting CLARIN standard formats with respect to the CLARIN centres means:

- making tools available to CLARIN community that support such formats
- CLARIN funded tools should support these formats
- important legacy formats should be supported and/or automatically converted
- important "reality" formats (like HTML, PDF, Word etc.) should be supported and/or automatically converted

CLARIN should stimulate the centers to deliver this type of support. Therefore, CSC would like to invite the CLARIN centers to help defining a procedure to support these sets of standards. At the same time, CSC requests CLARIN BoD to ensure the use of these recommended standards in the national consortia.

---

[1] data types, for instance:

- (annotated) natural language texts
- lexical conceptual data (dictionaries, terminological databases, WordNets
- multimodal data (speech, conversational analysis, sign language data, gesture)

[2] Cookbook with recipes