Common Language Resources and Technology Infrastructure

CLARIN

February 2009

Standards and Best Practices

What is it?

An integrated and interoperable domain of language resources and technology as envisaged by CLARIN cannot be built up without agreeing on standards and best practices (SBP) where standardization is yet not far enough. Wherever two or more parties interact with each other such agreements are a necessary prerequisite. The specification of SBP will influence the internal mechanisms of a service provider only to a limited extent. These may play a role when for example the quality of services is evaluated. Due to the importance CLARIN will announce the standards and best practices which are obligatory and which are recommended explicit via the web site. CLARIN will also push new standards via ISO TC37/SC4 where necessary.

What is it for?

Users increasingly often want to combine resources and services at a virtual level, i.e. integrate resources offered by different providers and created by different linguists and projects. This can only be done when barriers can be overcome at all levels by using appropriate standards and best practices. Standards will serve different functions: (1) Some may be used as pivot formats or point of reference for all; (2) Some may be applied by new tools. (3) Some may guide the interaction in form of protocols or interchange formats. In many cases CLARIN will specify SBP which have been created by other initiatives, some SBP which emerge from the CLARIN work may be of use for other initiatives.

For those standards and best practices which are obligatory or which are being recommended, CLARIN will offer tools and frameworks that support them. If a pivot format is chosen to represent lexica, for example, there should be converters from recommended formats to this pivot format. Supporting all SBP this way will take much time.

Who can use it?

- Primarily the CLARIN standards and best practice guidelines are meant for the CLARIN community.

- However, some standards which emerge from the LRT domain are useful for other initiatives as well.

- In those areas that are relevant for many disciplines and where further standardization is required CLARIN will look for collaboration.

When can it be used?

Henry Thompson recently said that there is no agreed descriptive system in the linguistic community except than UNICODE and XML. This statement points to the core problem for all integration and interoperability work. We need fast progress in defining such standards and best practices. But the variety of areas where SBP are required is rather broad and the state of discussions and agreement building is widespread. Therefore, CLARIN will maintain a list of obligatory, recommended and known standards and best practices on its website where also the state of the process/agreement will be indicated.
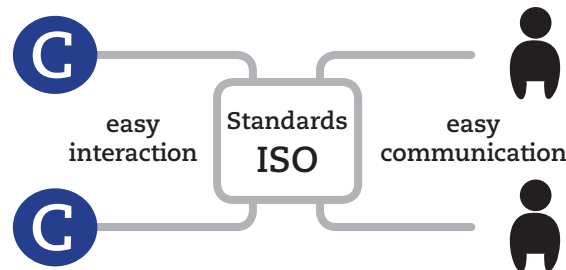
CLARIN

How does it work?

CLARIN is specifying standards and best practices at a number of levels:

- general SBP such as UNICODE, XML, RDF, OWL, SKOS, Handles, ISO Language Codes etc

- protocols such as OAI PMH, APIs, WSDL, SOAP, REST etc

- terminology/concept registries such as ISO 12620 etc

- metadata concepts and profiles such as DCMI, OLAC, IMDI, TEI Header, METS etc

- media standards such as MPEGx, H.264, JPEG, MP3, WAV etc

- text formats such as LMF, XCES, CHAT, etc

Two criteria will play a role for making statements: (1) Is the specification open and already used by a large community, i.e. is it "best practice" so that CLARIN cannot ignore it. (2) Is the specification subject of a standardization process or a recommendation by a large initiative and is it relevant for the CLARIN intentions. Of course the answers to these questions will vary over time, therefore all statements from CLARIN need to be evaluated at regular moments.

The CLARIN working groups are the forums to discuss the issue of standards and best practices in detail. The WGs will take up requests and discuss them, make specifications and where necessary be initiative with respect to the ISO process. In those cases where CLARIN is foreseeing the need of a standard, it will publish a specification as early as possible as a recommendation knowing that for example ISO standardization procedures take time.

Who is responsible?

Within CLARIN mainly two work packages are responsible in maintaining the information at the web Site:
- WP2 will argue from the perspective of technical integration and interoperability
- WP5 will argue from the perspective of deep knowledge about language resources and tools

CLARIN will closely collaborate with initiatives such as FLARENET, TEI and ISO TC37/SC4.

Whom to contact?

For all CLARIN related questions the following persons can be contacted:

Dieter van Uytvanck (MPI): dieter.vanuytvanck@mpi.nl

Iris Vogel (U Tübingen); ivogel@sfs.uni-tuebingen.de

Where to find more information?

There is so much information about the various standards and best practices that in this document we will only specify the sites from CLARIN, TEI and ISO TC37/SC4. At the CLARIN web page about standards links will be maintained.

CLARIN: http://www.clarin.eu/recommendations

TEI: http://www.tei-c.org

ISO TC37/SC4:   http://www.tc37sc4.org/

For questions about TEI and ISO several persons can be contacted also within CLARIN. Laurent Romary is chairing both initiatives.

Responsible for the content:
Dieter van Uytvanck
MPI for Psycholinguistics
Wundtlaan 2, 6525 XD Nijmegen, NL
Website: www.clarin.eu
Email: dieter.vanuytvanck@mpi.nl

The last version of this document is maintained at the CLARIN Web-Site under documents: www.clarin.eu/documents