D5.3-7

Adoption and Implementation of Standards

Document information

Title	**Adoption and Implementation of Standards**
ID	CLARINPLUS-D5.3-7
Author(s)	Claus Povlsen, Lene Offersgaard
Contractual Delivery Date	2017-06-30
Actual Delivery Date	2017-06-26
Distribution	PUBLIC
Document status in workplan	Deliverable

Project information

Project name	CLARIN-PLUS
Project number	676529
Call	H2020-INFRADEV-1-2015-1
Duration	2015-09-01 – 2017-08-31
Website	www.clarin.eu
Contact address	contact-clarinplus@clarin.eu

Table of contents

1 Executive Summary

This report is one of a series of best practice documents that aims at supporting those countries who are preparing to become a CLARIN member as well as for countries that recently joined CLARIN. The target audience is both the national coordinators and future national coordinators, *i.e.* those representatives that are going to manage their country's participation in CLARIN ERIC.

This report focuses on the adoption and implementation of standards within the CLARIN community in terms of recommendations for both existing resources and resources to be newly created.

There are two main CLARIN websites that offer knowledge and advice with respect to finding standards and formats supported by the CLARIN community. The guidelines in this report are based on these two websites.

2 Introduction

This report is one of a series of best practice documents that aims at helping new national consortia in the initial phase of their creation. In this context, the topic is best practice in connection with the adoption and implementation of standards.

In broad terms, standards serve two overall functions with respect to the creation and curation of language resources. In the resource creation process, standards must be followed as much as possible, as it will subsequently minimize the work in connection with interoperable tasks. Furthermore, for already existing resources embedded in diverse and less homogenous formats, the use of standards is advantageous when converting and exchanging from one format to another.

The outline of this report is as follows. In Sect. 3, we give two examples that illustrate the benefits of using existing standards, *e.g.,* as pivot formats in connection with supporting the exchange of less established data formats.

In Sect. 4, we describe the standards and formats used within the CLARIN community. First, the work and aims of the CLARIN Standards Committee are outlined. Then, we guide users in selecting a standard that fits their purpose. Finally, overall recommendations are given for adopting and implementing standards within the CLARIN community.

All information is based on the knowledge that is available on two main websites (see below).

3 Using Standards in Practice

There are standards for different aspects of language-related resources. For instance, there are standards for the description of resources with metadata, for the resources' storage formats, or for annotation schemes used to enrich resources with additional layers of scientific description.

Even before the CLARIN community had formed the CLARIN ERIC, the goal was facilitating the sharing of resources in the research community. Therefore, a key issue for CLARIN is to promote sharing of metadata about resources and making them searchable in the Virtual Language Observatory (VLO).[1] This has been the driving force in the CLARIN work on the Component MetaData Infrastructure (CMDI)[2] which is briefly introduced below. In another example, we show the benefits of using standards to define a pivot format for data exchange.

3.1 Component MetaData Infrastructure - CMDI

The language resources available in repositories across Europe use a number of different formats for resource metadata that they make available. Rather than promoting a single metadata format, the Component MetaData Infrastructure[3] provides a way to create and use self-defined metadata formats in a well-defined framework[4]. It relies on a modular model of so-called metadata components, which can be assembled

[1] See https://vlo.clarin.eu.
[2] See https://www.clarin.eu/content/component-metadata.
[3] See an extensive introduction to CMDI at http://media.dwds.de/clarin/userguide/text/metadata_CMDI.xhtml.
[4] See https://www.clarin.eu/content/cmdi-examples.

together, to improve the reuse, interoperability and cooperation among metadata providers[5].

Most existing metadata schemas for language resources seemed to be too superficial (*e.g.,* OLAC) or too much tailored towards specific research communities or use cases (*e.g.,* IMDI). The new metadata formalism promoted by CLARIN defines a common ground on which metadata can be searchable to a central metadata repository. It allows the user to integrate existing schemas (IMDI, OLAC) as components and thus offers interoperability to the existing base of metadata standards. In this way the CMDI framework defines a setup where the providers of metadata can easily convert their metadata to a common format that are well-defined, and documented.

The CMDI framework offers a way to semantically define metadata descriptions using references to CLARIN Concept Registry[6]. The VLO then implements a generic mapping of the semantics of the metadata elements defined via the CMDI framework to the metadata search facets in the VLO.

3.2 Interoperability

Interoperability is a key concept within the CLARIN community striving for a common infrastructure of language resources. Standards are not applied for their own sake but primarily to achieve interoperability. A requirement for maximising syntactic and semantic interoperability is the use of common data formats that are compliant with established standards and/or de-facto standards. The following description of a concrete exchange task illustrates the advantages of using standardized formats as pivot formats.

When the EU institutions in 1999 decided to collect EU institutions' terminology resources and create one single central terminology database IATE (Inter-Active Terminology for Europe)[7], which would be accessible for all the translation units of the institutions, they faced the large task of harmonising the existing terminology resources. The following existing term databases were to be imported into IATE:

- Eurodicautom (European Commission)
- TIS (Council of the European Union)
- Euterpe (European Parliament)
- Euroterms (Translation Centre for the Bodies of EU)
- CDCTERM (European Court of Auditors)

Not only the terminology concepts used by these institutions differed (semantic interoperability), but also the database formats differed (syntactic interoperability) .In order to facilitate the syntactic and semantic harmonization work, it was decided to use the ISO standard ISO-12620 (defining the procedure of registering and using data categories[8]) as a basis for defining a pivot exchange format. Having defined an exchange format it was also quite easy to implement software converting to and from the involved term data bases.

[5] See the CMDI Component Registry at https://catalog.clarin.eu/ds/ComponentRegistry.
[6] See https://www.clarin.eu/ccr and the CLARIN Concept Registry Browser at https://openskos.meertens.knaw.nl/ccr/browser/.
[7] See http://iate.europa.eu .
[8] Cf. https://www.iso.org/standard/37243.html.

Bearing in mind how much manpower can be saved, it is strongly recommended to use standardised exchange formats already when creating or formatting resources for the first time.

4 CLARIN and Standards

4.1 The CLARIN Standards Committee

The main tasks in the CLARIN standards committee are, *cf.*, https://www.clarin.eu/governance/standards-committee:

- to collect, consolidate and prepare for publication in a single place the findings and recommendations related to standards emerging from the CLARIN preparatory phase project
- to maintain the set of standards supported by CLARIN and adapt them to new developments within or outside CLARIN
- to develop and implement procedures for the discussion and adoption of new recommendations for standards
- to ensure harmonisation of standards between CLARIN ERIC and related initiatives, such as (but not restricted to) the META project
- to ensure communication with international standards bodies such as (but not restricted to) ISO
- to publish and promote the standards supported by CLARIN
- to advise the Board of Directors in all matters related to standards

At the annual CLARIN meeting in late October 2016, it was decided that the overall evaluation process of standards and formats should be organized via the *CLARIN trac wiki*.[9] Furthermore, it was agreed that the day-to-day work was taken over by the CLARIN-D centre IDS. In the next section, focus will be on the standards and formats that are considered relevant for the CLARIN community.

4.2 The CLARIN Standards

At the CLARIN ERIC website at https://www.clarin.eu/content/standards-and-formats you will find a (non-exhaustive) list of standards that are considered relevant for the CLARIN community.

The webpage lists CLARIN's principles on standards:

- Open standards are preferred over proprietary standards
- Formats and protocols should be:
  - well-documented
  - verifiable
  - proven (being used in practice)
- Text-based formats are (where possible) preferred over binary formats
- In the case of digitisation of an analogue signal, using no or lossless compression is recommended

The standards and formats selection are based on the CLARIN Standards Guide at http://clarin.ids-mannheim.de/standards/[10]. Some of the standards are annotated as *fully recommendable* while others are tagged as *acceptable*, or *not recommended*.

[9] See https://trac.clarin.eu/wiki/StandardsCommittee and https://www.clarin.eu/dev for instructions on how to access the wiki.

[10] Some of the standards listed at the website of the CLARIN Standards Guide are still not represented in the list at the CLARIN website (for instance the standard covering the Penn

An example of a recommended standard is the TEI Guidelines, a standard used in the Humanities to describe the annotation or mark up of electronic texts. An example of an acceptable standard is the OLAC Metadata scheme used for the description of language resources.

4.3 The CLARIN Standards Guide for Standards Selection

There are two options when using the CLARIN Standards Guide platform during the process of choosing an adequate and suitable standard for a given data set:

The webpage, http://clarin.ids-mannheim.de/standards/views/list-topics.xq offers a list of topics that can guide you to find an appropriate standard for your specific purposes, be it linguistic annotation of text collections or metadata.

The webpage http://clarin.ids-mannheim.de/standards/views/list-specs.xq?sortBy=name&page=1 lists 94 standards, together with metadata information about the person, organization or standard body that has developed or currently maintains the standard.

5 Lessons Learnt

The use of standards supports the interoperability and sustainability of resources. In connection with harmonization tasks, their use as pivot formats is extremely useful.

CLARIN strongly recommends to use the standards listed at the CLARIN ERIC website. In case your needs are not met by this list of those standards, please consider the more extensive list of standards on the CLARIN Standards Guide. Use the website at http://clarin.ids-mannheim.de/standards/views/list-topics.xq where you find standards categorized by topics (*e.g.*, linguistic annotation of a corpus, description of lexica).

If you have a choice between multiple standard, please use the ones used by your (potential) cooperation partners.

Note that during the discussions in the CLARIN Standards Committee, it has been proposed to develop tools that convert between (a limited number of) formats. A national standards committee member shall be elected to help with this task (choice of standards and implementation of tools to convert from one standard to another).

6 References and Further Reading

- CLARIN ERIC website about standards at https://www.clarin.eu/content/standards-and-formats.
- The CLARIN Standards Guide at http://clarin.ids-mannheim.de/standards/views/list-specs.xq?sortBy=name&page=1.
- The CLARIN Standards Guide on topics and standards at http://clarin.ids-mannheim.de/standards/views/list-topics.xq.
- Best Practice Documents CLARINPLUS-D5.3 in this series can be found at https://www.clarin.eu/content/information-potential-new-members.

Treebank data). The reason for this discrepancy is that the recently added standards at the CLARIN Standards Guide have not yet been through an evaluation process.