Short Guide

Common Language Resources and Technology Infrastructure

CLARIN

March 2009

Relevance

CLARIN community

for all communities

# Create CLARIN Metadata Now

## What is it?

Descriptive Metadata is used to characterize data resources and tools to facilitate discovery and management in large (virtual) infrastructures and repositories, i.e. they make resources visible to everyone. Since on the one hand the development of the coming CLARIN component framework (http://www.clarin.eu/documents/short-guides) will need some development time and on the other hand various institutes want to deliver metadata already now this short guide presents a transition scenario accommodating for both goals. This short guide is complementary to the one about the component metadata model. All statements about the component metadata model are valid and we expect that the new infrastructure will come in place in 2010.

## What is it for?

Users increasingly often want to

- search for specific data resources or tools in a rapidly increasing domain;
- create views and apply filters on large numbers of metadata descriptions to simplify navigation;
- combine metadata queries with content queries to get answers to research questions;
- build virtual collections and virtual workflows by combining data resources resp. tools;
- easily manage large collections by grouping the resources and carry out management operations;
- enable machines to automatically find appropriate resources for a given task.

It is widely agreed that high quality metadata is the only way to support re-usage in an era with an extreme growth of data resources and tools of all sorts. Since a metadata description can be seen as a kind of incarnation of a resource and since it contains additional information it can also be used for all types of automatic manipulations in the emerging eScience scenario.

## Who can use it?

- This guide is especially directed to those institutes that want to deliver metadata NOW to be harvested by CLARIN service providers that are offering a portal allowing to do searches.
- The procedure recommended in this guide can be applied by all institutions that are able to produce either IMDI schema based or Dublin Core / OLAC compliant metadata descriptions extracted from some internal representations.

## When can it be used?

The suggested procedure can be started NOW, but it should be synchronized closely with the MPI who will act as the site housing the CLARIN metadata catalog in the transition phase.

## How does it work?

Institutes have metadata in various formats partly even embedded in database fields or as TEI headers included in the data resources. The same holds for tools and services that are described in various formats. The question now frequently addressed is what these institutes should do when they want to have their metadata be registered now. What are the options and when will they become available:

**A.** Already now the MPI is able to harvest IMDI schema based or Dublin Core/OLAC based metadata. It should be noted that IMDI is richer, i.e. in future rich metadata will be expected in semantic web scenarios.

**B.** In April we will have the formal specifications for metadata components ready, i.e. people could wait, design their components with the help of MPI and generate CLARIN compliant metadata. We cannot yet harvest them or show them in a catalog, so practical value for those institutes that want their data visible within short is not given at this moment. Therefore we will discuss this option only at the web-site.

### Option A1 - IMDI Generation

- The institute creates IMDI schema based metadata, i.e. with the help of XSLT scripts or other methods metadata is extracted so that it complies with the IMDI schema, the semantics of its elements and vocabularies. For elements with a semantic scope that cannot be mapped on IMDI elements the key-value pairs of IMDI can be used. To determine proper mappings MPI is willing to give help.
- The created metadata descriptions should be checked against the IMDI schema to verify their correctness. Then they should be made available as linked XML files with a root URL.
- IMDI descriptions can also be offered via the OAI PMH (V2.0) protocol, but then two extra steps need to be carried out:
  - The institute needs to install an OAI PMH data provider software component and to create a mapping from elements to the limited Dublin Core or OLAC elements.
  - Having done this, both IMDI as well as DC/OLAC metadata descriptions can be offered via OAI PMH.

### Option A2 - Dublin Core / OLAC Generation

- The institute creates metadata descriptions that adhere to the DC/OLAC semantic and structure specifications.
- These descriptions are offered via the OAI PMH data provider component.

Summarizing we can say that the fastest way to integrate your metadata descriptions is to create IMDI schema based or DC/OLAC descriptions which MPI can harvest. The CLARIN metadata infrastructure will provide facilities that will safeguard any investments you do creating for example IMDI/DC metadata. It will provide special tools for transforming IMDI/DC metadata into the metadata components format.

## Who is responsible?

MPI is leading work package 2 of CLARIN and will implement this procedure and is for the moment acting as the central metadata catalog site that will harvest the metadata and integrate it into a portal.

## Whom to contact?

For information about the fast integration of metadata, please, contact the WP2 address:

Dieter van Uytvanck: dieter.vanuytvanck@mpi.nl

## Where to find more information?

The official CLARIN web-site is the source of all information:

| | |
|---|---|
| IMDI: | http://www.mpi.nl/IMDI/ |
| OLAC: | http://www.language-archives.org/ |
| DC: | http://dublincore.org/ |
| OAI PMH: | http://www.openarchives.org/pmh/ |
| OAI tutorial: | http://www.oaforum.org/tutorial/ |

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