Common Language Resources and Technology Infrastructure

CLARIN

February 2009

Component Metadata

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. The need for a component based metadata framework resulted from the experience in the LRT community that fixed schema solutions hamper a broad usage due to the different needs and terminologies of sub-communities. The component metadata framework allows users or projects to design their own set as long as they make use of widely agreed concepts that are stored in the ISOcat registry and therefore guarantee interoperability.

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?

- Metadata descriptions are open, i.e. all interested users can carry out browse and search operations. Thus metadata are facilitating interdisciplinary.

- In addition metadata descriptions can be harvested by service providers to create informative portals. These service providers are not limited to the discipline, but can be initiatives started by the national libraries or even companies adding value for example.

- Increasingly often we can assume that software will make use of metadata to find suitable resources for a specific task. For many operations fast access to the ISOcat concept registry is required, since this will guarantee interoperability.

- In contrast to a solution with a fixed schema where the discipline semantics are built in in the schema, the component based approach allows to build a discipline independent framework.

When can it be used?

The CLARIN component metadata framework is being defined and built right now. It requires that the metadata concepts are being registered in the ISOcat
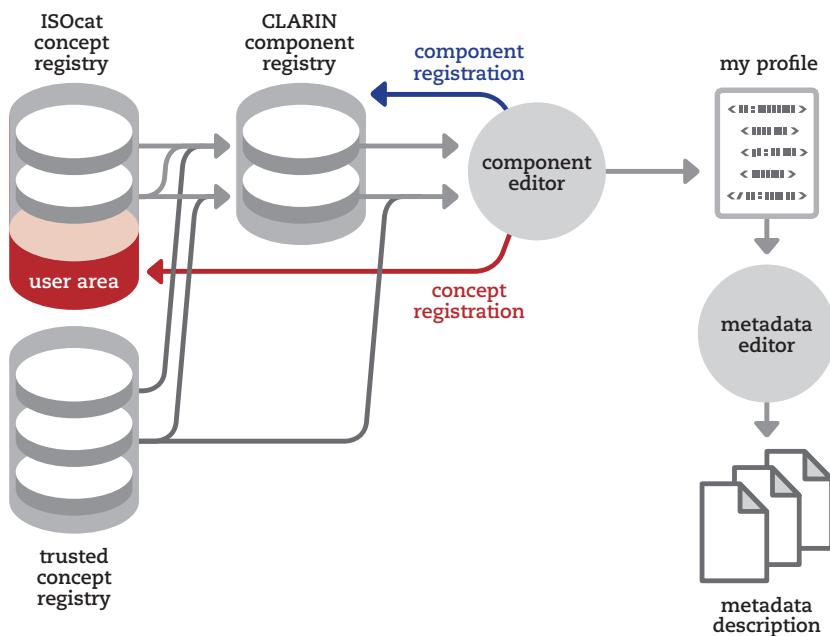
registry which will be done in first half of 2009 so that everyone can use them. First versions of the tools enabling the component framework (editors, search engines, browsers etc) will become ready in 2010. We recommend using the currently well established sets dependent on the specific requirements: (1) IMDI and its special profiles including TEI elements or (2) OLAC. CLARIN will take care that the installed based will be integrated in the component solution.

How does it work?

In the CLARIN component metadata framework the anchor for interoperability is not anymore a standardized profile (schema) that specifies syntax and the elements to be used, but the ISOcat data category registry that will include all metadata concepts that are seen as being necessary by the community. The user now first needs to specify his own components and profile which are then tailor-made for his specific purposes. The component/profile editors will take care that only registered concepts can be used. Then flexible editors will allow users to create metadata descriptions according to the profile. Search engines and other operators will exploit the references to the ISOcat registry to carry out operations across descriptions created based on a variety of profiles.

In doing so various sub-communities can easily tune the metadata descriptions to their specific needs as for example the Sign Language or the "Virtual Reality" communities. CLARIN will facilitate the creation of profiles by suggesting a number of ready-made components that are addressing typical parts of descriptions such as the location and project descriptions and profiles that are representing the needs of certain communities such as a lexicon profile often used by NLP (natural language processing) experts. Also components and profiles will be offered that contain IMDI, TEI and OLAC specifications to take care of the already existing metadata records. All these components and profiles will be registered as well as new ones in supported registries so that component/profile editors offer these options to the users.

ISOcat concept registry — CLARIN component registry — component registration — component editor — concept registration — user area — trusted concept registry — my profile — metadata editor — metadata description

Who is responsible?

Work package 2 of CLARIN is leading the design and implementation of the component framework based on the requirements from the language resource and technology community. Also the specification of the metadata concepts to be entered in the ISOcat concept registry will be taken care of by WP2.

Whom to contact?

For information about the component based metadata framework, please, contact the WP2 address:

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

Where to find more information?

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

CLARIN: http://www.clarin.eu

CLARIN Component Metadata Document:
          http://www.clarin.eu/specification-documents

IMDI: http://www.mpi.nl/IMDI

OLAC: http://www.language-archives.org

TEI: http://www.tei-c.org/release/doc/tei-p5-doc/en/
          html/HD.html

DC: http://dublincore.org

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