A Data Category Registry- and Component-based Metadata Framework

Daan Broeder et al.
Max-Planck Institute for Psycholinguistics
The CLARIN project is a large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily useable for Language & SSH (Social Sciences & Humanities) researchers.

CLARIN EU project and different national CLARIN projects

CLARIN EU WP2 since 2007 investigated and creates (prototypical) solutions for:

- Common AAI infrastructure
- Single system of persistent identifiers (PID) for resources
- Common metadata domain
- ...
Current Metadata Situation

Fragmented landscape

- Metadata sets, schema & infrastructures in our domain:
  - IMDI, OLAC/DCMI, TEI

- Problems with current solutions:
  - Inflexible: too many (IMDI) or too few (OLAC) metadata elements
  - Limited interoperability (both semantic and functional)
  - Problematic (unfamiliar) terminology for some sub-communities.
  - Limited support for LT tool & services descriptions
Why a common metadata domain:

- Finding and sharing resources housed at all archives & repositories participating in CLARIN
- Specify distributed heterogeneous collections of LRs and processing these collections
- In general, a common metadata domain helps bringing along a single domain of LRs
CLARIN chose for a component approach: CMDI
- NOT a single new metadata schema
- but rather allow coexistence of many (community/researcher) defined schemas
- with explicit semantics for interoperability

How does this work?
- Components are bundles of related metadata elements that describe an aspect of the resource
- A complete description of a resource may require several components.
- Components may use and contain other components
- Components should be designed for reusability
Metadata Components

Technical Metadata

Sample frequency
Format
Size
...

Let's describe a speech recording
Let's describe a speech recording
Metadata Components

Let's describe a speech recording

- Actor
  - Name
  - Age
  - Sex
  - Language
  - ...

- Language

- Technical Metadata
Metadata Components

- Location
  - Continent
  - Country
  - Address
- Actor
- Language
- Technical Metadata

Let's describe a speech recording.
Let's describe a speech recording
Let's describe a speech recording.
Metadata Components

- Project
- Location
- Actor
- Language
- Technical Metadata

Metadata profile

Lets describe a speech recording

Metadata schema

Metadata description
Metadata Components

Project
Location
Actor
Language
Technical Metadata

Metadata profile

Let's describe a speech recording

Profile definition XML

Metadata schema
W3C XML Schema

Metadata description
XML File
User selects appropriate components to create a new metadata profile or an existing profile.

Selecting metadata components from the registry.
Concept registries

- Basically a list with concepts and their descriptions where every concept has a unique identifier.
- Some have a complicated structure and are associated with elaborate (administrative) processes to determine the status and acceptance of concepts in the registry. e.g. ISO-DCR.
- Others are static and simple lists of concepts and descriptions e.g. DCTERMS
ISO DCR

- ISO-DCR is important for more CLARIN objectives than metadata and is under control of the linguistic community (ISO-TC37)
- is an implementation of the model defined in ISO 12620, offering a GUI and programming APIs
- Every DC is subject to a standardization process and carries information on the status of that process
- Metadata is just one of 13 Thematic Domains in the DCR
- Can contain no relations between the DCs, only a value domain relation is possible.
User selects appropriate components to create a new metadata profile or an existing profile.

Semantic interoperability **partly** solved via references to ISO DCR or other registry.

Selecting metadata components from the registry.
Perform search/browsing on the metadata catalog using the ISO DCR and other concept registries and CLARIN relation registry.

Create metadata schema from selection of existing components. Allow creation of new components if they have references to ISOcat.

Metadata harvesting by OAI-PMH protocol.

Metadata descriptions created.

Metadata component profile was selected from metadata component registry.
CMDI Architecture I

- Division into:
  - MD Producer components
  - MD Exploitation or consumer components
  - OAI-PMH components
  - Knowledge components: DCR, Relation Registry
- The CMDI takes an archivist or “production” first viewpoint
  - Prioritize that the metadata can be of good quality: consistent, coherent, correctly linked to the concept registries
  - The consumer side can be more “experimental” and diverse.
  - Many MD exploitation “stacks” or consumers can work in parallel on the same metadata
CMDI contributors

Collaboration on the CMDI implementation

- MPI for Psycholinguistics: metadata modeling and editing facilities
- Språkbanken, University of Gothenburg: Joint CLARIN metadata repository
- Austrian Academy: Metadata catalog, metadata & semantic mapping services
- IDS: Virtual Collection Registry
- MPG / CLARIN NL: ISO-DCR
- DFKI: Relation Registry
CMDI Architecture II

[Diagram showing relationships between various components such as user, MD Catalog, Virtual Collection Registry, MD Services, Semantic mapping Services, ISO-Cat DCR, Relation Registry, MD Comp. Editor, MD Comp. Registry, MD Editor, CLARIN Joint MD Repository, OAI-PMH Service Provider, OAI-PMH Data provider, Local MD Repository, ISO TDG, and Metadata modeler.]
Current CMDI status I

- ISO-DCR: 218 metadata concepts
- CMDI component registry: 135 components, 19 profiles

Produced & inspired by:
- Deconstructing existing metadata schema IMDI, OLAC, TEI
- Considering requirements of other CLARIN activities like profile matching
- CLARIN NL metadata project tested the CMDI model and delivered components and profiles for the resources in two major Dutch Language Resource centers
Current CMDI status II

Operational or test phase:
- ISOCat DCR
- Component registry & editor
- ARBIL metadata editor

Still working on:
- *Joint Metadata Repository, Metadata Catalog, Semantic Mapping, Relation Registry*

Expect a usable first version in third quarter 2010
CMDI: Browsing the Component Registry

### Clarin Component Registry

**Clarion Component Browser**

<table>
<thead>
<tr>
<th>Name</th>
<th>Creator</th>
<th>Description</th>
<th>Registration Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynasand</td>
<td>Folkert de Vriend</td>
<td>Profile voor Dynasand</td>
<td>26 April 2010 13:15:55</td>
</tr>
<tr>
<td>LexicalResourceProfile</td>
<td>Laura van Eerten</td>
<td>a profile for describing a lexical resource</td>
<td>26 April 2010 12:13:06</td>
</tr>
<tr>
<td>SessionProfile</td>
<td>Patrick Duin</td>
<td>Clarin-NL profiles</td>
<td>21 April 2010 16:18:10</td>
</tr>
<tr>
<td>SpeechCorpusProfile</td>
<td>Patrick Duin</td>
<td>Clarin-NL profiles</td>
<td>21 April 2010 16:18:10</td>
</tr>
<tr>
<td>SpeechCorpusProfile_Autonomata_speaker</td>
<td>Patrick Duin</td>
<td>Clarin-NL profiles</td>
<td>21 April 2010 16:18:10</td>
</tr>
<tr>
<td>TextCorpusProfile</td>
<td>Patrick Duin</td>
<td>Clarin-NL profiles</td>
<td>21 April 2010 16:18:10</td>
</tr>
<tr>
<td>TextProfile</td>
<td>Patrick Duin</td>
<td>Clarin-NL profiles</td>
<td>21 April 2010 16:18:10</td>
</tr>
<tr>
<td>avatech-recognizer</td>
<td>Patrick Duin</td>
<td>Avatech</td>
<td>21 April 2010 16:19:47</td>
</tr>
<tr>
<td>cmdi-virtual-collection</td>
<td>Patrick Duin</td>
<td>Clarin vc profile</td>
<td>21 April 2010 16:18:30</td>
</tr>
<tr>
<td>component-dcterms</td>
<td>Patrick Duin</td>
<td>DC metadata</td>
<td>21 April 2010 16:19:45</td>
</tr>
<tr>
<td>component-dcterms-modular</td>
<td>Patrick Duin</td>
<td>DC metadata</td>
<td>21 April 2010 16:19:45</td>
</tr>
</tbody>
</table>

**cmdi-virtual-collection**

<table>
<thead>
<tr>
<th>Description: A component for describing personalized sets of metadata descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Element:</strong> Name string  1 - 1</td>
</tr>
<tr>
<td><strong>Element:</strong> Description string  1 - 1</td>
</tr>
<tr>
<td><strong>Element:</strong> CreationDate date  1 - 1</td>
</tr>
<tr>
<td><strong>Element:</strong> Status draft  1 - 1</td>
</tr>
<tr>
<td><strong>Element:</strong> Reproducability certain  1 - 1</td>
</tr>
</tbody>
</table>
CMDI: Editing a Component
Thank you for your attention

CLARIN has received funding from the European Community's Seventh Framework Programme under grant agreement n° 212230