CLARIN Concept Registry: the new semantic registry

Ineke Schuurman, Menzo Windhouwer, Oddrun Ohren, Daniel Zeman

cr@clarin.eu
www.clarin.eu/ccr
www.clarin.eu/conceptregistry
In the tools and resources offered by CLARIN many *(de facto)* standards are being referred to, concerning both metadata and content data, but …

- What do they mean?
- Do they mean the same in the various tools and resources?

CMDI (CLARIN Metadata Infrastructure)

- Makes use of several registries
The metadata provided in CMDI should be clear, i.e., unambiguous, in order to be useful.

The building blocks, components, elements, attributes and values, of a CMDI profile should be clearly defined in a Concept or, for value ranges, Vocabulary registry.

Registries used:
- Dublin Core
- ISOcat (in the past)
- CLARIN Concept Registry (CCR)
- CLAVAS (in the future, CMDI 1.2)
- Too much proliferation
  - everybody could enter stuff
  - entries quite often not meeting our standards
  - entries were out of control
- Too complex
  - data category type, data type
  - while several ‘problematic’ fields were not useful for our (CLARIN) purposes

In addition: last year ISOcat had to be migrated (decision Registration Authority) and became static

- CLARIN decided to look for another solution.
New approach: CCR

⇒ **CCR** (CLARIN Concept Registry)

**SIMPLIFIED**
**CONTROLLED**

**Simplified**: several ‘fields’ not adopted from ISOcat

**Controlled**: national CCR-coordinators will filter the input

http://www.clarin.eu/conceptregistry/
Characteristics CCR

- **Browser:**
  Accessible for everybody

- **Editor:**
  Just for CCR-coordinators to insert new entries

- **API:**
  For tools, e.g., the Component Registry

**Browser:** easy search for

- Label (name)
- Definition
- Other text fields (example, history, …)
Definitions should be ‘as general as possible, as specific as necessary’, therefore they should be

1. Unique
2. Meaningful
3. Reusable
4. Concise
5. Unambiguous

Also in other fields characteristic nr 5 is to be obeyed!
Entries are ‘for ever’

Trust and reliability

- Issue in ISOcat!
- CCR ➔ controlled
  - Definitions cannot be updated in a way that changes their meaning
    - Only typos etc can be corrected
  - Preferred label (name) will not be changed
  - Instead a new entry will be created, the old one being expired if necessary
- what can be added: examples, alternative labels, ‘higher’ status, notes, additional scheme and/or collection
OpenSKOS

- Existing OpenSKOS infrastructure was adapted.
  - Already available
    - API to access, create, share thesauri and vocabularies
    - Editor
  - New
    - Concepts have a handle as Persistent Identifier
    - Faceted browser
    - Support for SKOS collections
    - Shibboleth-based access
From ISOcat to the CCR

Imported in CCR
- Entries used in CLARIN, e.g., in CMDI
- Entries recognized as belonging to a standard
- Entries selected by the national CCR coordinators

ISOcat: over 5000 entries
CCR: 3139 entries (for CLARIN)

We will perform a clean-up action before adding new entries, in order to remove duplications, project or language specific definitions, empty definitions, misspellings (organization vs organisation), …
<table>
<thead>
<tr>
<th>URI</th>
<th>Label</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://hdl.handle.net/11148/CCR_C-085_fa814390-4841-47cb-6dcb-52ce7d36b277">http://hdl.handle.net/11148/CCR_C-085_fa814390-4841-47cb-6dcb-52ce7d36b277</a></td>
<td>common noun</td>
<td>A noun or adjective denoting a class of objects. (source: ISO12620)</td>
</tr>
<tr>
<td>Field</td>
<td>Value</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>class</td>
<td>Concept</td>
<td></td>
</tr>
<tr>
<td>status</td>
<td>candidate</td>
<td></td>
</tr>
<tr>
<td>preLabel@en</td>
<td>common noun</td>
<td></td>
</tr>
<tr>
<td>definition@en</td>
<td>A noun or adjective denoting a class of objects. (source: ISO12620)</td>
<td></td>
</tr>
<tr>
<td>notation</td>
<td>commonNoun</td>
<td></td>
</tr>
<tr>
<td>example@en</td>
<td>continent (source: Mitre; TEI (green text))</td>
<td></td>
</tr>
<tr>
<td>scopeNote@en</td>
<td>... (source: Mitre; TEI (green text))</td>
<td></td>
</tr>
<tr>
<td>changeNote</td>
<td>This concept is based on the ISOcat data category: <a href="http://www.isocat.org/datcat/DC-385">http://www.isocat.org/datcat/DC-385</a></td>
<td></td>
</tr>
<tr>
<td>inScheme</td>
<td>Terminology</td>
<td></td>
</tr>
<tr>
<td>inSkosCollection</td>
<td>Uby 2012</td>
<td></td>
</tr>
<tr>
<td>deleted</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>toBeChecked</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>uri</td>
<td><a href="http://hdl.handle.net/11459/CCR_C-385_d5af2f91-c926-3747-cc35-1d21aef2717">http://hdl.handle.net/11459/CCR_C-385_d5af2f91-c926-3747-cc35-1d21aef2717</a></td>
<td></td>
</tr>
<tr>
<td>license</td>
<td>Creative Commons Attribution (CC BY) (use the uri above for the attribution)</td>
<td></td>
</tr>
</tbody>
</table>
CCR Coordinators

- If you need
  - a new concept, or
  - want to change an existing concept
- contact your CCR coordinator:
  http://clarin.eu/content/concept-registry-coordinators
- If no CCR coordinator is appointed for your country:
  ccr@clarin.eu

- For information on the CCR, the coordinators and the (upcoming) procedures see
  http://www.clarin.eu/ccr
Decision procedure

- All ‘ERIC countries’ appointed a CCR content coordinator

Wrt decisions about entries
- All CCR content coordinators (or deputies) are involved
- We aim for unanimity
- If necessary we will vote
  - A change in CCR (like adding specific new entry) is accepted when 70% or more of the coordinators represented agree
- All changes are recorded in the CLARIN CCR-section
Correction current entries, new entries

There still are ‘incorrect’ entries, i.e., entries not meeting our demands

- We are working on these.

- In the future we will have 2 weeks to come to an agreement on a batch of entries,

- The same holds wrt proposals for new entries

Exceptions: holiday season, and the initial period (=now!)
Moving to the CCR

- If you
  - have resources that contain references to ISOcat data categories which you want to replace by their CCR concept handles (if available), or
  - want to know which ISOcat data categories are imported into the CCR
- Visit https://github.com/TheLanguageArchive/ISOcat2CCR
- where you can find
  - mapping files, and
  - a tool to use those files to replace ISOcat data category references by CCR concept handles
- If you run into problems contact your national CCR coordinator or, if necessary, ccr@clarin.eu
Thank you for your attention!

(There will be a demo later today)