CLARIN
Common Language Resources and Technology Infrastructure

CLARIN Concept Registry: the new semantic registry

Ineke Schuurman, Menzo Windhouwer, Oddrun Ohren, Daniel Zeman

ccr@clarin.eu
www.clarin.eu/ccr
www.clarin.eu/conceptregistry

CLARIN Annual Conference
October 15-17 2015
Wroclaw, Poland

Background

- 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

Clear metadata

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)

Drawbacks ISOcat

- 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, …)

High quality concepts

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), …

CLARIN Concept Registry Browser help

Please type one or more space separated search terms

common noun Search Reset all

Search terms mode
○ Or **(22)** ◉ And **(2)**

Search terms matching
○ Part of word **(2)** ◉ Whole word **(2)**

Search field filters

Search exclusively in these fields

- ☑ Labels
- ☐ Definition
- ☐ Default documentation fields

Concepts found: 1 to 2 of 2 concepts

URI	Label	Definition
http://hdl.handle.net/11148/CCR_C-385_fa814390-4841-47cb-6dcb-52ce7d36b277	common noun	A noun or adjective denoting a class of objects. (source: ISO12620)
http://hdl.handle.net/11148/CCR_C-1256_8d65f6d2-9dfb-4326-f61b-63ed3673ee6d	common noun	Noun that signifies a non-specific member of a group. (source: www.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsACommonNoun.htm)

More details …

CLARIN Concept Registry Browser

Please type one or more space separated search terms

Search Reset all

Search terms mode
◉ Or ○ And

Search terms matching
○ Part of word ◉ Whole word

Search field filters

Search exclusively in these fields
- ☐ Labels
- ☐ Definition
- ☐ Default documentation fields

clear all search field filters

Facet filters

Status
- ○ Approved (233)
- ○ Candidate (2899)
- ◉ Any

Concept Schemes
- ☐ Dialogue Acts (1)
- ☐ Language Codes (4)
- ☐ Language Resource Ontology (4)
- ☐ Lexical Resources (15)

back

Field	Value
class	Concept
status	candidate
prefLabel@en	common noun
definition@en	A noun or adjective denoting a class of objects. (source: ISO12620)
notation	commonNoun common noun
example@en	continent (source: Mitre; TEI(green text))
scopeNote@en	... (source: Mitre; TEI(green text))
changeNote	This concept is based on the ISOcat data category: http://www.isocat.org/datcat/DC-385
inScheme	**Terminology**
inSkosCollection	**Uby 2012**
deleted	---
toBeChecked	---
uri	**http://hdl.handle.net/11459/CCR_C-385_d5af2f91-c926-3747-cc35-1d21aefd2717**
license	**Creative Commons Attribution (CC BY)** (use the uri above for the attribution)

back

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 concent 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)