
# CLARIN NL – ISOcat DCR EXPERIENCES

Ineke Schuurman, U Utrecht & KU Leuven
Menzo Windhouwer, TLA - DANS
Daan Broeder, TLA - MPI

Utrecht 2013-12-09

# CLARIN


CLARIN:

**C**ommon **La**nguage **R**esources and Technology **In**frastructure

➔European Research Infrastructure

Intended users:

researchers and students in Human and Social Sciences (HSS)

# “HSS”

**Very broad domain:**

linguistics, literature, archeology, sociology, psychology, …

➔Dutch, English, … (all European languages)

=> Latin, Greek, Gothic, … (classical languages)

=> Hebrew, Chinese, Arabic, Russian, …

=> Middle Dutch, … (i.e., older instantiations)

➔Speech, sign language, ‘video’, …

# Needs


When a researcher formulates a request

- Machines need to find the resources/tools wanted, even when formulated implicitly
  - A tagger for German, a parser for Roman languages, documents on Berlin (DDR), all documents referring to Tony Blair, all videos on Marilyn Monroe and Kennedy

➔Metadata needed, such that they for **allow semantic interoperability**

# Content


More content related (annotations!) requests should also be possible in the future:

- find me a resource in which ‘events’ are marked, show all instantiations
- same, but with ‘events’ à la ISOTimeML
- find me a resource (EN) with marking of nouns à la CGN

# Standards

CLARIN needs standards, but

- Several *(de facto)* standards (per domain)
  - Same names of these in several (sub)domains, or
  - Different names in several domains
- Role of language / culture
- Lots of legacy data

plus

- Theory-dependent concepts (esp. content related)

# CLARIN


➔Supporting **interoperability** is becoming even harder

CLARIN approach: **have them all defined** (and relate them elsewhere)

- concepts for metadata (CMDI) are already being defined (ISOcat),
- concepts for (linguistic) content should be defined as well (faltering start)

# Semantic registry

CLARIN needs a semantic registry

- Metadata
- Content

**ISOcat:**

**Data Category Registry** defining widely accepted data categories (DCs)

http://www.isocat.org

# ISOcat and CLARIN


**ISOcat** certainly meets CLARIN demands **to a certain extent**

➔ ISOcat ‘as is’ does have some serious disadvantages for us

# CLARIN environment


- **Non-technical** users
- **Need** for **stable** entries (esp definitions)
- **not just official standards**, also *de facto* ones and other instantiations (legacy!)
- **Explicit / unambiguous** formulation of definitions

# ISOcat: open registry


Disadvantages according to users:

- **Unreliable**
  - Essential changes in definitions
    - **CLARIN groups do not trust DC's they do not own !!**
- **Messy**
  - People are not inclined to insert their data in ISOcat
  - CLARIN-users are to do that themselves
    - sub-optimal input (they may have to formulate definitions themselves)

# Open and/or closed?


ISOcat is an **open** registry

- Everybody can register as expert user and add entries

ISOcat is at the same time **closed**

- Only the owner can change an existing entry

And this has positive and negative aspects

# Open

**Positive:**

•ISOcat is not static – new ‘standards’ can be inserted, while the ‘old’ ones remain

• ‘double’ entries can be inserted

**Negative:**

•Proliferation of entries, due to

– Owners being ‘out of reach’

– Distrust (content out of control)

# “Closed” aspects


**Positive**

- Third parties can not corrupt your entries
- You as owner are in control
  - Entries are *more* stable

**Negative**

- You as owner are to be contacted for things beyond your control (translation in Finnish, DEN (alsoKnownAs) used in Norwegean corpus, …)
- laborious (for owners)

# Preliminary ‘conclusion’

Mixed feelings:

Semantic registry should not be closed, but neither ‘too open’

That being said: the current balance in ISOcat is not considered a ‘perfect’ one!

- For existing DCs the “EN-definition” should remain closed, other parts open (or done away with)
- However: open means ADDING stuff, not deleting

# ISOcat – CLARIN


- Proliferation
  - Almost the same DCs are entered several times,
    - Sometimes even within one profile, and with same owner
- Hugh difference in quality
  - Seemingly different starting points, from
    - Definitions formulated to cover all languages,

  to
    - Definitions covering just one language / application

**CLARIN needs something 'in-between':**

**As general as possible, as specific as necessary**

# Definitions


More serious, however,

- **Ambiguous** definitions
  - Concepts used in definition are not explained, for example by linking them to their PID
    - This really makes many existing definitions ‘useless’

Consequence:

Lots of available definitions can not be re-used in CLARIN

# therefore


➔Many definitions are to be made anew for CLARIN (but can be reused by others)

➔Users are to go over lots of existing definitions to check to see whether these are re-usable

➔time-consuming / boring

- "Views" are organized (an ISOcat within ISOcat) in which a selected part of ISOcat is available. This has improved the conditions a lot (when used)
- "Recommended by …" status (kind of community standard)

# Worries wrt CLARIN


- Quality of DCs (from CLARIN point of view)
- Number of DCs
- Semantic consistency of DCs
- User-friendliness of ISOcat when creating new entries

Not sure whether our people would insert data into ISOcat when not forced by CLARIN (currently NL and VL)

# Consequence

When nothing is done, ISOcat will not be very useful for our CLARIN purposes

**Need**: **user-friendly web interface**, also for new DCs

Current problems:

- DC type (5 types, such as closed and open)
- Data type (some 40 possibilities)
- Linking closed and simples (can be looong list)

# Stable entries

- Too often, definitions are changed in a meaningful way after a DC has been made public

- Was difficult to keep up with

➔ track history made easier

- Still, people should create new entry instead of changing the old one (which is to be made superseded), sometimes users really need the older version!

# Why this complex?


Users ask themselves why they are to fill out specific fields, like type or relation

- Info available in manuals
- Better done elsewhere (RELcat)

We could do with a lighter version of a semantic registry

# Desiderata wrt ISOcat


**ISOcat-light:**

- Major changes ‘forbidden’ in public DCs
- Less (or no) types of DCs
- Less (or no) data types
- Easier combination of or no simple-closed DCs
- More fields writeable for ‘others’
- DEN per language
- No IsA relation ( ➔ Relation registry)
- (Cleanup)

# Costs/profits for CLARIN

Development by TLA/MPI -- 0.7 fte since 2007

Coordination (NL) -- 0.2 fte since 2011

Recently: national coordinators for other CLARIN countries as well, to work with their communities + full one

Profits: not that many up till now with ISOcat-as-is. It is considered a burden by many of our people. Some adaptations have been made, that helps. But we would need more *(to be discussed tomorrow)*


Thanks !