ISOcat and Semantic Operability

Ineke Schuurman
ISOcat content coördinator CLARIN-NL

Nijmegen 13-09-2012

Overview

- Context
- ISOcat
  - general
  - use in CLARIN
- Some metadata examples
- Do’ s and don’ ts

Uhhh?

“Give me a list with all forms of ‘wijf’ in 14th century documents in Dutch by female authors, the same for the 16th,18th and 20th century. Contrast them with documents by male authors and by unknown authors. Present the results ordered per region and per genre.”

- How to find data that could answer such a research question?

Metadata and machine

- Not ‘just by hand’ ►machine
- Subset selection ►metadata

Some problem(s):

- question not formulated in ‘Metadatish’
- What is clear for us is not clear for a machine
- What is meant by the concepts used (‘author’, ‘region’, ‘Dutch’)
- Several ‘definitions’ / ‘encoding schemes’ in use

CLARIN

- Not one metadata scheme favoured
- You may combine elements of several schemes

► “semantic interoperability” is to be ensured

  – Is a ‘kopiist’ an author?
  – What defines a ‘genre’, a ‘region’?

May differ in various metadata schemes coming with documents!

Consequence

Within CLARIN, metadata concepts are to be

- **defined**

CMDI, ISOcat

- **related**

RELcat

ISOcat

ISOcat:

Data Category Registry defining widely accepted data categories (DCs)

http://www.isocat.org

Registry that stores DCs for language resources and their metadata, together with properties of the DCs (definition, administration, examples, etc.)

A good example

A Pan European Initiative

NEHOL project

- *Alphabet* (DC-4143)
  - – any set of characters representing the simple sounds used in a language or in speech generally

In principle good because:

- No language / project dependency
- No tautology
- Reusable (not too strict)

Some ‘rules’

- Adopt an existing entry,
  if not possible
- create a new entry

In all cases: the entries should be GOOD ones

- But: what makes an entry a good one, one that you can (re)use?

What defines a good DC?

Reusable definition

NOT

- *conversation* (DC-2661)
  - Communication event with more than two participants
- *mother tongue* (DC-2955)
  - […] a speaker’s mother tongue

What defines a good DC?

Correct definition

NOT (?)

- *Actor* (DC-4146)
  - a participant in an action or process

Question: is an **addressee** to be considered an actor? (used in DC-4158, no proper definition yet)

What defines a good DC?

Meaningful definition

NOT

- *annotation format* (DC-2562)
  - – Specifies the annotation format that is used …

- *source language* (DC-2494)
  - – Indicates if a language is a source language

Not that good examples

- Mother tongue (DC-2955)
  - – Specifies whether the language is a speaker’s mother tongue
- Mother’ s language (DC-4516)
  - – […] NOT necessarily the mother tongue […]

- There is no definition of concept ‘mother tongue’

(Relation with /home language/ , /primary language/, /heritage language/?)

- And why ‘speaker’ ?

Rule

Make your definition

- as general as possible
- as specific as necessary

Do’ s

Do’ s:

- Create a DCS for your scheme (name project, annotation scheme, …)
- Provide clear definition (short, to the point) for your scheme, application, ….
- Take care not to leave concepts used in your definition undefined or vague
- Use appropriate profile (NOT: ‘private’)
- Use appropriate vocabulary (per profile)
- Check ‘adopted’ DC’ s regularly till standardization

Don’ ts

- Be (too) language specific in definition
- Mention scheme in definition
- Use several definitions in one DC
- Circular definitions
- Rely on authority
- Rely on standardized status
  - – Definition should fit YOUR scheme, etc

Athens Core

DO USE THESE DCs !!

- We will take care of those definitions that are
  - tautological
  - too strict
  - . . .

Tautological DCs are easy to spot, when you spot DCs (esp. owned by ‘Athens Core’) that are imperfect in another way, let us know!

Flagged DCs

- Try to avoid linking with ‘deprecated’ or ‘superseded’ DCs !
  - do not use DCs with 2 definitions!!
- In other cases the flags show whether the DC specification is correct from a purely technical point of view
- Note that only DCs with a green marking are qualified for standardization

CLARIN-NL

Thank you for your attention.

Any questions?

ineke@ccl.kuleuven.be