CLARIN -NL
Common Language Resources and Technology Infrastructure

ISOcat usage in CLARIN NL/VL: results and challenges

Ineke Schuurman
ISOcat coordinator CLARIN-NL
Utrecht University & KU Leuven

Overview

1. User's perspective
    1. CLARIN
2. Why (not to) use ISOcat, some issues
    1. Standardization
    2. Harmonization
3. View
4. Do's and don'ts
5. Conclusion

CLARIN

CLARIN (ESFRI-project)

- building a research infrastructure at European level, for
- researchers in Human and Social Sciences (HSS) working with text (spoken, written, …), allowing them to make use of
- tools and resources already used by 'specialists' (in Language and Speech Technology)

CLARIN NL/VL

Up till now some 35 Dutch and Flemish projects use ISOcat

- Linguistic concepts
- Metadata

Broad range of domains, such as

- Syntax, Part of Speech, Named Entity Recognition, Co-reference, Sign Language, Speech, …

CLARIN NL/VL

Our material: existing corpora, tools , ...

- i.e. not necessarily using the latest (versions of) standards, if any,
- but widely used, in several cases having become *de facto* standards

► Expecting that DCs as defined in ISOcat can just be reused is not realistic in all cases

Common desiderata HSS

Tools and resources that are easy to use

Crucial:

- Easy to combine
- Easy to compare results

➔What does *X* mean ? Is it used the same way in *A* and *B* ? How is it related to *Z*?

➔**ISOcat** !

Our desiderata (providers)

Systematic sharing of resources should be possible

- 'Referability' of used linguistic concepts
- Long term preservation of these
- Interoperability

=> ISOcat (+RELcat and SCHEMAcat) !!

Role of ISOcat

ISOcat offers means

- to determine the meaning of a linguistic 'concept' in a document
- to compare several uses of a specific concept
  - over several documents

  and/or
  - over several languages

Some issues

However,

- CLARIN groups are not always really satisfied with ISOcat
- Some researchers/groups, also outside NL/VL, are not really prepared to make use of ISOcat

1. **Standardization / harmonization**
2. **reliability / trust**
3. usefulness in actual practice
4. user-friendliness

Standardization/standards

- Hardly any standardization after 4 years, plus
- Few existing standards, even ISO ones, incorporated (like TEI, EAGLES, …)

- Some (potential) users expect to find THE one and only (everlasting) standard in ISOcat, whereas
- others expect ISOcat to contain very specific (language/project) descriptions

➔Somewhere in the middle?

Harmonization

- Many existing DCs are **almost** identical (principled/pragmatic/arbitrary reasons)
- There are **many** similar DCs in ISOcat
- Several DCs in ISOCAT are not defined clearly

Standardization/harmonization

ISOcat:

- Not necessarily **just one** DC per concept over all profiles
- So, why would there be just one per profile?
  - Theories, schools, …
- On the other hand, proliferation should be prohibited!

HOW??

Reliability / trust

Current DCs are not stable (1)

- Minor changes (correction of typing errors, etc), but also
- **Major, content related changes**
  - In definition
  - In profile

    private ► more specific
  - In administrative status (superseded, deprecated)
  - (In scope – 'withdrawing')

As a consequence,

a specific DC you adopted may *have become* unsuitable for you !!

Plus

another DC may have become standardized

➔Currently you are to check all DCs you adopted time and again!

- ➔People prefer to come up with their own definitions
- ➔Proliferation of DCs

Reliability / trust - 2

- DCs are not stable - 2
  - They may be withdrawn (by becoming ‘private’ after having been ‘public’)

As a consequence, a DCS may contain a ‘untraceable’ (for guest) DC

➔Currently you are to check all DCs you are adopted time and again!

Reliability / trust

Currently, major semantic changes are not always easily traceable in case the original is not deprecated or superseded!

Same holds for standardization of related DC

HOW ??

Preliminary conclusion:

ISOcat currently is indeed not as useful and userfriendly as desired when applications are concerned:

- Procedure wrt semantic changes
- Avoidance of proliferation

- Our CLARIN NL/VL approach:
  - **View**
  - **Do's and don'ts**

CLARIN NL/VL

CLARIN NL/VL **view**:

➔ Users not confronted with all ISOcat entries (esp. the not-reusable ones)

- too specific (language, project)
- Incorrect (no proper definition, …)

➔ Instead creation of set of 'recommended' DCs which are to be adopted

- unless shown 'unfit' in a particular case

Avoidance of proliferation-1

Course of events in CLARIN NL/VL:

- Select an **existing** DC for a specific concept which suits your needs
  - for a specific tagset, domain, application, language, …
- preferably a standardized one,
- if not available or reuseable

➔create a new DC

Avoidance of proliferation -1

When a new DC is created:

➔Make clear why a standardized / recommended DC could not be adopted

➔For the time being: DCs by ‘authorities’

➔In case of minor discrepancies the co-ordinator will contact the owner to look whether the original could be adapted

Avoidance of proliferation - 2

- Standards:
  - very useful
- Items created for ISOcat in order to be standardized, covering as many standards, languages, theories etc as possible:
  - less useful

The latter tend to become rather trivial, noncommittal, vague, …, and therefore not really useful in a specific application

Avoidance of proliferation -3

Other DCs are far too specific

- Language
- Project
- Application

These are also of no use for most other applications, such DCs are to be avoided in the CLARIN NL/VL view:

be as generic as possible!

Consequences

1. new DCs are being added, resulting in several DCs for a specific data category (such as *noun, token, foreign word, manuscript*)
   - Not all instantiations contained in our view !
2. But also: references in definitions etc are no longer self-evident
3. When not disambiguated, such a definition is bound to remain vague (and therefore not useful)

Some do's and don'ts

Do's

- Disambiguate linguistic notions used in definition by mentioning their PID (note section)
- Explain why existing (standardized) DC can not be reused (explanatory comments (part of adm. section))
- Make definitions short and to the point

Some do's and don'ts -2

Do's

- Note in a separate schema the relations between DCs used (for RELcat)
- Mention a 'parent' DC whenever possible
- Make a DCS (with project name) containing **ALL** DCs you are using (adopted ones, new ones, 'linked' ones)

Some do's and don'ts

Don'ts

- Mention project, language etc in definition, name, …
- Use 'circular' definitions
- Use definitions á la 'a definite article is an article that is definite' unless both 'article' and 'definite' have a DC of their own (mentioned in note section)

Conclusion

ISOcat has lots of potential, but needs

- Standardization,
- Harmonization,
- Cleaning
  - Try outs, etc

Thanks for your attention!