(Semantic) Interoperability in the CLARIN Infrastructure

Menzo Windhouwer
The Language Archive - DANS
menzo.windhouwer@dans.knaw.nl

Succeed Interoperability Workshop
2nd October 2014
KB, The Hague, Netherlands
CLARIN

- CLARIN = Common Language Resources and Technology Infrastructure = an european ESFRI infrastructure project
- Aims at providing easy and sustainable access for scholars in the humanities and social sciences to digital language data (in written, spoken, video or multimodal form) and advanced tools to discover, explore, exploit, annotate, analyze or combine them, independent of where they are located.

http://www.clarin.eu/
CLARIN & Interoperability

- CLARIN Centre Committee
  - Sets technical standards for the infrastructure to operate
- CLARIN Standards Committee
  - Standards/Recommendations for linguistic resources
    - http://clarin.ids-mannheim.de/standards/
- Various (national) coordinators
  - Recommendations and best practices for specific topics
    - CMDI and ISOcat
- My own focus: achieve a level of semantic interoperability
(meta)data schema building blocks have semantics
- Make semantics explicit and share them, i.e., identify common (emerging) concepts
- The concept registries used by CLARIN are lightweight
  - CLARIN doesn’t try to construct a ‘global’ domain ontology
  - Relationships are on the CLARIN scale suitable for resource discovery, not for formal reasoning
- Technology
  - Resources are taken as they are, i.e., no CLARIN scale conversion to RDF
  - Metadata is XML-based, but some RDF experiments are underway
ISOcat Concept Registry

http://www.isocat.org/
CMDI: Component Editor

http://catalog.clarin.eu/ds/ComponentRegistry/
Virtual Language Observatory

http://catalog.clarin.eu/vlo/
Good vs Bad & the future

The CMD Infrastructure is very flexible with regard to metadata structures, but also provides an integrated light weight semantic layer to achieve semantic interoperability and overcome the structural differences 😊

- But sometimes more context needs to be taken into account
- Do deal with emerging semantics and specific needs many registries are open, which leads to proliferation 😞

- The ISOcat registry operated together with ISO TC 37 is too complex 😞
  - Move to a simpler model based on SKOS
  - Move to a simpler workflow based on CLARIN recommendations
    - The organisation problem is harder than the technical one!

- CLARIN-NL experimented with the same kind of semantic interoperability for linguistic resources
  - Might be exploited by a higher level of the CLARIN Federated Content Search (currently supports full text search)
  - How to make semantic annotation a part of a researcher’s workflow?
    - Embed interaction with the registries in their tools