



Lightweight Semantics in the CLARIN Infrastructure

Menzo Windhouwer

The Language Archive - DANS

menzo.windhouwer@dans.knaw.nl

Matej Durco

Institute for Corpus Linguistics and Text Technology

matej.durco@assoc.oeaw.ac.at

EUON workshop 2014

Amsterdam, Netherlands

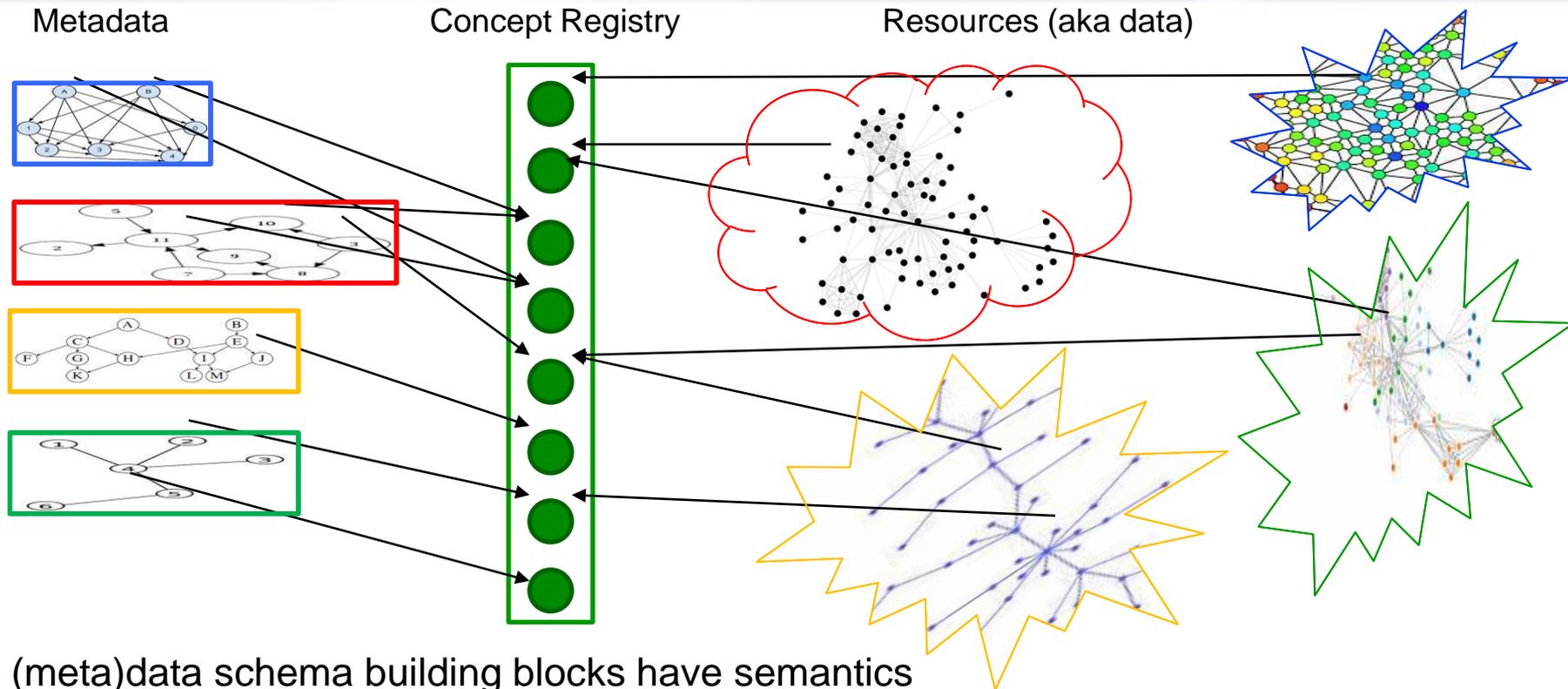
CLARIN



- 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.
 - Building a networked federation of European data repositories, service centers and centers of expertise.

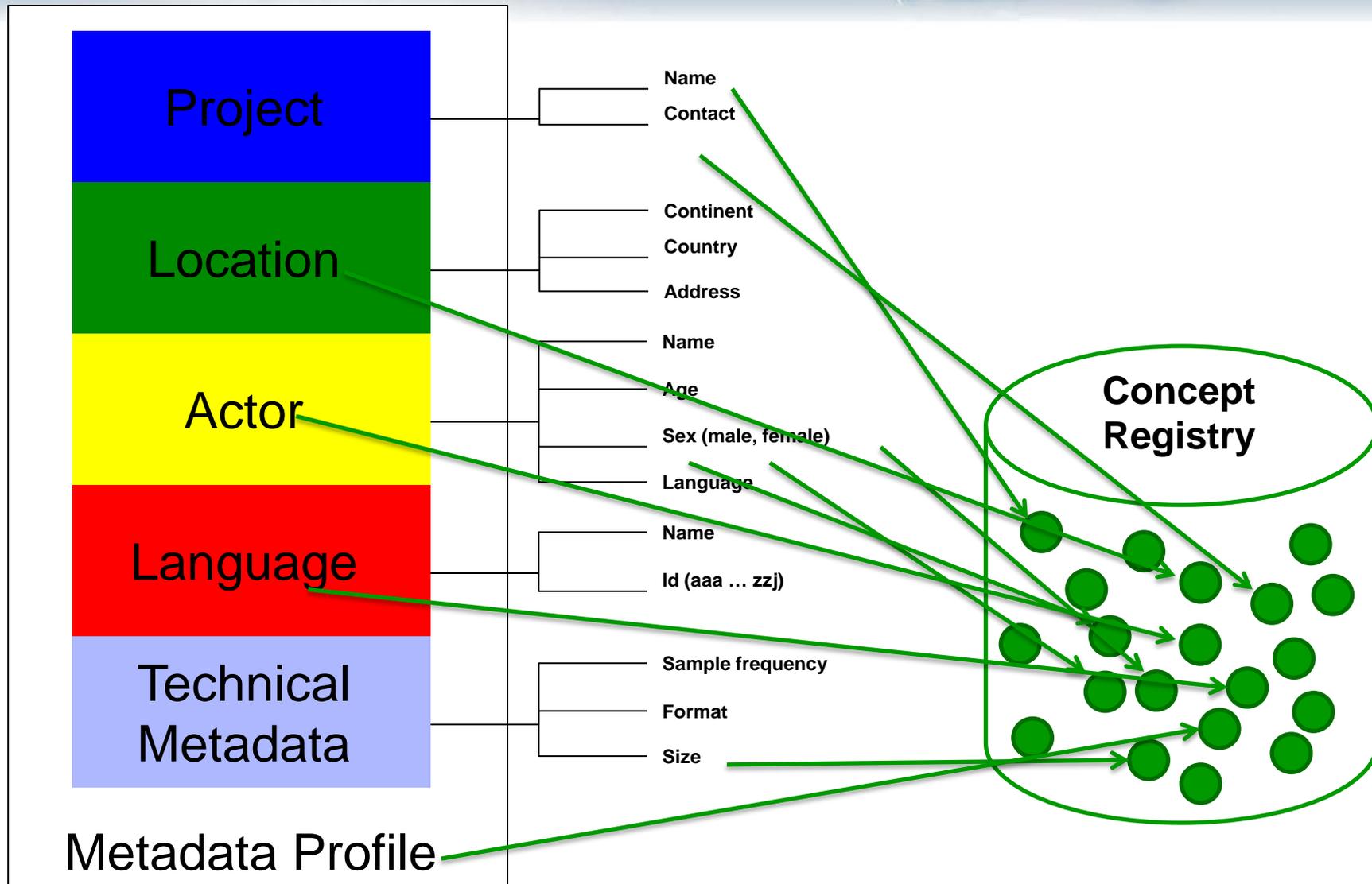
<http://www.clarin.eu/>

CLARIN and Semantics

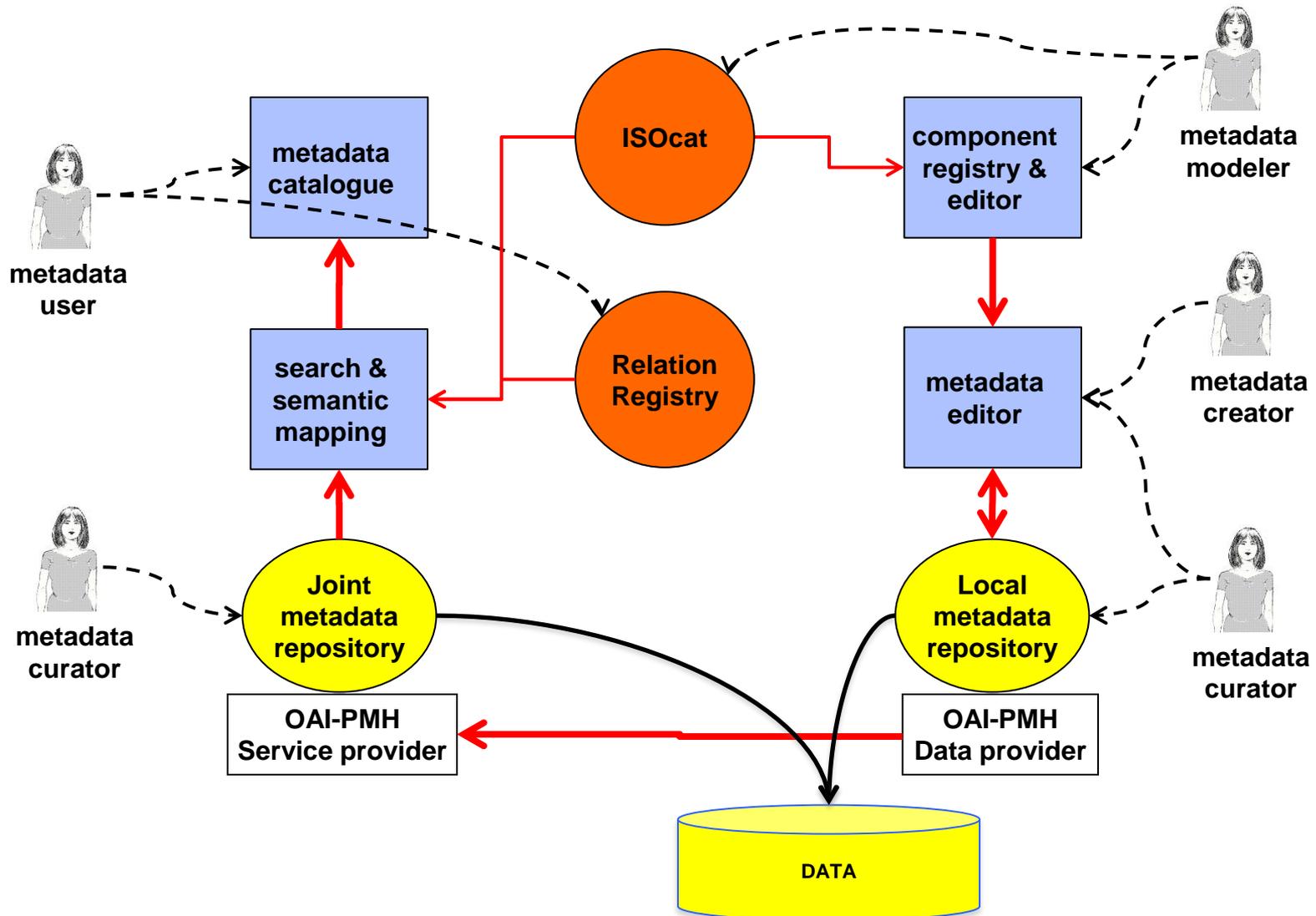


- (meta)data schema building blocks have semantics
 - Make semantics explicit and share them, i.e., identify common (emerging) concepts
- The concept registries used by CLARIN are lightweight
 - CLARIN doesn't try to construct a 'global' domain ontology
 - Relationships are on the CLARIN scale suitable 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

Component Metadata Infrastructure



CMDI workflow



Virtual Language Observatory



Virtual Language Observatory 

Explore the world of language resources and technology from different perspectives

CLARIN WALS THE KONINKLIJKE BIBLIOTHEEK DER CONINKLIJKE AKADEMIE VAN LETTEREN DOBES

VLO > [Faceted search](#) > [Selections](#): Dutch x [Link](#) | [Report](#) | [Help](#)

SEARCH

SEARCH RESULTS

36987 results << < 1 2 3 4 5 6 7 8 9 10 >>> Showing 1 to 10

" De professionaliteit van medisch specialisten onder de loep"

Resources: | 1 other |

"...Een canaal ende deurganck..." : een onderzoek naar de binnenvaart door Holland in de lange Gouden Eeuw (1572-1700)

Aan de hand van de tol- en sluisrekeningen van Gouda is veel te zeggen over de sociale en economische betekenis van de binnenvaart door Holland.

Resources: | 1 image |

"A Change is Gonna Come" Stax Records en de zwarte vrijheidsstrijd

Resources: | 1 image |

"A nation stays alive when its culture stays alive" - Het belang van culturele noodhulp geïllustreerd aan de hand van drie projecten van Cultural Emergency Response

In oorlogen en andere gewapende conflicten, aardbevingen, stormen en andere natuurrampen wordt altijd verwoesting aangericht. Niet alleen bij mensen, ook bij goederen. Wat onderbelicht blijft is dat bij al deze rampen ook vaak onvoorstelbare – al dan niet bewuste – schade is toegebracht aan cultuurgoederen. Pas sinds d...

Resources: | 1 image |

"Aanloopstraten in binnensteden De invloed van fysieke en organisatorische

NARROW DOWN

Use the categories below to limit the search results to those matching the selected value(s).

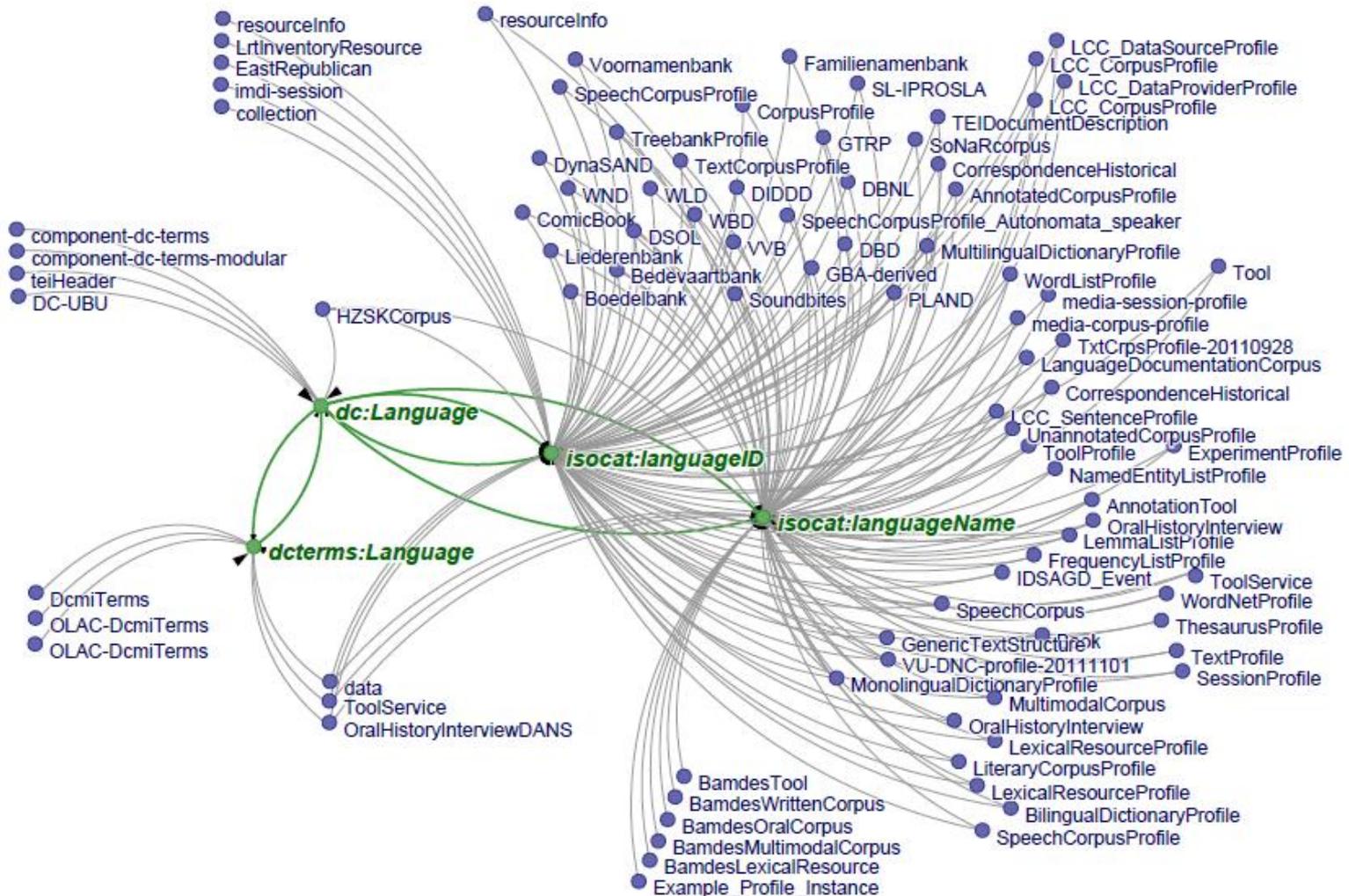
- + LANGUAGE
Dutch x
- + COLLECTION
- + RESOURCE TYPE
- + CONTINENT
- + COUNTRY
- + MODALITY
- + GENRE
- + SUBJECT
- + FORMAT
- + ORGANISATION
- + NATIONAL PROJECT
- + KEYWORD
- + DATA PROVIDER

+
 -
 x

+ SEARCH OPTIONS

<http://catalog.clarin.eu/vlo/>

CMD Semantic Mapping in CLARIN



<http://clarin.aac.ac.at/smc-browser>

Conclusion & future work



- 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
- Move to a next version of the Concept and Relation Registry
 - Simpler model (opposed to a Data Category Registry)
 - Simpler workflow (no formal standardization, but CLARIN recommendations)
- 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)