(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.

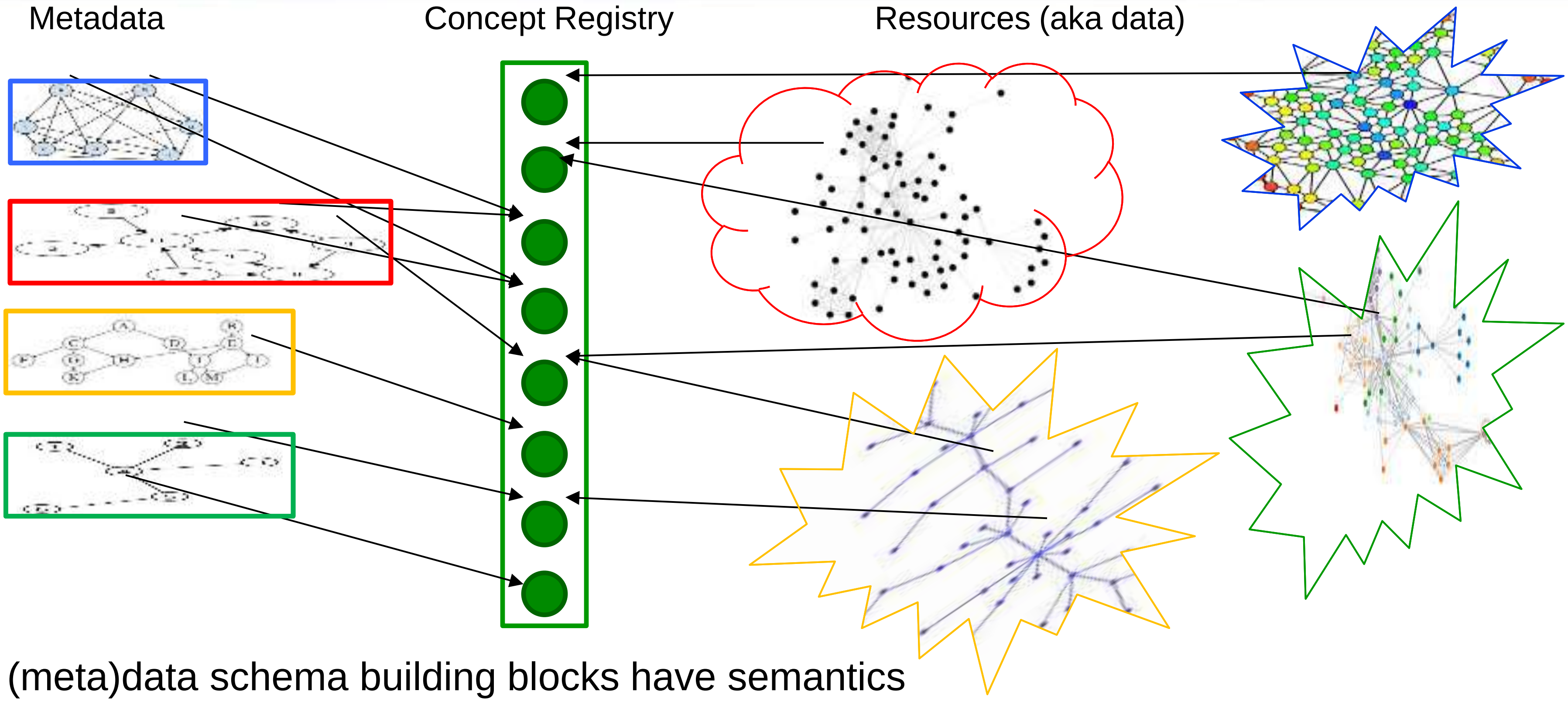
http://www.clarin.eu/

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

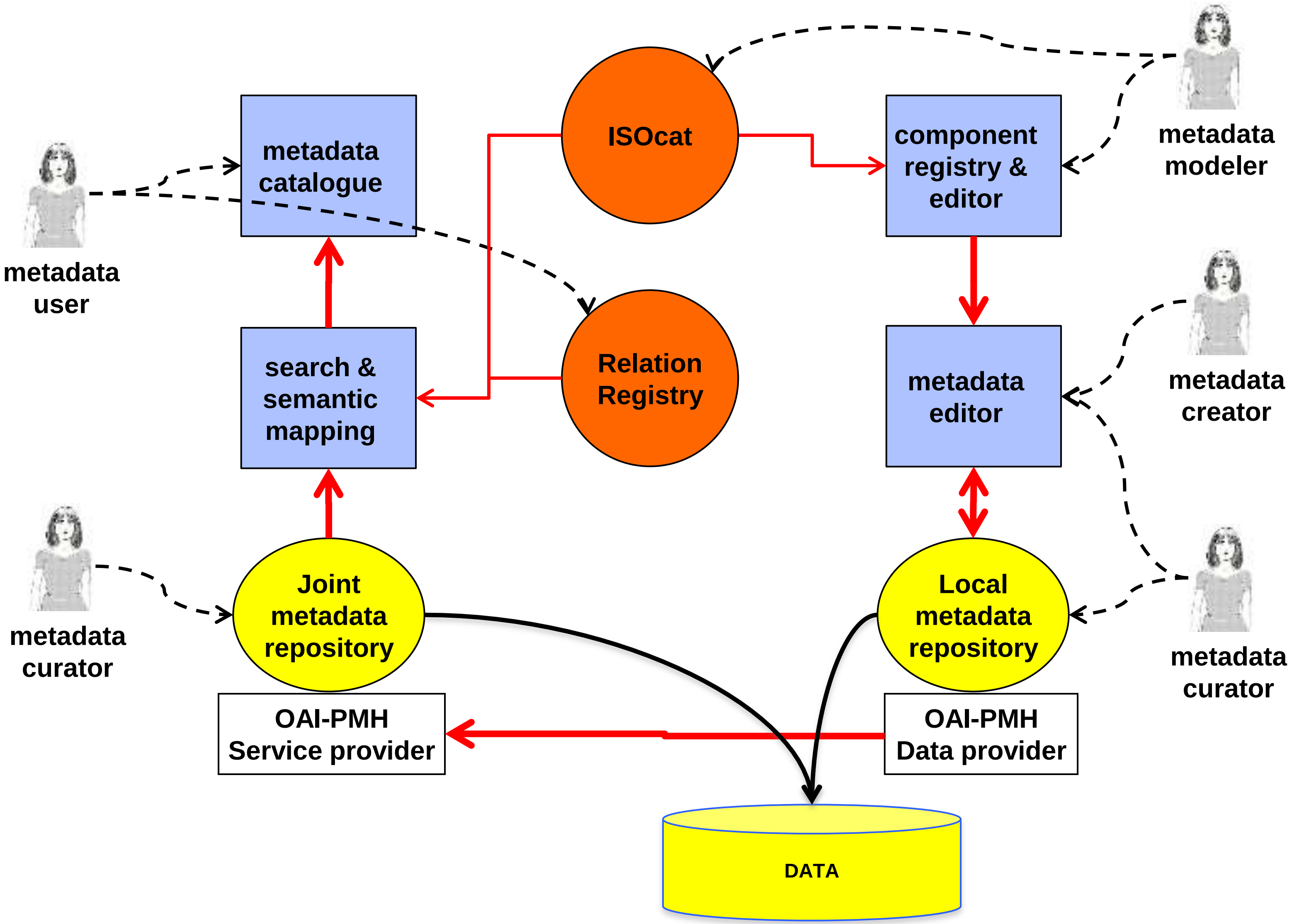
- My own focus: achieve a level of semantic interoperability

CLARIN and Semantics

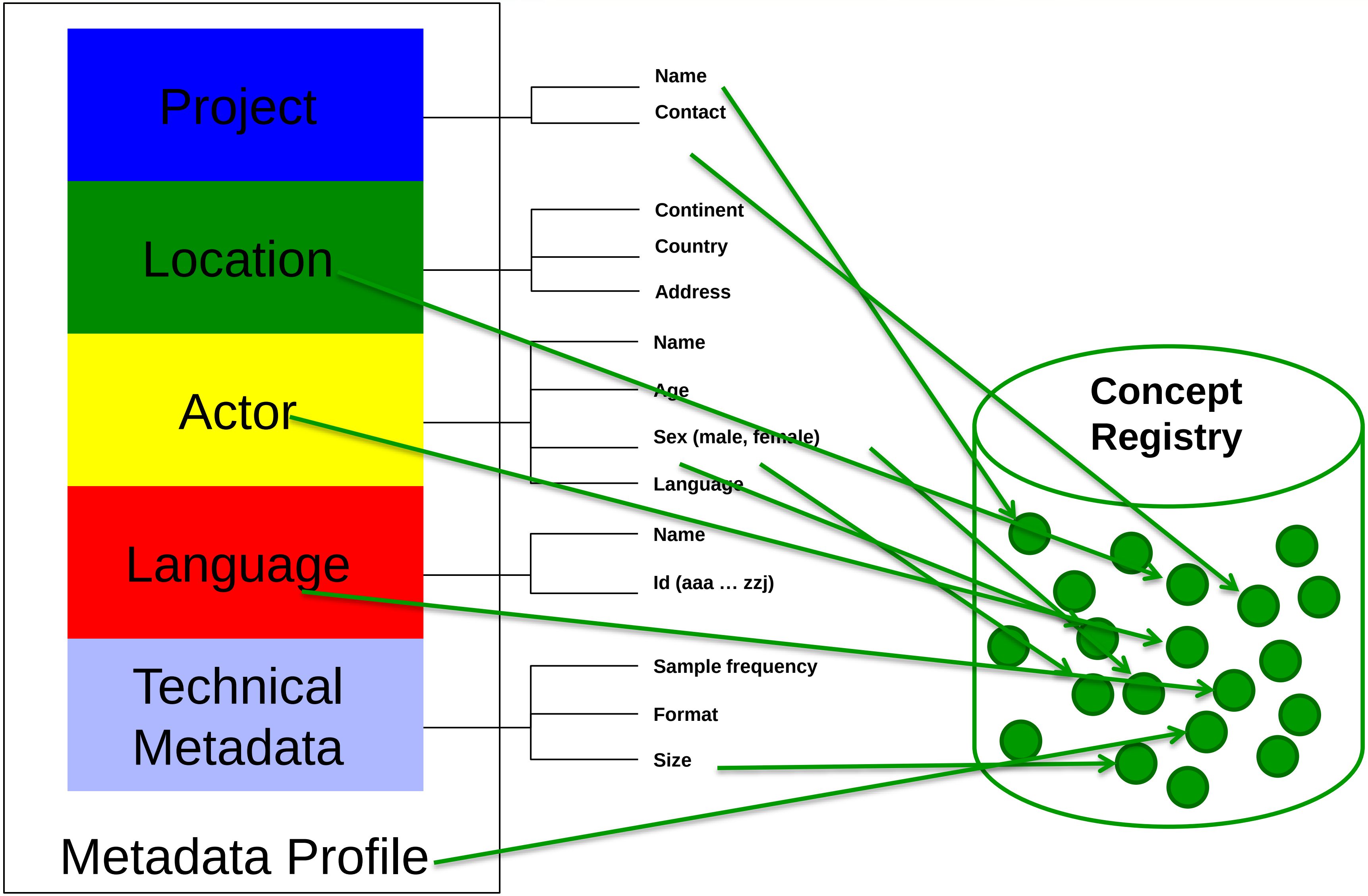


- (meta)data schema building blocks have semantics
  - Make semantics explicit and share them, i.e., identify common (emerging) concepts
- The concept registries used by CLARIN are lightweight
  - CLARIN doesn’t try to construct a ‘global’ domain ontology
  - Relationships are on the CLARIN scale suitable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

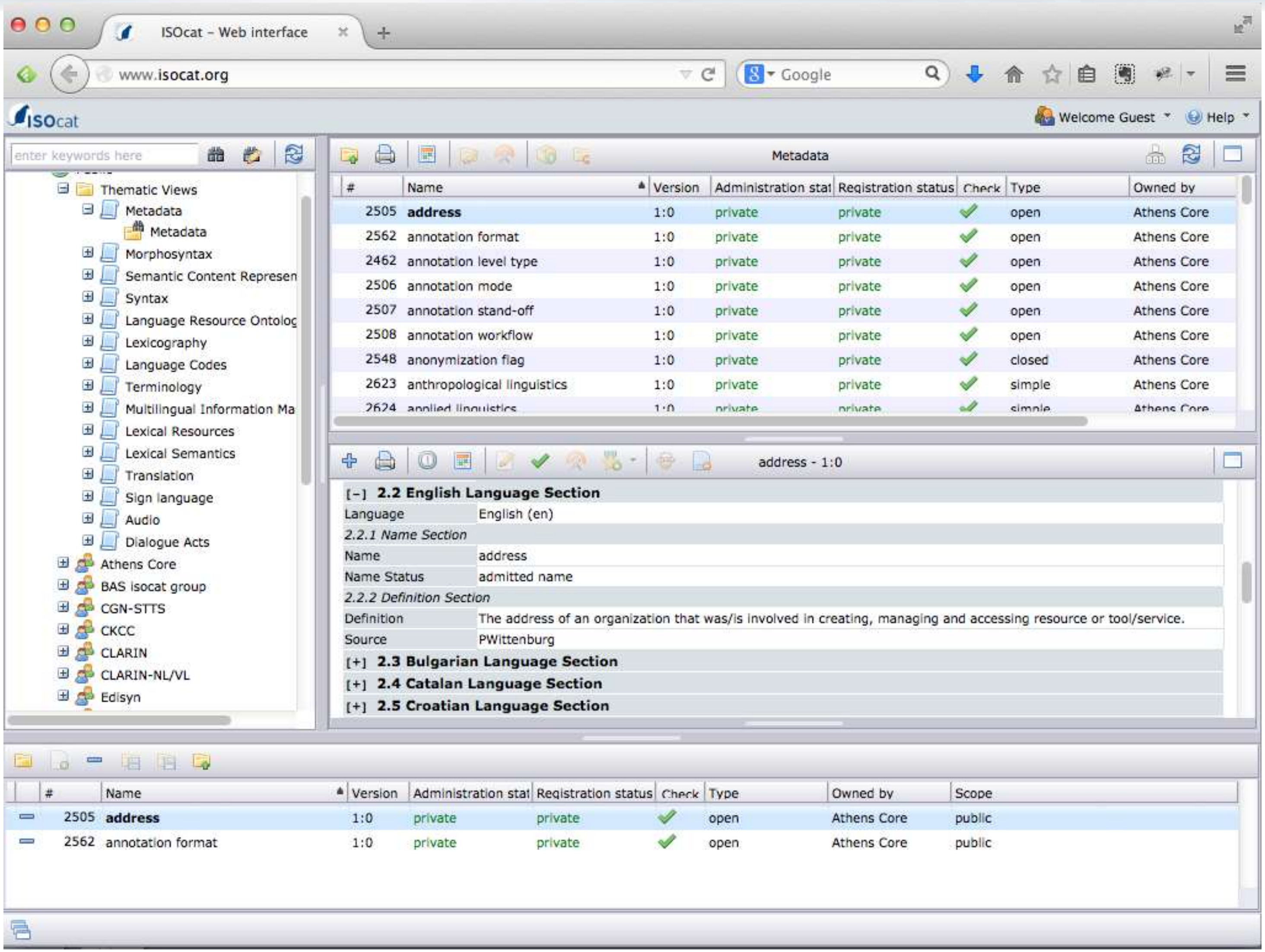


CMDI



Project
Location
Actor
Language
Technical Metadata
Metadata Profile
Name
Contact
Continent
Country
Address
Name
Age
Sex (male, female)
Language
Name
Id (aaa … zzj)
Sample frequency
Format
Size
Concept Registry

ISOcat Concept Registry



- Thematic Views
  - Metadata
    - Metadata
  - Morphosyntax
  - Semantic Content Represen
  - Syntax
  - Language Resource Ontolog
  - Lexicography
  - Language Codes
  - Terminology
  - Multilingual Information Ma
  - Lexical Resources
  - Lexical Semantics
  - Translation
  - Sign language
  - Audio
  - Dialogue Acts
- Athens Core
- BAS isocat group
- CGN-STTS
- CKCC
- CLARIN
- CLARIN-NL/VL
- Edisyn

#	Name	Version	Administration stat	Registration status	Check	Type	Owned by
2505	**address**	1:0	private	private	✓	open	Athens Core
2562	annotation format	1:0	private	private	✓	open	Athens Core
2462	annotation level type	1:0	private	private	✓	open	Athens Core
2506	annotation mode	1:0	private	private	✓	open	Athens Core
2507	annotation stand-off	1:0	private	private	✓	open	Athens Core
2508	annotation workflow	1:0	private	private	✓	open	Athens Core
2548	anonymization flag	1:0	private	private	✓	closed	Athens Core
2623	anthropological linguistics	1:0	private	private	✓	simple	Athens Core
2624	applied linguistics	1:0	private	private	✓	simple	Athens Core

address - 1:0

[-] **2.2 English Language Section**

Language	English (en)
2.2.1 Name Section	
Name	address
Name Status	admitted name
2.2.2 Definition Section	
Definition	The address of an organization that was/is involved in creating, managing and accessing resource or tool/service.
Source	PWittenburg

[+] **2.3 Bulgarian Language Section**

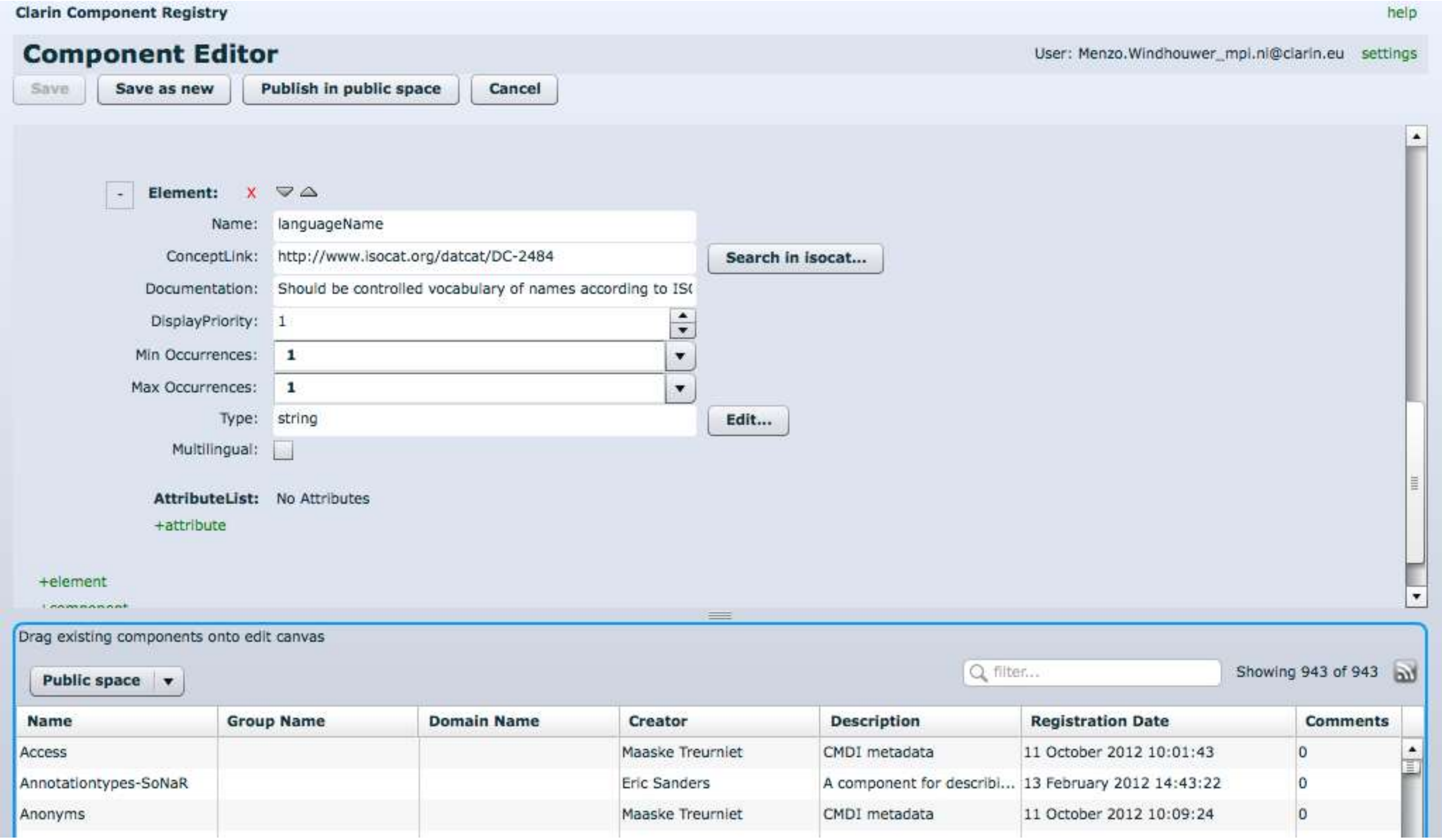
[+] **2.4 Catalan Language Section**

[+] **2.5 Croatian Language Section**

	#	Name	Version	Administration stat	Registration status	Check	Type	Owned by	Scope
−	2505	**address**	1:0	private	private	✓	open	Athens Core	public
−	2562	annotation format	1:0	private	private	✓	open	Athens Core	public

http://www.isocat.org/

CMDI: Component Editor



Component Editor

User: Menzo.Windhouwer_mpi.nl@clarin.eu

Element:

Name: languageName

ConceptLink: http://www.isocat.org/datcat/DC-2484

Documentation: Should be controlled vocabulary of names according to IS(

DisplayPriority: 1

Min Occurrences: 1

Max Occurrences: 1

Type: string

Multilingual: □

AttributeList: No Attributes

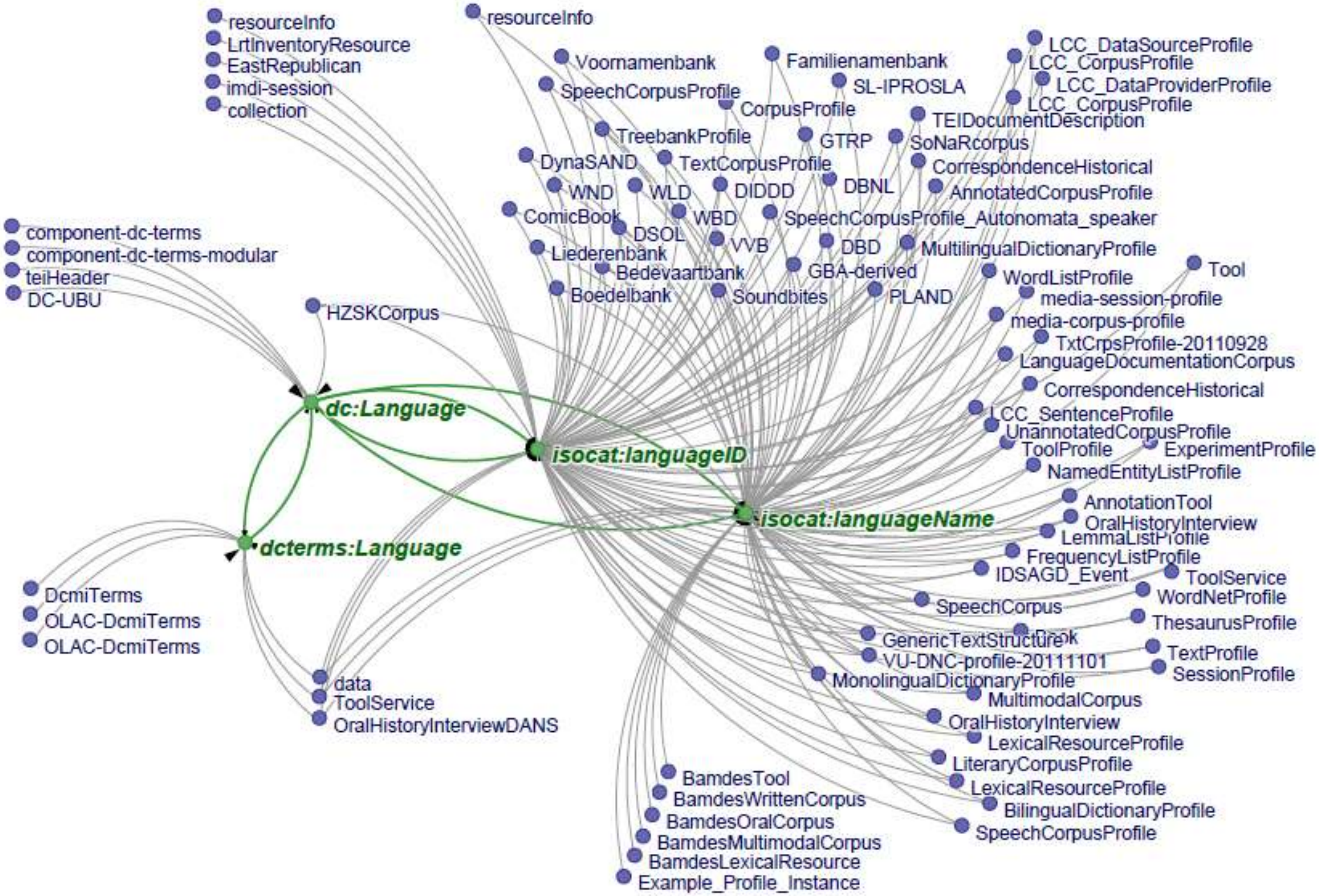
Drag existing components onto edit canvas

Showing 943 of 943

Name	Group Name	Domain Name	Creator	Description	Registration Date	Comments
Access			Maaske Treurniet	CMDI metadata	11 October 2012 10:01:43	0
Annotationtypes-SoNaR			Eric Sanders	A component for describi...	13 February 2012 14:43:22	0
Anonyms			Maaske Treurniet	CMDI metadata	11 October 2012 10:09:24	0

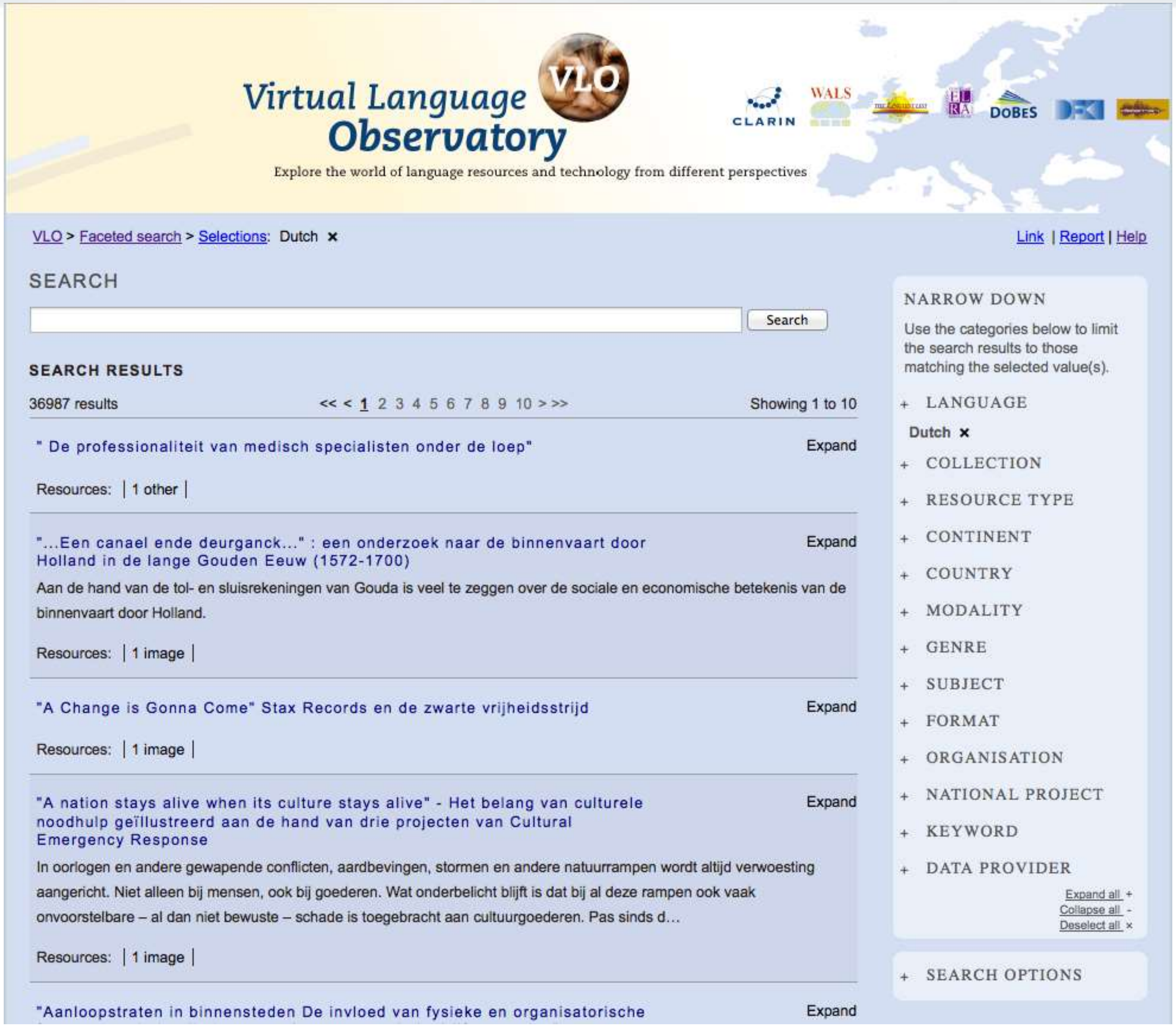
http://catalog.clarin.eu/ds/ComponentRegistry/#

CMDI: Semantic Mapping



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

Virtual Language Observatory


Virtual Language Observatory
VLO
Explore the world of language resources and technology from different perspectives
CLARIN
WALS
DoBES

VLO > Faceted search > Selections: Dutch ×

Link | Report | Help

SEARCH

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" Expand

Resources: | 1 other |

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

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 Expand

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 Expand

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 Expand

NARROW DOWN

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

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

Expand all +
Collapse all -
Deselect all ×

+ SEARCH OPTIONS

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