CLARIN
Common Language Resources and Technology Infrastructure

The CMD Cloud

Matej Ďurčo, Menzo Windhouwer
Institute for Corpus Linguistics and Text Technology, Vienna, AT
The Language Archive - DANS, The Hague, NL
matej.durco@oeaw.ac.at, menzo.windhouwer@dans.knaw.nl

Context & Problem Statement

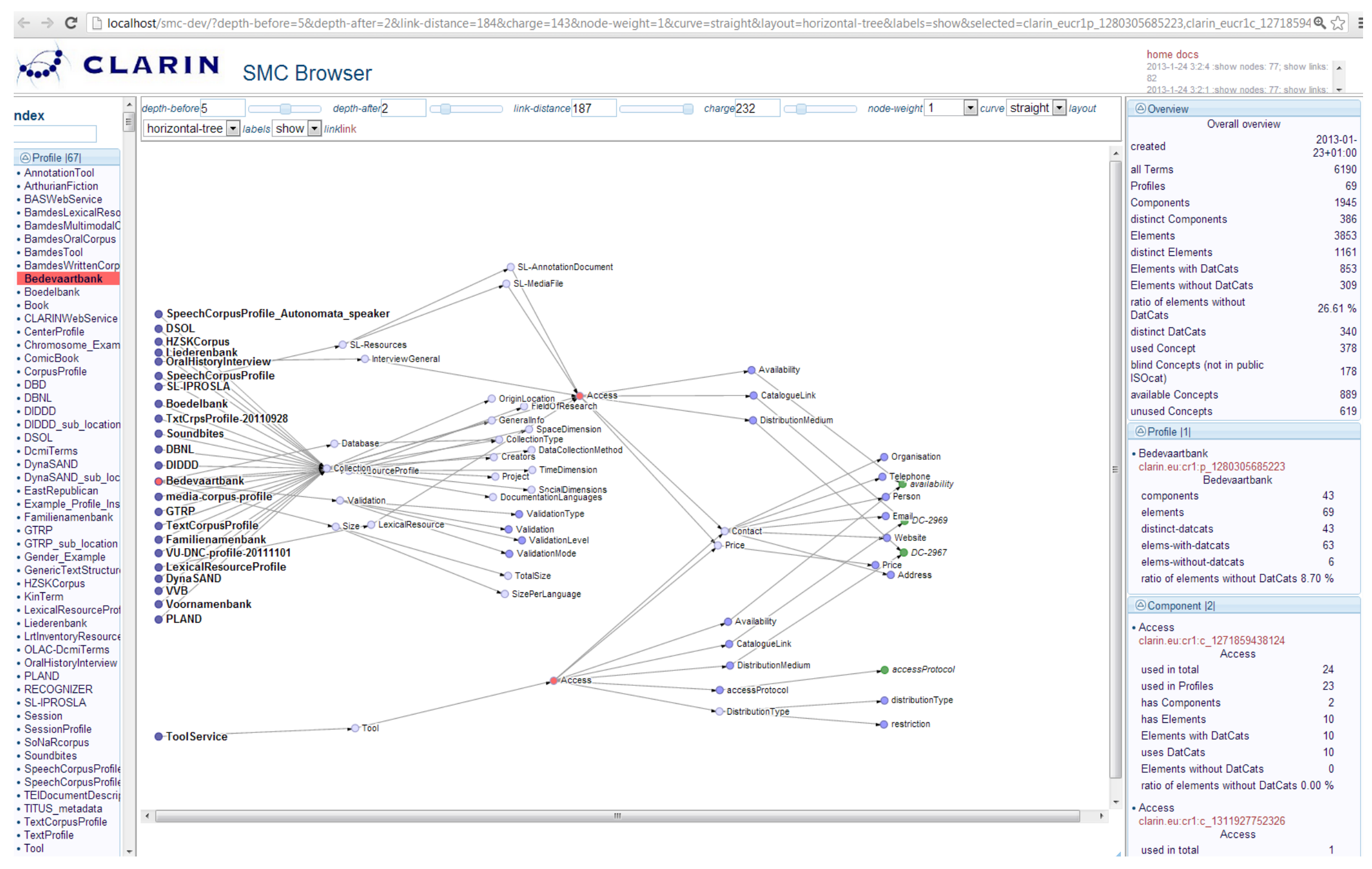
CLARIN Component Metadata Infrastructure (CMDI) established means for flexible resource descriptions for the domain of language resources with sound provisions for semantic interoperability weaved deeply into the meta model and the infrastructure. The data domain rapidly growing in both size and complexity requires advanced means for inspection and analysis of the data on schema and instance level to be used by the metadata modellers, editors and curators .

CMD Graph

The reuse of components and data categories yields the CMD data as a graph blending the component trees of individual profiles. The resulting graph consists of over 4.600 nodes and 7500 edges requiring an interactive interface that allows to select nodes of interest and dynamically investigate the contextual subgraph.

SMC Browser

Semantic Mapping Component (SMC) is one module within CMDI designed to overcome the semantic interoperability problem stemming from the heterogeneity of the resource descriptions, by harnessing the provisions for sharing semantics built into the CMDI.
One part of the SMC module is the SMC Browser, a web application that visualizes the CMD entities (profiles/schemas, components, elements and data categories) as an interactive graph enabling the metadata modeller to examine the reuse of components or data categories in different profiles/schemas.



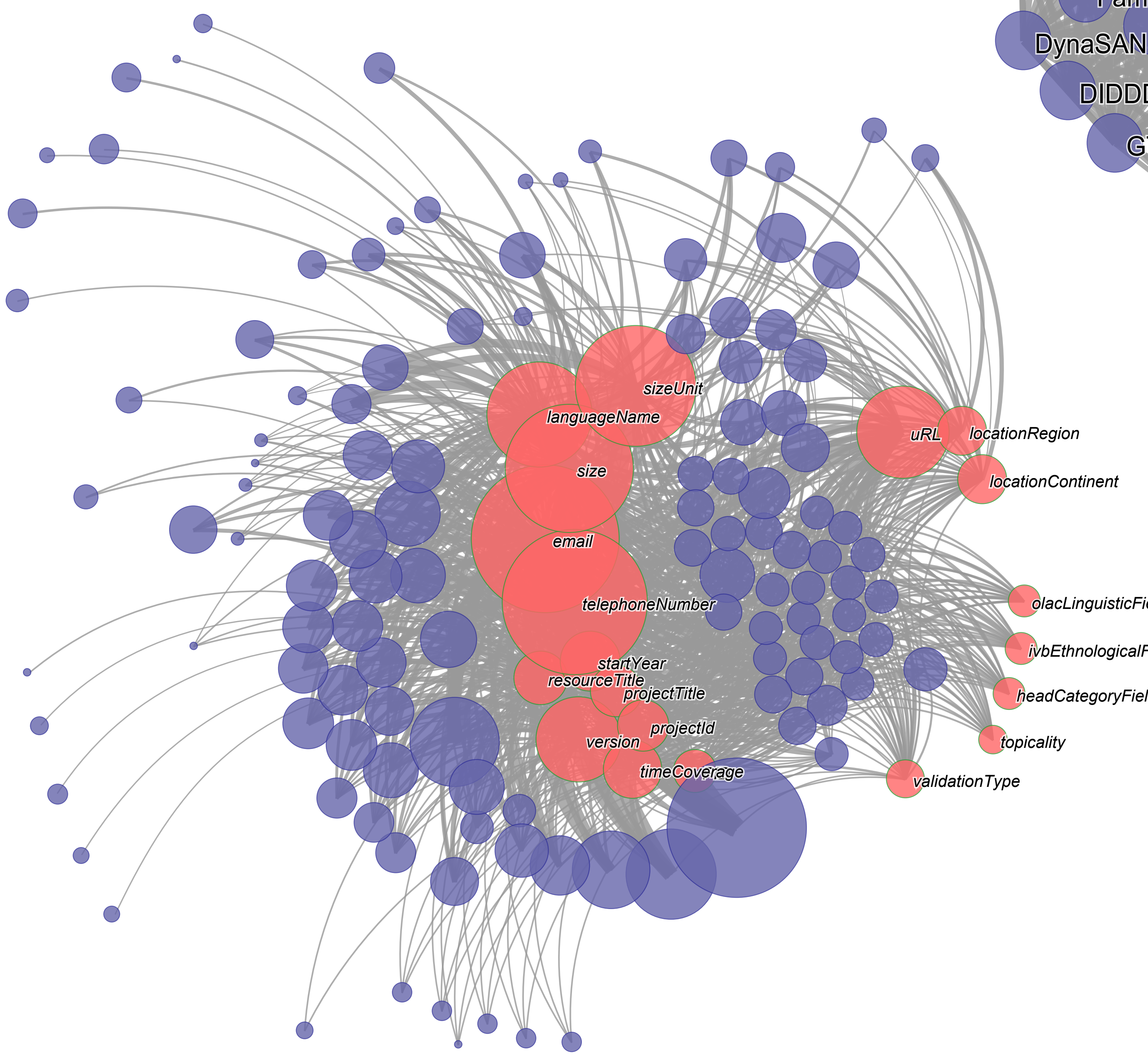
Profile similarity

The ratio of data categories shared by profiles implies their semantic proximity. In the basic setup, the pairwise similarity ratio is computed as the average of the quotients of matching distinct data categories of the two profiles.

$$sim_{p1} := \frac{count(distinct(Datcats_{match}))}{count(distinct(Datcats_{p1}))}$$

$$sim_{p2} := \frac{count(distinct(Datcats_{match}))}{count(distinct(Datcats_{p2}))}$$

$$sim := \frac{(sim_{p1} + sim_{p2})}{2}$$



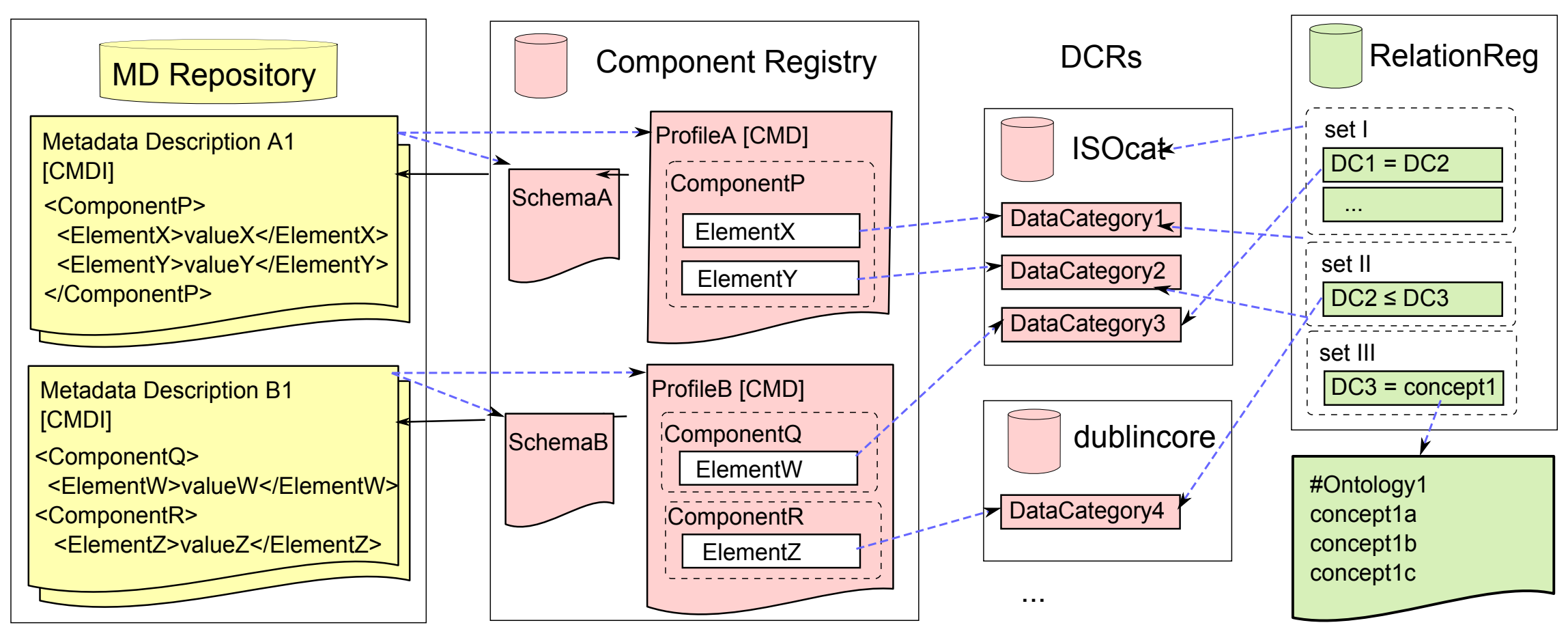
The reuse of some very common data categories by different schemas

SMC is part of Austrian contribution to CLARIN

CMD - Data Domain

Within CMDI, metadata records are based on XML schemas generated from profiles maintained in the Component Registry. Profiles are constructed out of reusable components and elements, linked to data categories – well-defined concepts maintained in a data category registry – for semantic grounding. This setup allows for high flexibility in modelling the metadata structures, while establishing a shared semantics layer.
Additionally, in the RELcat Relation Registry links between near equivalent concepts can be created, introducing another ad hoc mapping layer.

Relations between pieces of data in the CMD data domain and corresponding CMDI modules



CMD data domain over time: # of public CMD profiles, components and elements

	2011-01	2012-06	2013-01	2013-06	2014-03
Profiles	40	53	87	124	153
Components	164	298	542	828	1.110
Elements	511	893	1.505	2.399	3.101
Distinct DCs	203	266	436	499	737
Elements without DCs	24,70%	17,60%	21,50%	26,50%	24,20%

Data Categories

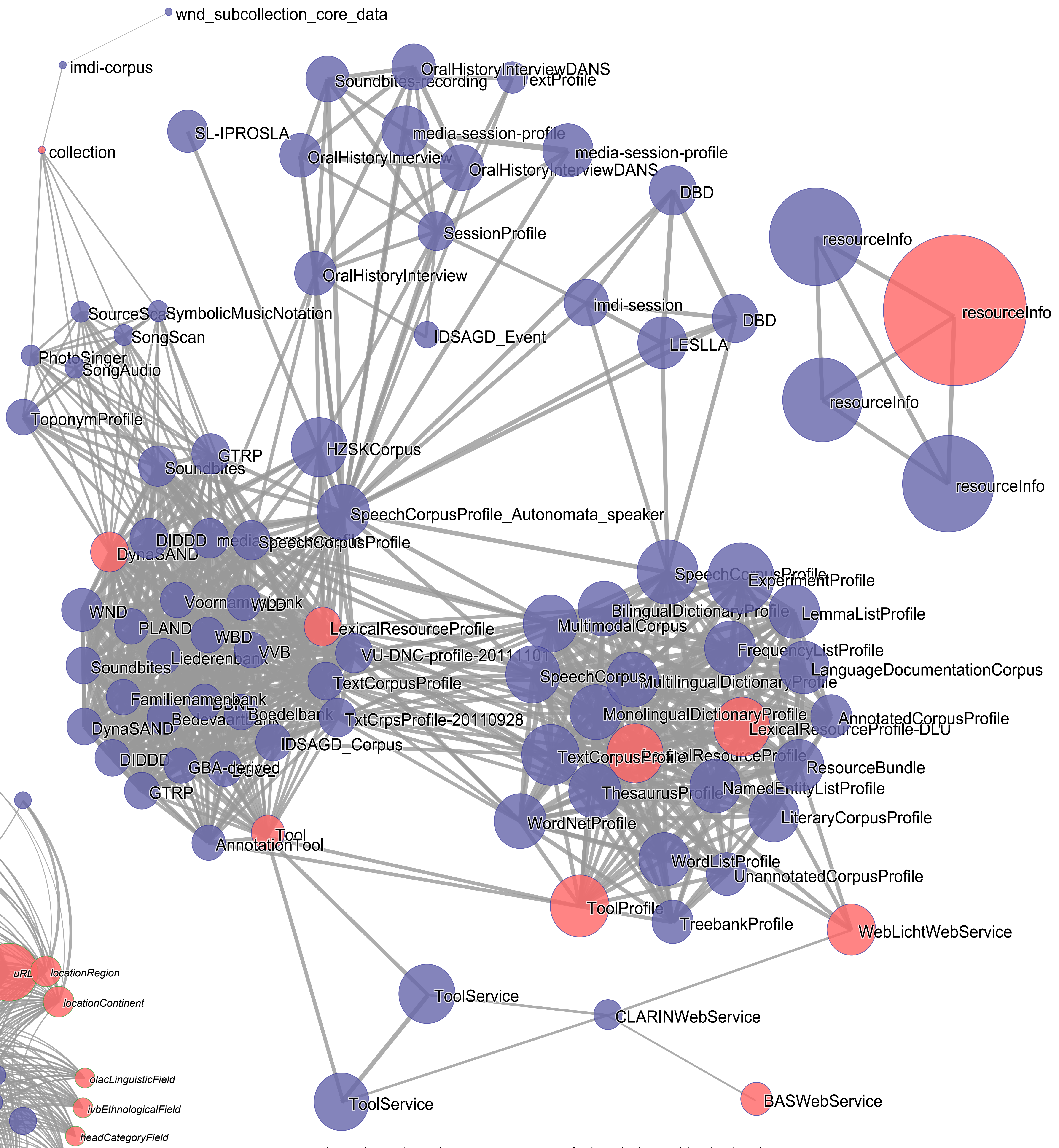
# of CMD profiles and elements referencing a DC [2014-05]

132/2363	description [isocat:DC-2520]
119/373	languageID [isocat:DC-2482]
117/322	languageName [isocat:DC-2484]
115/477	email [isocat:DC-2521]
113/115	resourceTitle [isocat:DC-2545]
111/160	resourceName [isocat:DC-2544]
110/237	mimeType [isocat:DC-2571]
106/534	address [isocat:DC-2505]
103/456	telephoneNumber [isocat:DC-2461]
101/366	size [isocat:DC-2580]
100/449	Organisation [isocat:DC-2979]
99/356	Person [isocat:DC-2978]
94/144	availability [isocat:DC-2453]
88/216	version [isocat:DC-2547]
88/112	publicationDate [isocat:DC-2538]
87/92	projectName [isocat:DC-2536]
87/91	projectTitle [isocat:DC-2537]
84/102	timeCoverage [isocat:DC-2502]
84/89	projectId [isocat:DC-2535]
80/117	completionYear [isocat:DC-2509]
79/115	startYear [isocat:DC-2539]
79/79	legalOwner [isocat:DC-2956]
77/651	url [isocat:DC-2546]
77/81	funder [isocat:DC-2522]
76/105	DistributionMedium [isocat:DC-2967]
76/132	CatalogueLink [isocat:DC-2969]
75/314	sizeUnit [isocat:DC-2583]
75/87	quality [isocat:DC-2574]
74/86	price [isocat:DC-2460]
68/91	resourceClass [isocat:DC-3806]
68/93	modalities [isocat:DC-2490]
65/75	characterEncoding [isocat:DC-2564]
65/167	locationCountry [isocat:DC-2532]
60/60	PID [isocat:DC-2573]
58/209	uRL [isocat:DC-63]
57/59	dominantLanguage [isocat:DC-2468]
54/62	locationAddress [isocat:DC-2528]

Profiles

# of metadata records per CMD profile [2014-03]

155.403	Song
138.821	Session
104.991	OLAC-DcmiTerms
70.577	mods
46.157	DcmiTerms
31.827	media-session-profile
28.448	SongScan
21.256	SourceScan
16.519	Source
14.811	imdi-corpus
8.508	IDSAGD_Speaker
8.109	IDSAGD_Event
7.961	SongAudio
7.810	teiHeader
7.557	SymbolicMusicNotation
4.485	LCC_DataProviderProfile
4.417	Text
2.950	ArthurianFiction
2.183	LrtInventoryResource
1.982	Soundbites-recording
1.952	SL-IPROSLA
1.530	Performer
1.466	DiscAn_Case
1.303	teiHeader
998	Etstoel
916	teiHeader
775	OLAC-DcmiTerms-ref-DWR
697	OLAC-DcmiTerms-ref
613	GTRP_sub_location
583	JacobsstafVerhaal
443	GBA-derived_sub_municipality
399	ToponymProfile
399	Communication_Transcript
397	Communication_Recording
333	DIDDD_sub_location
267	DynaSAND_sub_location
187	data



Sample graph visualizing the semantic proximity of selected schemas (threshold=0.6)

Future Work

- apply other factors for similarity computation
  (label string distance, structure, value domain)
- integration of instance data
  (to analyze actual use of components and elements)
- integration with a continuous metadata curation process
- refactor SMC Browser as a generic interactive graph viewer

ICLTT The Language Archive DANS

http://clarin.oeaw.ac.at/smc/