Semantic Mapping in CMDI

2012-05-21 - LREC 2012 - ISOcat-workshop

Matej Ďurčo, ICLTT, Vienna;

Introduction

- CLARIN – Common Language Resources and Technology Infrastructure

- CMDI – Component Metadata Infrastructure
  CLARIN‛s technical heart delivers a
  heterogeneous collection of metadata (about language resources)

- ISOcat
  framework within ISO TC 37 for defining:

- data categories
  definitions of widely accepted linguistic concepts

- Task:
  enhance the search in the heterogeneous collection
  employing the defined data categories

  => semantic search

CMDI modules

- DataCategoryRegistry – ISOcat (, dublincore, …)

  define/standardize a reusable set of (basic) data categories

- CMDI - ComponentRegistry

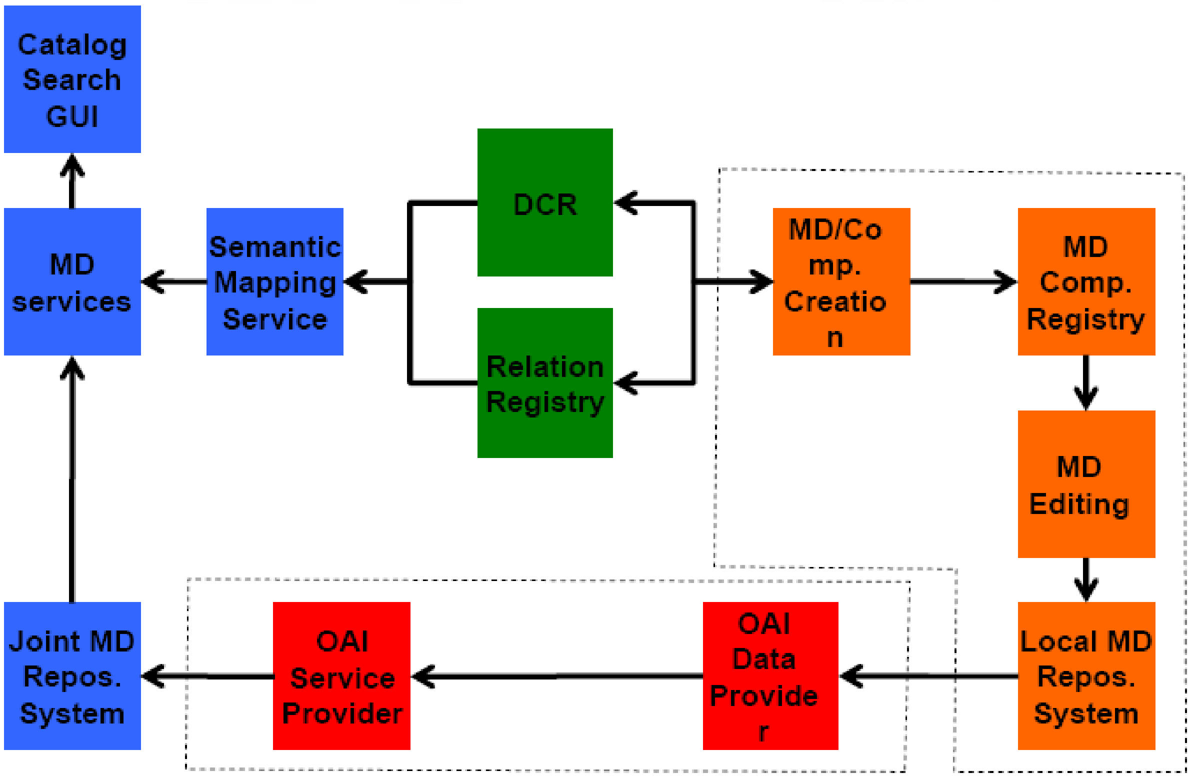
  define profiles/schemas at will, but reference data categories!

- RelationRegistry

  allows defining relations between data categories

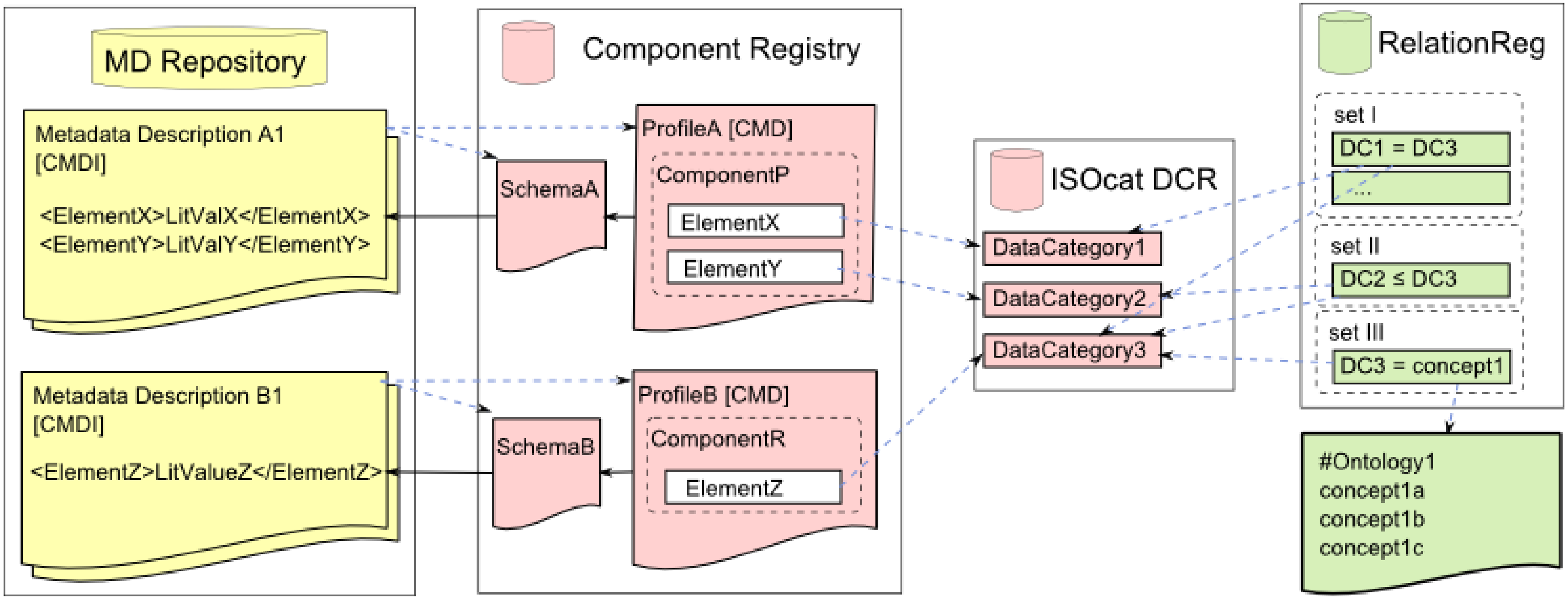
exploitation side:

- MD Repository
- MD Service
- MD Browser
- VLO, …



CMDI linking

- components and elements in CMD profiles are bound to data categories
- the CMD records reference their profiles
- in Relation Registry data categories are related to each other
  in separate (possibly overlapping/contradicting) relation sets

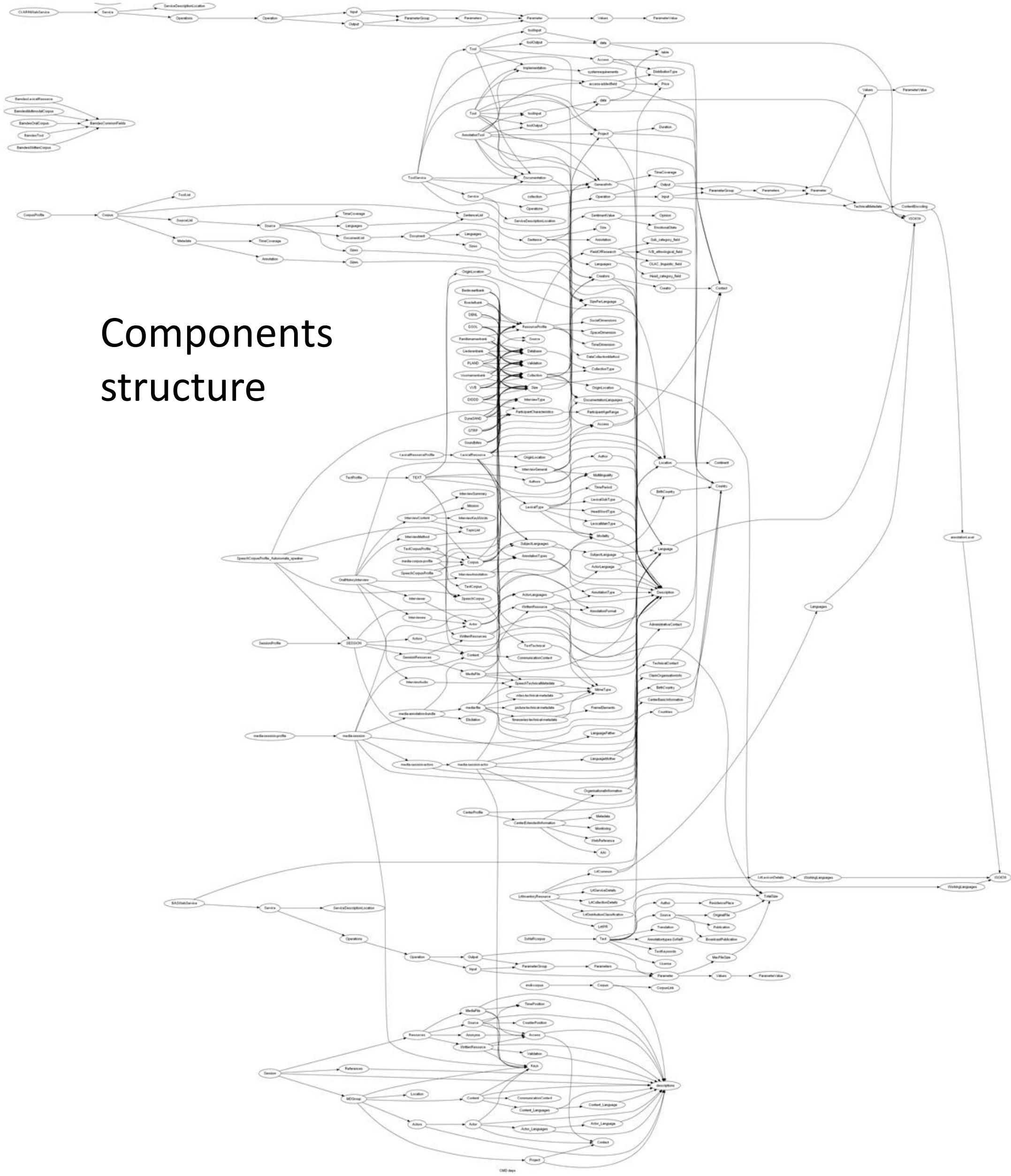


DCR usage in Component Registry

Data Categories Sets	827
isocat (Metadata Profile#5)	712
dublincore elements	16
dublincore terms	99

Component Registry	
CMD-Profiles	53
standalone Components	235*)
overall Components	298
distinct Elements	893
all Elements	3.030
all paths (profile/comp/elem	4.565

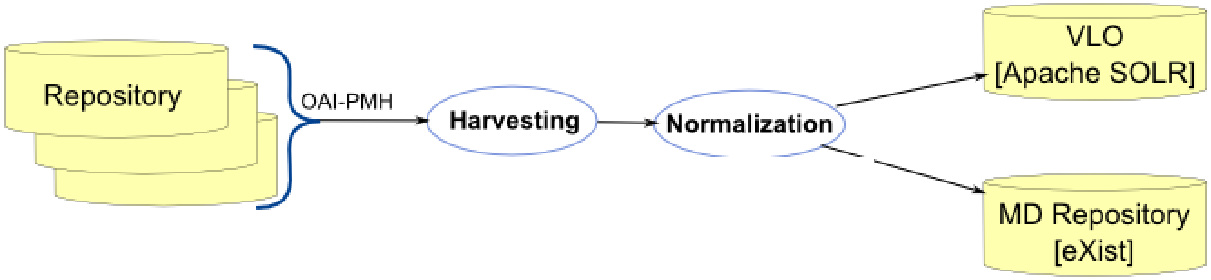
Datcats in CompReg	288
ISOcat	164
dc-elems	15
dc-terms	55
private ISOcat DatCats (?)	54
Elements with Datcats	82,38%
Components with Datcats	67



Components structure

Data in MDRepository

- Ingesting data from MPI CMDI Harvester



Providers	# Records
LRT-Invetory	892
MPI Language Archives	136.338
OLAC (52 Providers)	34.058
Uni Tübingen	19.138
Uni Leipzig	(3.901)
Uni Saarland	9
CNRTL	228
Meertens Institute	246.728
AAC test corpus	(476)
all	**441.788**

Profile |32|	# Records
CorpusProfile	7
DIDDD	1
DIDDD_sub_location	333
DynaSAND	1
DynaSAND_sub_location	267
GTRP	1
GTRP_sub_location	613
imdi-corpus	11.000
LexicalResourceProfile	1
LiteraryCorpusProfile	19.059
OLAC-DcmiTerms	34.067
Performer	1.530
PhotoSinger	38
ResourceBundle	1
Session	125.336
Song	155.403
SongAudio	7.961
SongScan	28.448
Soundbites	1
Soundbites-recording	1.982
Source	16.519
SourceListProfile	8
SourceProfile	3.886
SourceScan	21.256
SymbolicMusicNotation	7.557
TCOF-profile	228
teiHeader	467
Text	4.417
TextCorpusProfile	10
ToolService	2
ToponymProfile	400
WebLichtWebService	76

Semantic Mapping

- metadata fields in (completely) different profiles
  bound to data categories (ConceptLinks)
- use this linkage when searching in the data
  i.e. allow the user to search

  a) „in the data category“

  b) in a MD field but also all related fields from other profiles

- Multiple mapping levels:

  1. just mapping based on the ConceptLink resolvable via ComponentRegistry
     different elements pointing to the same DatCat
  2. use equivalence relations between DatCats from Relation Registry
  3. use equivalence relations also between Container DatCats
  4. use also other relations in Relation Registry (subClassOf, almostSameAs, …)
  5. apply selected (user defined) relation sets from Relation Registry

Semantic Mapping Component

- separate CMDI module
- relies on information from ComponentRegistry, DCR, RelationRegistry
- is used by Metadata Repository / Service / Browser
- Task:

  resolution: dcrIndex ↔ cmdIndex

  dcrIndex :: (abstract) concept defined in DCR
  cmdIndex :: path to a field in a MDRecord

- (different from

  - query expansion: CQL(datcat) → CQL(cmdIndex[])
  - query translation: e.g. CQL → XPath

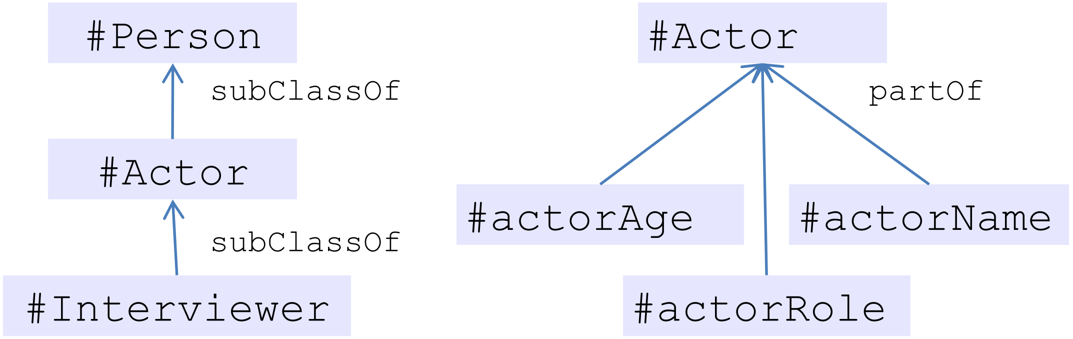
Input			Output	
dcrIndex	isocat.DC-2545 (= isocat.resourceTitle)	=>	cmdIndex[]	[BamdesCommonFields.resourceTitle, imdi-corpus.Corpus.Title, …]
cmdIndex	Actor.Role	=>	dcrIndex	isocat:DC-2559 (participantRole)

Relations

```
#sameAs (#dc:title, #isocat:resourceTitle)
```

```
#almostSameAs (#resourceTitle, #resourceName)
```

- types
  - equivalence
  - lax equivalence, synonymy
  - generalization
  - part of relation
    -> Container DatCats



- currently two relation sets relevant for CMDI:
  - **cmdi** - 14 relations between isocat and dublincore elements
  - **dc** - 15 (trivial) relations between dublincore elements and terms
- use in a query (optional, relation set has to be selectable)

```
dc:title =/relset/cmdi "syntax"
```

or

```
rr-cmdi:title = "syntax"
```

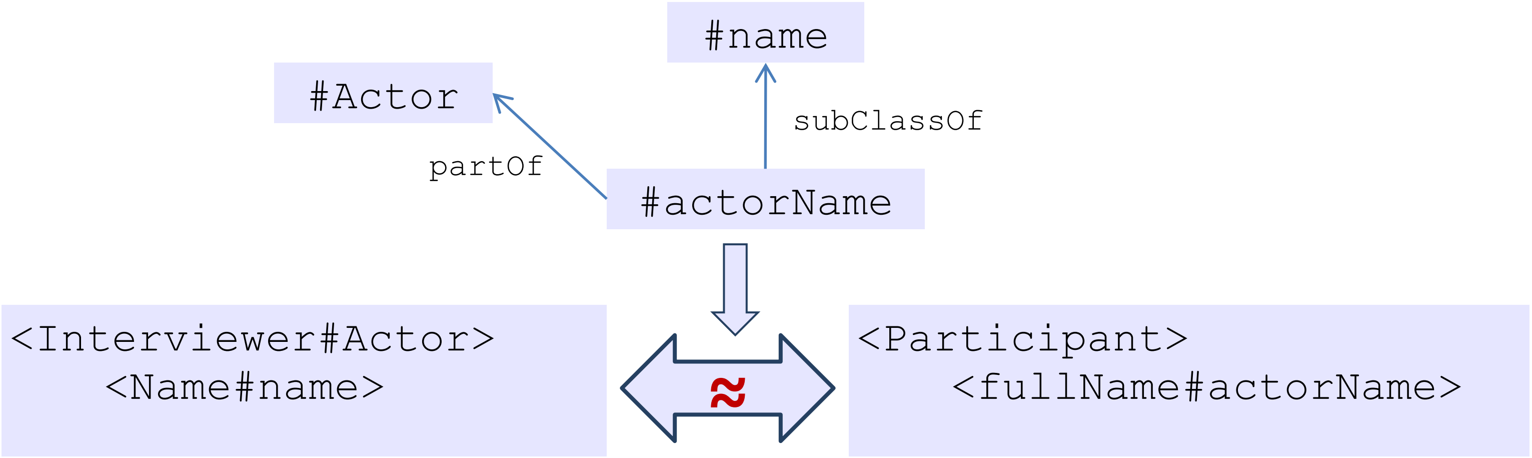
would match:

```
<* #dc:title> or <* #isocat:resourceTitle>
```

(all MD-fields bound to dc:title or isocat:resourceTitle)

Container data categories

- distinguish between (CMD-)components and container data categories
- currently 67 Components with data categories
  ?! but do they use container datcats?
- basic use in search: just as field `isocat:Actor`
- with `partOf`-relations, new mappings would be deducible:



#name

#Actor

subClassOf

partOf

#actorName

<Interviewer#Actor>
    <Name#name>

≈

<Participant>
    <fullName#actorName>

- difficult to express `partOf`-relation in a query:

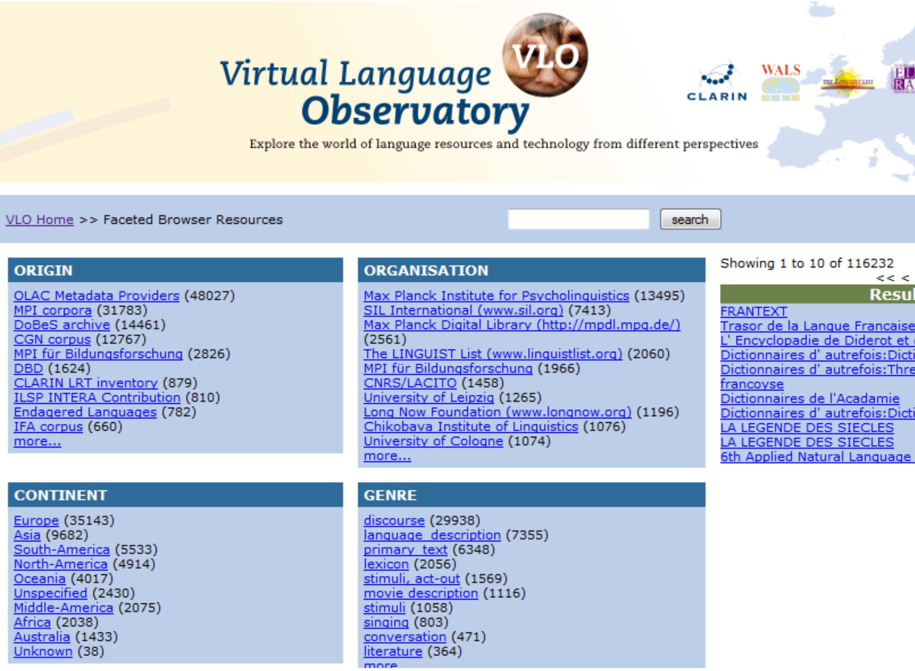
`isocat:Actor.isocat:Name` = `isocat:actorName` = `cmd:Interviewer.Name`

DCR/SMC use in CMDI exploitation tools

- VLO – Virtual Language Observatory
  uses data categories to automatically
  map fields to facets

- MD Browser
  query input widget for complex queries
  uses SMC for autocompletion

- MD Repository
  uses SMC to resolve queries with dcrIndexes
  (in development)



Explore the world of language resources and technology from different perspectives

VLO Home >> Faceted Browser Resources

search

ORIGIN

OLAC Metadata Providers (48027)
MPI corpora (31783)
DoBeS archive (14461)
CGN corpus (12767)
MPI für Bildungsforschung (2826)
DBD (1624)
CLARIN LRT inventory (879)
ILSP INTERA Contribution (810)
Endagered Languages (782)
IFA corpus (660)
more...

ORGANISATION

Max Planck Institute for Psycholinguistics (13495)
SIL International (www.sil.org) (7413)
Max Planck Digital Library (http://mpdl.mpg.de/) (2561)
The LINGUIST List (www.linguistlist.org) (2060)
MPI für Bildungsforschung (1966)
CNRS/LACITO (1458)
University of Leipzig (1265)
Long Now Foundation (www.longnow.org) (1196)
Chikobava Institute of Linguistics (1076)
University of Cologne (1074)
more...

Showing 1 to 10 of 116232

<< <

Resul

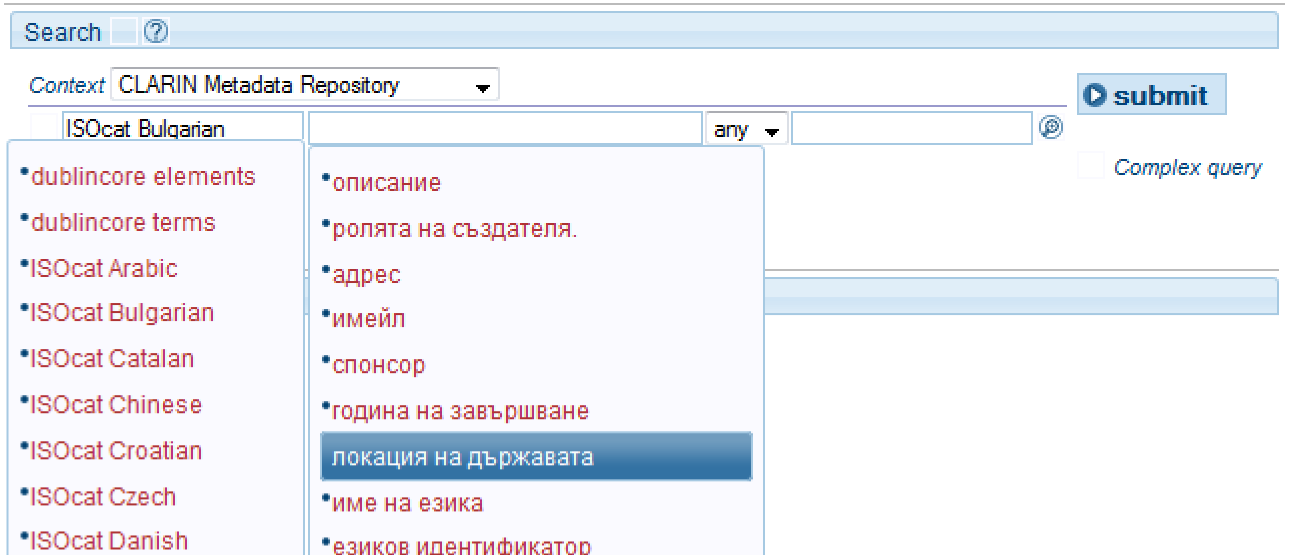
FRANTEXT
Trasor de la Langue Francaise
L' Encyclopadie de Diderot et
Dictionnaires d' autrefois:Dicti
Dictionnaires d' autrefois:Thre
francoyse
Dictionnaires de l'Acadamie
Dictionnaires d' autrefois:Dicti
LA LEGENDE DES SIECLES
LA LEGENDE DES SIECLES
6th Applied Natural Language

CONTINENT

Europe (35143)
Asia (9682)
South-America (5533)
North-America (4914)
Oceania (4017)
Unspecified (2430)
Middle-America (2075)
Africa (2038)
Australia (1433)
Unknown (38)

GENRE

discourse (29938)
language_description (7355)
primary_text (6348)
lexicon (2056)
stimuli, act-out (1569)
movie description (1116)
stimuli (1058)
singing (803)
conversation (471)
literature (364)
more



Context CLARIN Metadata Repository

ISOcat Bulgarian | | any |

□ *Complex query*

- dublincore elements
- dublincore terms
- ISOcat Arabic
- ISOcat Bulgarian
- ISOcat Catalan
- ISOcat Chinese
- ISOcat Croatian
- ISOcat Czech
- ISOcat Danish

- описание
- ропята на създателя.
- адрес
- имейл
- спонсор
- година на завършване
- покация на държавата
- име на езика
- езиков идентификатор

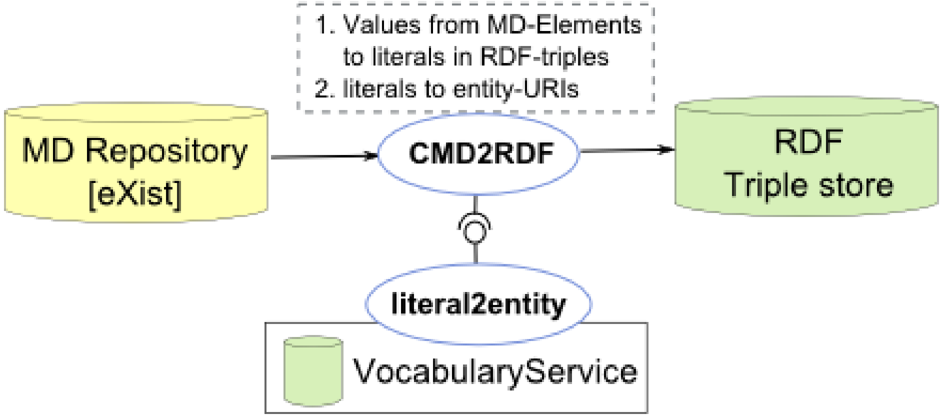
Next Steps

- support and test complex queries in MD Repository (employing SMC)
- finish and publish MDBrowser
- integrate relations and container data categories

- long term: Semantic Mapping on instance level

  = bind also values in MD fields to concepts

  - requires further Vocabularies (=> Vocabulary Service)
  - leads to expressing MD records in RDF/LOD


1. Values from MD-Elements to literals in RDF-triples
2. literals to entity-URIs

Thank you