SMC4LRT - Master Outline

Matej Durco

March 10, 2013

Contents

List of Figures

1 Introduction

Title: Semantic Mapping Component for Language Resources and Technology

1.1 Main Goal

This work proposes a component that shall enhance search functionality over a *large heterogeneous collection of metadata descriptions* of Language Resources and Technology (LRT). By applying semantic web technology the user shall be given both better recall through *query expansion* based on related categories/concepts and new means of *exploring the dataset* via ontology-driven browsing.

Alternatively/ that allows query expansion by providing mappings between search indexes. This enables semantic search, ultimately increasing the recall when searching in metadata collections. The module builds on the Data Category Registry and Component Metadata Framework that are part of CMDI.

Following two examples for better illustration. First a concept-based query expansion: Confronted with a user query: `Actor.Name = Sue` and knowing that `Actor` is synonym to `Person` and `Name` is synonym to `FullName` the expanded query could look like:

```
Actor.Name = Sue OR Actor.FullName = Sue OR
Person.Name = Sue OR Person.FullName = Sue
```

And second, an ontology-driven search: Starting from a list of topics the user can browse an ontology to find institutions concerned with those topics and retrieve a union of resources for the resulting cluster. Thus in general the user is enabled to work with the data based on information that is not present in the original dataset, but rather in external linked-in semantic resources.

Such **semantic search** functionality requires a preprocessing step, that produces the underlying linkage both between categories/concepts and on the instance level. We refer to this task as **semantic mapping**, that shall be realized by corresponding `Semantic Mapping Component`. In this work the focus lies on the method itself – expressed in the specification and operationalized in the (prototypical) implementation of the component – rather than trying to establish a final, accomplished alignment. Although a tentative, naïve mapping on a subset of the data will be proposed, this will be mainly used for evaluation and shall serve as basis for discussion with domain experts aimed at creating the actual sensible mappings usable for real tasks.

In fact, due to the great diversity of resources and research tasks, a "final" complete alignment does not seem achievable at all. Therefore also the focus shall be on "soft" dynamic mapping, i.e. to enable the users to adapt the mapping or apply different mappings depending on their current task or research question essentially being able to actively manipulate the recall/precision ratio of the search results. This entails an examination of user interaction with and visualization of the relevant additional information in the user search interface. However this would open doors to a whole new (to this work) field of usability engineering and can be treated here only marginally.

1.2 Method

We start with examining the existing data and describing the evolving infrastructure in which the components are to be embedded. Then we formulate the function of **Semantic Search** distinguishing between the concept level – using semantic relations between concepts or categories for better retrieval – and the instances level – allowing the user to explore the primary data collection via semantic resources (ontologies, vocabularies).

Subsequently we introduce the underlying **Semantic Mapping Component** again distinguishing the two levels - concepts and instances. We describe the workflow and the central methods, building upon the existing pieces of the infrastructure (See *Infrastructure Components* in 2). A special focus will be put on the examination of the feasibility of employing ontology mapping and alignment techniques and tools for the creation of the mappings.

In the practical part - processing the data - a necessary prerequisite is the dataset being expressed in RDF. Independently, starting from a survey of existing semantic resources (ontologies, vocabularies), we identify an intial set of relevant ones. These will then be used in the exercise of mapping the literal values in the by then RDF-converted metadata descriptions onto externally defined entities, with the goal of interlinking the dataset with external resources (see *Linked Data* in 2).

Finally, in a prototypical implementation of the two components we want to deliver a proof of the concept, supported by an evaluation in which we apply a set of test queries and compare a traditional search with a semantically expanded query in terms of recall/precision indicators. A separate evaluation of the usability of the Semantic Search component is indicated, however this issue can only be tackled marginally and will have to be outsourced into future work.

- a) define/use semantic relations between categories (RelationRegistry)
- b) employ ontological resources to enhance search in the dataset (SemanticSearch)
- c) specify a translation instructions for expressing dataset in rdf (LinkedData)

1.3 Expected Results

The primary concern of this work is the integrative effort, i.e. putting together existing pieces (resources, components and methods) especially the application of techniques from ontology mapping to the domain-specific data collection (the domain of LRT). Thus the main result of this work will be the *specification* of the two components `Semantic Search` and the underlying `Semantic Mapping`. This theoretical part will be accompanied by a proof-of-concept *implementation* of the components and the results and findings of the *evaluation*.

One promising by-product of the work will be the original dataset expressed as RDF with links into existing external resources (ontologies, knowledgebases, vocabularies), effectively laying a foundation for providing this dataset as *Linked Open Data*[1] in the *Web of Data*.

Specification definition of the mapping mechanism

Prototype proof of concept implementation

Evaluation evaluation results of querying the dataset comparing traditional search and semantic search

LinkedData translation of the source dataset to RDF-based format with links into existing datasets/ontologies/knowledgebases

[1] `http://linkeddata.org/`

1.4 Keywords

Metadata interoperability, Ontology Mapping, Schema mapping, Crosswalk, Similarity measures, LinkedData Fuzzy Search, Visual Search?

Language Resources and Technology, LRT/NLP/HLT

Ontology Visualization

2 Previous Work

Infrastructure Components

In recent years, multiple large-scale initiatives have been set out to combat the fragmented nature of the language resources landscape in general and the metadata interoperability problems in particular. A comprehensive architecture for harmonized handling of metadata – the Component Metadata Infrastructure (CMDI)[2] [1] – is being implemented within the CLARIN project[3]. This service-oriented architecture consisting of a number of interacting software modules allows metadata creation and provision based on a flexible meta model, the *Component Metadata Framework*, that facilitates creation of customized metadata schemas – acknowledging that no one metadata schema can cover the large variety of language resources and usage scenarios – however at the same time equipped with well-defined methods to ground their semantic interpretation in a community-wide controlled vocabulary – the data category registry [2, 3].

Individual components of this infrastructure will be described in more detail in the section 5.

LRT Resources

The CLARIN project also delivers a valuable source of information on the normative resources in the domain in its current deliverable on *Interoperability and Standards* [4]. Next to covering ontologies as one type of resources this document offers an exhaustive collection of references to standards, vocabularies and other normative/standardization work in the field of Language Resources and Technology.

Regarding existing domain-specific semantic resources `LT-World`[4], the ontology-based portal covering primarily Language Technology being developed at DFKI[5], is a prominent resource providing information about the entities (Institutions, Persons, Projects, Tools, etc.) in this field of study. [5]

Ontology Mapping

As the main contribution shall be the application of *ontology mapping* techniques and technology, a comprehensive overview of this field and current developments is paramount. There seems to be a plethora of work on the topic and the difficult task will be to sort out the relevant contributions. The starting point for the investigation will be the overview of the field by Kalfoglou [6] and a more recent summary of the key challenges by Shvaiko and Euzenat [7].

In their rather theoretical work Ehrig and Sure [8] elaborate on the various similarity measures which are at the core of the mapping task. On the dedicated platform OAEI[6] an ongoing effort is being carried out and documented comparing various alignment methods applied on different domains.

One more specific recent inspirative work is that of Noah et. al [9] developing a semantic digital library for an academic institution. The scope is limited to document collections, but nevertheless many aspects seem very relevant for this work, like operating on document metadata, ontology population or sophisticated querying and searching.

[2] `http://www.clarin.eu/cmdi`
[3] `http://clarin.eu`
[4] `http://www.lt-world.org/`
[5] *Deutsches Forschungszentrum für Künstliche Intelligenz* - `http://www.dfki.de`
[6] Ontology Alignment Evalution Intiative - `http://oaei.ontologymatching.org/`

Linked Open Data

As described previously one outcome of the work will be the dataset expressed in RDF interlinked with other semantic resources. This is very much in line with the broad *Linked Open Data* effort as proposed by Berners-Lee [10] and being pursuit across many discplines. (This topic is supported also by the EU Commission within the FP7.[7]) A very recent comprehensive overview of the principles of Linked Data and current applications is the book by Heath and Bizer [11], that shall serve as a practical guide for this specific task.

———————-

2.1 Language Resources and Technology

While in the Digital Libraries community a consolidation generally already happened and big federated networks of digital libary repository are set up, in the field of Language Resource and Technology the landscape is still scattered, although meanwhile looking back at a decade of standardizing efforts. One main reason seems to be the complexity and diversity of the metadata associated with the resources, stemming for one from the wide range of resource types additionally complicated by dependence of different schools of thought.

Need some number about the disparity in the field, number of institutes, resources, formats.

This situation has been identified by the community and multiple standardization initiatives had been conducted/undertaken. This process seems to have gained a new momentum thanks to large Research Infrastructure Programmes introduced by European Commission, aimed at fostering Research communities developing large-scale pan-european common infrastructures. One key player in this development is the project CLARIN.

2.1.1 CLARIN

CLARIN - Common Language Resource and Technology Infrastructure - constituted by over 180 members from round 38 countries. The mission of this project is

create a research infrastructure that makes language resources and technologies (LRT) available to scholars of all disciplines, especially SSH large-scale pan-European collaborative effort to create, coordinate and make language resources and technology available and readily useable

This shall be accomplished by setting up a federated network of centers (with federated identity management) but mainly providing resources and services in an agreed upon / coherent / uniform / consistent /standardized manner. The foundation for this goal shall be the Common or Component Metadata infrastructure, a model that caters for flexible metadata profiles, allowing to accomodate existing schemas.

The embedment in the CLARIN project brings about the context of Language Resources and HLT (Human Language Technology, aka NLP - Natural Language Processing) and SSH (Social Sciences and Humanities) as the primary target user-group of CLARIN. CLARIN/NLP for SSH

2.1.2 Standards

ISO12620 Data Category Registry

LAF Linguistic Annotation Framework

[7] `http://cordis.europa.eu/fetch?CALLER=PROJ\_ICT&ACTION=D&CAT=PROJ&RCN=95562`

CMDI - (DC, OLAC, IMDI, TEI)

2.1.3 NLP MD Catalogues

LAT, TLA - Language Archiving Technology, now The Language Archive - provided by Max Planck Insitute for Psycholinguistics `http://www.mpi.nl/research/research-projects/language-archiving-technology`

OTA LR Archiving Service provided by Oxford Text Archive `http://ota.oucs.ox.ac.uk/`

OLAC

ELRA

LDC

DFKI/LT-World

2.2 Ontologies

2.2.1 Word, Sense, Concept

Lexicon vs. Ontology Lexicon is a linguistic object an ontology is not.[12] We don't need to be that strict, but it shall be a guiding principle in this work to consider things (Datasets, Vocabularies, Resources) also along this dichotomy/polarity: Conceptual vs. Lexical. And while every Ontology has to have a lexical representation (canonically: rdfs:label, rdfs:comment, skos:*label), if we don't try to force observed objects into a binary classification, but consider a bias spectrum, we should be able to locate these along this spectrum. So the main focus of a typical ontology are the concepts ("conceptualization"), primarily language-independent.

A special case are Linguistic Ontologies: isocat, GOLD, WALS.info ontologies conceptualizing the linguistic domain

They are special in that ("ontologized") Lexicons refer to them to describe linguistic properties of the Lexical Entries, as opposed to linking to Domain Ontologies to anchor Senses/Meanings. Lexicalized Ontologies: LingInfo, lemon: LMF + isocat/GOLD + Domain Ontology

a) as domain ontologies, describing aspects of the Resources
b) as linguistic ontologies enriching the Lexicalization of Concepts

Ontology and Lexicon [12]

LingInfo/Lemon [13]

We shouldn't need linguistic ontologies (LingInfo, LEmon), they are primarily relevant in the task of ontology population from texts, where the entities can be encountered in various word-forms in the context of the text. (Ontology Learning, Ontology-based Semantic Annotation of Text) And we are dealing with highly structured data with referenced in their nominal(?) form.

Another special case are Controlled Vocabularies or Taxonomies/Classification Systems, let alone folksonomies, in that they identify terms and concepts/meanings, ie there is no explicit mapping between the language represenation and the concept, but rather the term is implicit carrier of the meaning/concept. So for example in the LCSH the surface realization of each subject-heading at the same time identifies the Concept .

controlled vocabularies?

2.2.2 Semantic Web - Linked Data

RDF/OWL

SKOS

2.2.3 OntologyMapping

2.3 Visualization

2.4 FederatedSearch

2.4.1 Standards

Z39.50/SRU/SRW/CQL LoC

OAI-PMH

2.4.2 (Digital) Libraries

General (Libraries, Federations):

OCLC `http://www.oclc.org` world's biggest Library Federation

LoC Library of Congress `http://www.loc.gov`

EU-Lib European Library `http://www.theeuropeanlibrary.org/portal/organisation/handbook/accessing-collections\_en.htm`

europeana virtual European library - cross-domain portal `http://www.europeana.eu/portal/`

2.4.3 Content Repositories

PHAIDRA Permanent Hosting, Archiving and Indexing of Digital Resources and Assets, provided by Vienna University `https://phaidra.univie.ac.at/`

eSciDoc provided by MPG + FIZ Karlsruhe `https://www.escidoc.org/`

DRIVER pan-European infrastructure of Digital Repositories `http://www.driver-repository.eu/`

OpenAIRE - Open Acces Infrastructure for Research in Europe `http://www.openaire.eu/`

2.4.4 (MD)search frameworks:

Zebra/Z39.50 JZKit

Lucene/Solr

eXist - xml DB

2.4.5 Content/Corpus Search

Corpus Search Systems

DDC - text-corpus

manatee - text-corpus

CQP - text-corps

TROVA - MM annotated resources

ELAN - MM annotated resources (editor + search)

2.5 Summary

3 Definitions

We want to clarify or lay down a few terms and definition, ie explanation of our understanding

Concept sense, idea, philosophical problem, which we don't need to discuss here. For our purposes we say: Basic "entity" in an ontology? that of what an ontology is build

Ontology "an explicit specification of a conceptualization" [cite!], but for us mainly a collection of concepts as opposed to lexicon, which is a collection of words.

Word a lexical unit, a word in a language, something that has a surface Realization (writtenForm) and is a carrier of sense. so a Relation holds: hasSense(Word, Concept)

Lexicon a collection of words, a (lexical) vocabulary

Vocabulary an index providing mapping from Word (string) to Concept (uri)

(Data)Category (almost) the same as Concept; Things like "Topic", "Genre", "Organization", "ResourceType" are instantiations of Category

ConceptualDomain the Class of entities a Concept/Category denotes. For Organization it would be all (existing) organizations, CD(ResourceType)=Corpus, Lexicon, Document, Image, Video, ... Entities of the domain can itself be Categories (ResourceType:Image), but it can be also individuals (Organization University of Vienna)

Entity

Resource informational resource, in the context of CLARIN-Project mainly Language Resources (Corpus, Lexicon, Multimedia)

Metadata Description description of some properties of a resource. MD-Record

Schema - CMD-Profile

Annotation

4 Overview of the data landscape

This section gives an overview of existing metadata formats and a description of their characteristics and their usage

4.1 Metadata Schemas

4.1.1 CMD-Framework

created	2013-01-26
Profiles	87
Components	2904
distinct Components	542
Elements	5754
distinct Elements	1505
distinct DatCats	436
Elements with DatCats	1183
Elements without DatCats	323
ratio of elements without DatCats	21.46 %
available Concepts	893
used Concept	474
blind Concepts (not in public ISOcat)	190
Concepts not used in CMD	539

4.1.2 Dublin Core + OLAC

DC, OLAC

4.1.3 TEI / teiHeader

TEI/teiHeader/ODD,

4.1.4 ISLE/IMDI

4.1.5 MODS/METS

4.1.6 Europeana Data Model - EDM

4.1.7 Other

OAI-ORE - is this a schema?

4.2 Metadata collections

META-NET

4.2.1 CMDI

collections, profiles/Terms, ResourceTypes!

4.3 Content/Annotation Schemas

CHILDES, TEI, EAF! (CES/XCES) Open Annotation Collaboration (OAC)[8]

[8] http://openannotation.org/

4.4 Ontologies, Controlled Vocabularies, Knowledge Organizing Systems

4.4.1 Classification Schemes, Taxonomies

LCSH, DDC

4.4.2 Other controlled Vocabularies

Tagsets: STTS Language codes ISO-639-1

4.4.3 Domain Ontologies, Vocabularies

Organization-Lists LT-World !?

5 Underlying infrastructure

As stated before, the proposed module is part of CMDI and depends on multiple modules of the infrastructure. Before we describe the interaction itself in chapter 7.2, we introduce in short these modules and the data they provide:

- Data Category Registry,
- Relation Registry
- Component Registry
- Vocabulary Alignement Service (OpenSKOS)
- SchemaParser

?MDBrowser ?MDService

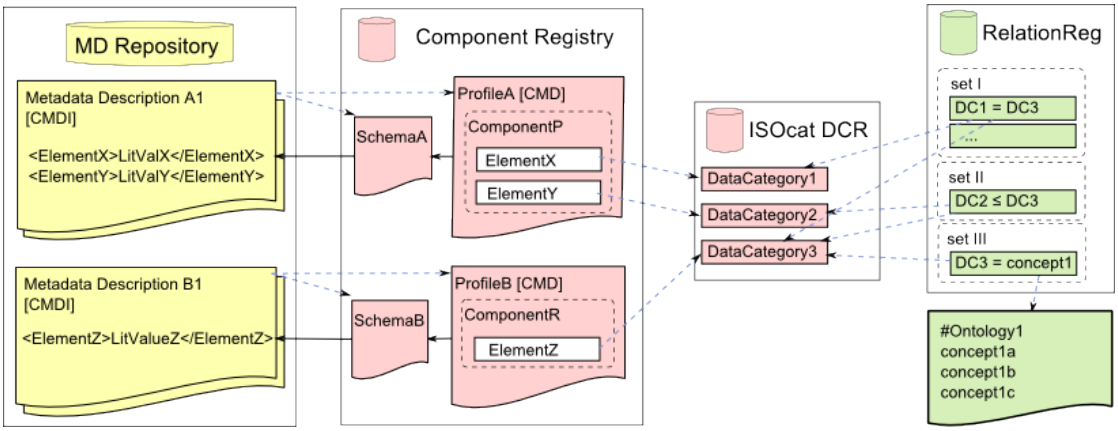

Figure 1: The diagram depicts the links between pieces of data in the individual registries that serve as basis for semantic mapping

5.1 CMDI - Production side

The *Data Category Registry* (DCR) is a central registry that enables the community to collectively define and maintain a set of relevant linguistic data categories. The resulting commonly agreed controlled vocabulary is the cornerstone for grounding the semantic interpretation within the CMD framework. The data model and the procedures of the DCR are defined by the ISO standard [14], and is implemented in *ISOcat*[9].

The *Component Metadata Framework* (CMD) is built on top of the DCR and complements it. While the DCR defines the atomic concepts, within CMD the metadata schemas can be constructed out of reusable components - collections of metadata fields. The components can contain other components, and they can be reused in multiple profiles as long as each field "refers via a PID to exactly one data category in the ISO DCR, thus indicating unambiguously how the content of the field in a metadata description should be interpreted" [3]. This allows to trivially infer equivalencies between metadata fields in different CMD-based schemas. While the primary registry used in CMD is the ISOcat DCR, other authoritative sources for data categories ("trusted registries") are accepted, especially Dublin Core Metadata Initiative [15].

[9] `http://www.isocat.org/`

The framework as described so far provides a sound mechanism for binding the semantic interpretation of the metadata descriptions. However there needs to be an additional means to capture information about relations between data categories. This information was deliberately not included in the DCR, because relations often depend on the context in which they are used, making global agreement unfeasible. CMDI proposes a separate module – the *Relation Registry* (RR) [16] –, where arbitrary relations between data categories can be stored and maintained. We expect that the RR should be under control of the metadata user whereas the DCR is under control of the metadata modeler.

There is a prototypical implementation of such a relation registry called *RELcat* being developed at MPI, Nijmegen. [17, 18], that already hosts a few relation sets. There is no user interface to it yet, but it is accessible as a REST-webservice[10]. This implementation stores the individual relations as RDF-triples

$$< subjectDatacat, relationPredicate, objectDatcat >$$

allowing typed relations, like equivalency (`rel:sameAs`) and subsumption (`rel:subClassOf`). The relations are grouped into relation sets that can be used independently.

!check DCR-RR/Odijk2010 -follow up !Cf. Erhard Hinrichs 2009

And a last relevant intiative to mention is that of a `Vocabulary Alignment Service` being developed and run within the Dutch program CATCH[11], which serves as a neutral manager and provider of controlled vocabularies. There are plans to reuse or enhance this service for the needs of the CLARIN project.
All these components are running services, that this work shall directly build upon.

This approach of integrating prerequisites for semantic interoperability directly into the process of metadata creation differs from the traditional methods of schema matching that try to establish pairwise alignments between schemas only after they were created and published.

Consequently, the infrastructure also foresees a dedicated module, *Semantic Mapping*, that exploits this novel mechanism to deliver correspondences between different metadata schemas. The details of its functioning and its interaction with the aforementioned modules is described in the following chapter 7.2.

5.2 CMDI - Exploitation side

Metadata complying to the CMD-framework is being created by a growing number of institutions by various means, automatic transformation from legacy data, authoring of new metadata records with the help of one of the Metadata-Editors (TODO: cite: Arbil, NALIDA,). The CMD-Infrastructure requires the content providers to publish their metadata via the OAI-PMH protocol and announce the OAI-PMH endpoints. These are being harvested daily by a dedicated CLARIN harvester[12]. The harvested data is validated against the schemas What about Normalization?. and made available in packaged datasets. These are being fetched by the exploitations side components, that index the metadata records and make them available for searching and browsing.

The first stable and publicly available application providing access to the collected metadata of CMDI has been the `VLO - Virtual Language Observatory`[13][19], being developed within the CLARIN project. This application operates on the same collection of data as is discussed in this work, however it employs a faceted search, mapping manually the appropriate metadata fields from the different schemas to 10? fixed facets.

[10] sample relation set: `http://lux13.mpi.nl/relcat/rest/set/cmdi`
[11] *Continuous Access To Cultural Heritage* - `http://www.catchplus.nl/en/`
[12] `http://catalog.clarin.eu/oai-harvester/`
[13] `http://www.clarin.eu/vlo/`

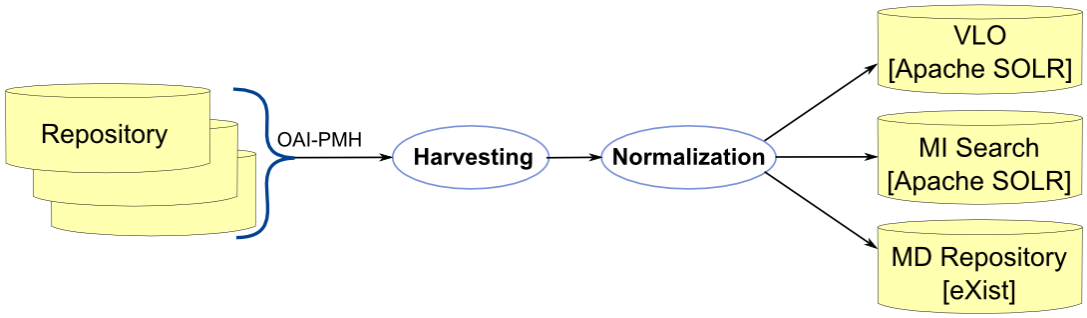

Figure 2: Within CMDI, metadata is harvested from content providers via OAI-PMH and made available to consumers/users by exploitation side components

Underlying search engine is the widely used full-text search engine Apache Solr[14]. Although this is a very reductionist approach it is certainly a great starting point offering a core set of categories together with an initial set of category mappings.

More recently, the team at Meertens Institute developed a similar application the `MI Search Engine`[15]. It too is based on the Apache Solr and provides a faceted search, but with a substantially more sophisticated both indexing process and search interface. describe indexing and search add citation

And finally, there is the *Metadata Repository* aimed to collect all the harvested metadata descriptions from CLARIN centers, and *Metadata Service* that provides search access to this body of data. As such, Metadata Service is the primary application to use Semantic Mapping, to optionally expand user queries before issuing a search in the Metadata Repository. [20]

[14] `http://lucene.apache.org/solr/`
[15] `http://www.meertens.knaw.nl/cmdi/search/`

Semantic Mapping Component

6 ?? DataModel

Terms ? move to SKOS ?

RDF

7 Semantic Mapping on concept level

merging the pieces of information provided by those, offering them semi-transaprently to the user (or application) on the consumption side.

a module of the Component Metadata Infrastructure performing semantic mapping on search indexes. This builds the base for query expansion to facilitate semantic search and enhance recall when querying the Metadata Repository.

7.1 smcIndex

In this section we describe *smcIndex* – the data type for input and output of the proposed application. An smcIndex is a human-readable string adhering to a specific syntax, denoting some search index. The generic syntax is:

$$smcIndex ::= context\ contextSep\ conceptLabel$$

We distinguish two types of smcIndexes: (i) *dcrIndex* referring to data categories and (ii) *cmdIndex* denoting a specific "CMD-entity", i.e. a metadata field, component or whole profile defined within CMD. The *cmdIndex* can be interpreted as a XPath into the instances of CMD-profiles. In contrast to it, the *dcrIndexes* are generally not directly applicable on existing data, but can be understood as abstract indexes referring to well-defined concepts – the data categories – and for actual search they need to be resolved to the metadata fields they are referred by. In return one can expect to match more metadata fields from multiple profiles, all referring to the same data category.

These two types of smcIndex also follow different construction patterns:

$$\begin{aligned}
smcIndex &::= dcrIndex \mid cmdIndex \\
dcrIndex &::= dcrID\ contextSep\ datcatLabel \\
cmdIndex &::= profile \\
&\quad \mid [\ profile\ contextSep\ ]\ dotPath \\
dotPath &::= [\ dotPath\ pathSep\ ]\ elemName \\
contextSep &::= \text{`.`} \mid \text{`:`} \\
pathSep &::= \text{`.`} \\
dcrId &::= \texttt{`isocat`} \mid \texttt{`dc`}
\end{aligned}$$

The grammar is based on the way indices are referenced in CQL-syntax[16] (`dc.title`) and on the dot-notation used in IMDI-browser[17] (`Session.Location.Country`).

dcrID is a shortcut referring to a data category registry similar to the namespace-mechanism in XML-documents. *datcatLabel* is the verbose Identifier- (e.g. `telephoneNumber`) or the Name-attribute (in any available translation, e.g. `numero di telefono@it`) of the data category. *profile* is the name of the profile. *dotPath* allows to address a leaf

[16] Context Query Language, `http://www.loc.gov/standards/sru/specs/cql.html`

[17] `http://www.lat-mpi.eu/tools/imdi`

element (`Session.Actor.Role`), or any intermediary XML-element corresponding to a CMD-component (`Session.Actor`) within a metadata description.

Generally, smcIndexes can be ambiguous, meaning they can refer to multiple concepts, or entities (CMD-elements). This is due to the fact that the names of the data categories, and CMD-entities are not guaranteed unique. The module will have to cope with this, by providing on demand the list of identifiers corresponding to a given smcIndex.

7.2 Function

In this section, we describe the actual task of the proposed application – **mapping indexes to indexes** – in abstract terms. The returned mappings can be used by other applications to expand or translate the original user query, to match elements in other schemas. [18]

Initialization

First there is an initialization phase, in which the application fetches the information from the source modules (cf. 5). All profiles and components from the Component Registry are read and all the URIs to data categories are extracted to construct an inverted map of data categories:

$$datcatURI \mapsto profile.component.element[]$$

The collected data categories are enriched with information from corresponding registries (DCRs), adding the verbose identifier, the description and available translations into other working languages.

Finally relation sets defined in the Relation Registry are fetched and matched with the data categories in the map to create sets of semantically equivalent (or otherwise related) data categories.

Operation

In the operation mode, the application accepts any index (*smcIndex*, cf. 7.1) and returns a list of corresponding indexes (or only the input index, if no correspondences were found):

$$smcIndex \mapsto smcIndex[\ ]$$

We can distinguish following levels for this mapping function:

(1) *data category identity* – for the resolution only the basic data category map derived from Component Registry is employed. Accordingly, only indexes denoting CMD-elements (*cmdIndexes)* bound to a given data category are returned:

```
isocat.size ↦
[teiHeader.extent,
 TextCorpusProfile.Number]
```

[18] Though tightly related, mapping of terms and query expansion are to be seen as two separate functions.

cmdIndex as input is also possible. It is translated to a corresponding data category, proceeding as above:

```
imdi-corpus.Name ↦
(isocat.resourceName) ↦
TextCorpusProfile.GeneralInfo.Name
```

(2) *relations between data categories* – employing also information from the Relation Registry, related (equivalent) data categories are retrieved and subsequently both the input and the related data categories resolved to cmdIndexes:

```
isocat.resourceTitle ↦     (+ dc.title) ↦
[imdi-corpus.Title,
 TextCorpusProfile.GeneralInfo.Title,
 teiHeader.titleStmt.title,
 teiHeader.monogr.title]
```

(3) *container data categories* – further expansions will be possible once the container data categories [18] will be used. Currently only fields (leaf nodes) in metadata descriptions are linked to data categories. However, at times, there is a need to conceptually bind also the components, meaning that besides the "atomic" data category for `actorName, there would be also a data category for the complex concept Actor.` Having concept links also on components will require a compositional approach to the task of semantic mapping, resulting in:

```
Actor.Name ↦
    [Actor.Name, Actor.FullName,
     Person.Name, Person.FullName]
```

Extensions

A useful supplementary function of the module would be to provide a list of existing indexes. That would allow the search user-interface to equip the query-input with autocompletion. Also the application should deliver additional information about the indexes like description and a link to the definition of the underlying entity in the source registry.

Once there will be overlapping[19] user-defined relation sets in the Relation Registry an additional input parameter will be required to *explicitly restrict the selection of relation sets* to apply in the mapping function.

Also, use of *other than equivalency relations will necessitate more complex logic in the query expansion and accordingly also more complex response of the SMC, either returning the relation types themselves as well or equip the list of indexes with some similarity ratio.*

8 Semantic Mapping on instance level

8.1 Mapping from strings to Entities

Based on the textual values in the Metadata-descriptions find matching entities in selected Ontologies.

Identify related ontologies: LT-World [5]

task:

[19] i.e. different relations may be defined for one data category in different relation sets

1. express MDRecords in RDF
2. identify related ontologies/vocabularies (category -¿ vocabulary)
3. use a lookup/mapping function (Vocabulary Alignement Service? CATCH-PLUS?)

$$lookup(Category, Literal) -> ConceptualDomain??$$

Normally this would be served by dedicated controlled vocabularies, but expect also some string-normalizing preprocessing etc.

8.2 Semantic Search

Main purpose for the undertaking described in previous two chapters (mapping of concepts and entities) is to enhance the search capabilities of the MDService serving the Metadata/Resources-data. Namely to enhance it by employing ontological resources. Mainly this enhancement shall mean, that the user can access the data indirectly by browsing one or multiple ontologies, with which the data will then be linked. These could be for example ontologies of Organizations and Projects.

In this section we want to explore, how this shall be accomplished, ie how to bring the enhanced capabilities to the user. Crucial aspect is the question how to deal with the even greater amount of information in a user-friendly way, ie how to prevent overwhelming, intimidating or frustrating the user.

Semi-transparently means, that primarily the semantic mapping shall integrate seamlessly in the interaction with the service, but it shall "explain" - offer enough information - on demand, for the user to understand its role and also being able manipulate easily.

? Facets Controlled Vocabularies Synonym Expansion (via TermExtraction(ContentSet))

9 Linked Data - Express dataset in RDF

Partly as by-product of the entities-mapping effort we will get the metadata-description rendered in RDF, linked with So theoretically we then only need to provide them "on the web", to make them a nucleus of the LinkedData-Cloud.

Practically this won't be that straight-forward as the mapping to entities will be a hell of a work. But once that is solved, or for the subsets that it is solved, the publication of that data on the "SemanticWeb" should be easy.

Technical aspects (RDF-store?) / interface (ontology browser?)

defining the Mapping:

1. convert to RDF translate: MDREcord -¿ [#mdrecord #property literal]
2. map: #mdrecord #property literal -¿ [#mdrecord #property #entity]

9.1 Content/Annotation

AF + DCR + RR

9.2 Visualization

Landscape, Treemap, SOM

Ontology Mapping and Alignement / saiks/Ontology4 4auf1.pdf

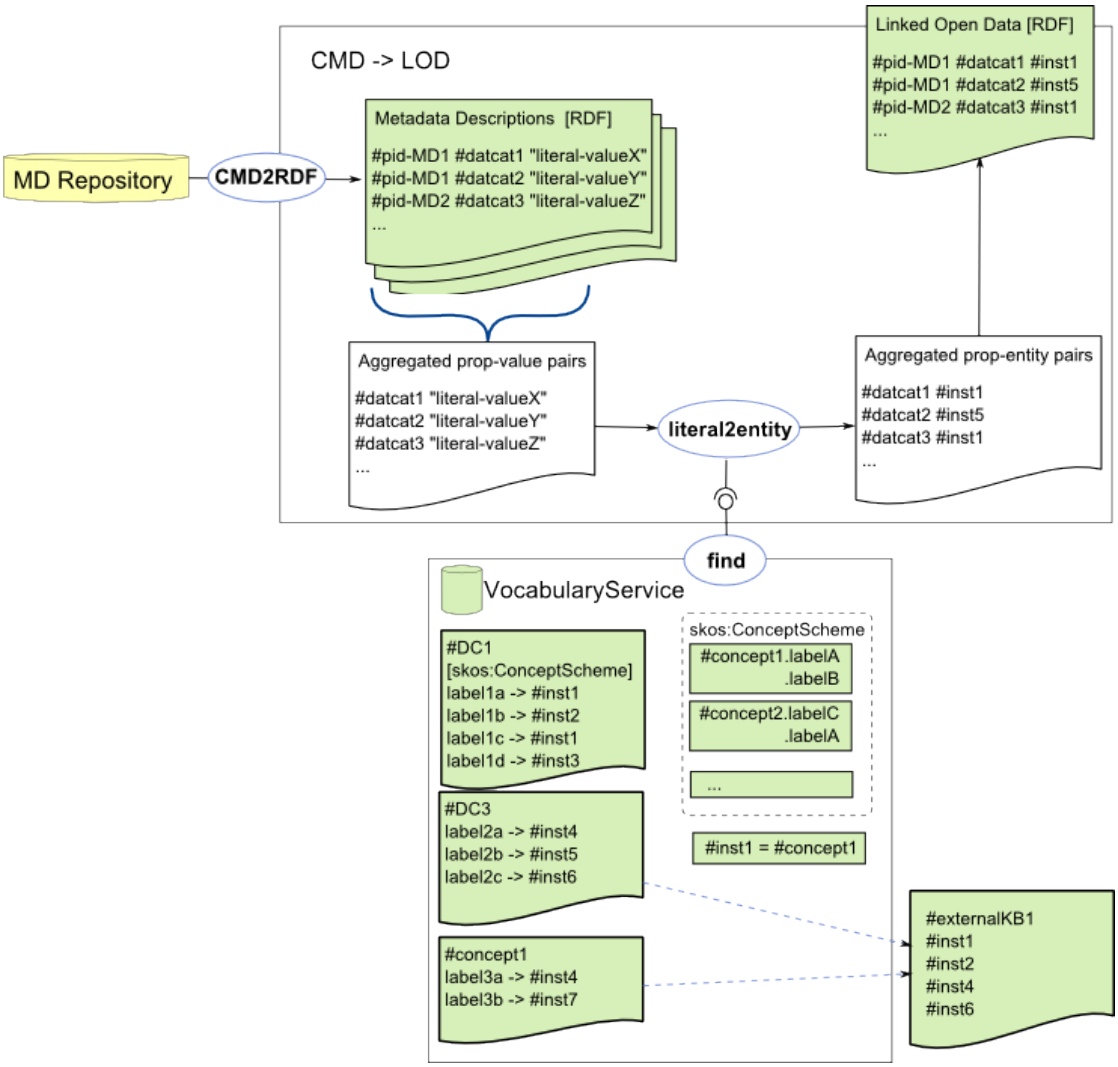

Figure 3: The process of transforming the CMD metadata records to and RDF representation

9.3 Implementation

The core function of the SMC is being implemented as a set of XSL-stylesheets, with auxiliary functionality (like caching or a wrapping web service) being provided by a wrapping application implemented in Java. There is also a plan to provide an XQuery implementation. The SMC module is being maintained in the CMDI code repository[20].

9.3.1 smc init

9.3.2 smc browser

Explore the Component Metadata Framework

In CMD, metadata schemas are defined by profiles, that are constructed out of reusable components - collections of metadata fields. The components can contain other components, and they can be reused in multiple profiles. Furthermore, every CMD element (metadata field) refers via a PID to a data category to indicate unambiguously how the content of the field in a metadata description should be interpreted (Broeder et al., 2010).

Thus, every profile can be expressed as a tree, with the profile component as the root node, the used components as intermediate nodes and elements or data categories as leaf nodes, parent-child relationship being defined by the inclusion (componentA -includes-¿ componentB) or referencing (elementA -refersTo-¿ datcat1).The reuse of components in multiple profiles and especially also the referencing of the same data categories in multiple CMD elements leads to a blending of the individual profile trees into a graph (acyclic directed, but not necessarily connected).

SMC Browser visualizes this graph structure in an interactive fashion. You can have a look at the examples for inspiration.

It is implemented on top of wonderful js-library d3, the code checked in clarin-svn (and needs refactoring). More technical documentation follows soon.

The graph is constructed from all profiles defined in the Component Registry. To resolve name and description of data categories referenced in the CMD elements definitions of all (public) data categories from DublinCore and ISOcat (from the Metadata Profile [RDF] - retrieving takes some time!) are fetched. However only data categories used in CMD will get part of the graph. Here is a quantitative summary of the dataset.

9.3.3 smc as mdrepo module

9.3.4 smc as VAS

9.4 User Interface

9.4.1 Query Input

9.4.2 Columns

9.4.3 Summaries

9.4.4 Differential Views

Visualize impact of given mapping in terms of covered dataset (number of matched records).

[20] `http://svn.clarin.eu/SMC`

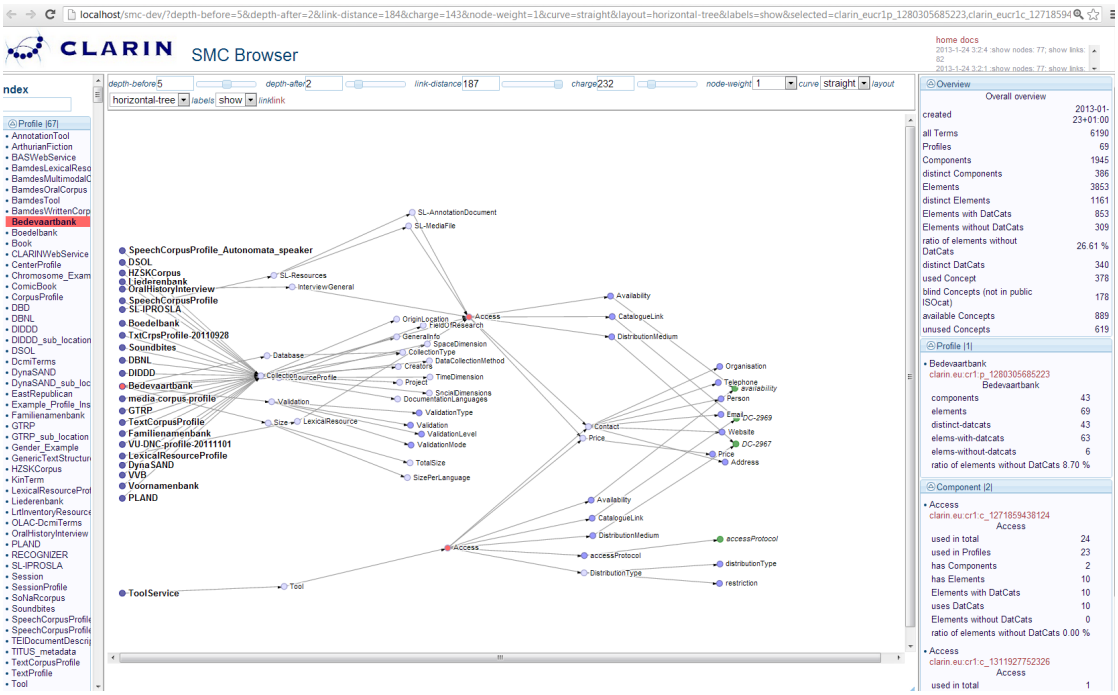

Figure 4: Screenshot of the SMC browser

10 Evaluation

10.1 Use Cases

- MD Search employing Semantic Mapping
- MD Search employing Fuzzy Search
- Visualization of the Results - ?

A trivial example for a concept-based query expansion: Confronted with a user query: `Actor.Name = Sue` and knowing that `Actor` is equivalent or similar to `Person` and `Name` is synonym to `FullName` the expanded query could look like: `Actor.Name = Sue OR Actor.FullName = Sue OR Person.Name = Sue OR Person.FullName= is Sue`

Another example concerning instance mapping: the user looking for all resource produced by or linked to a given institution, does not have to guess or care for various spellings of the name of the institution used in the description of the resources, but rather can browse through a controlled vocabulary of institutions and see all the resources of given institution. While this could be achieved by simple normalizing of the literal-values (and indeed that definitely has to be one processing step), the linking to an ontology enables to user to also continue browsing the ontology to find institutions that are related to the original institution by means of being concerned with similar topics and retrieve a union of resources for such resulting cluster. Thus in general the user is enabled to work with the data based on information that is not present in the original dataset.

10.2 Research Questions

10.3 Sample Queries

candidate Categories: ResourceType, Format Genre, Topic Project, Institution, Person, Publisher

10.4 Usability

11 Conclusions and Future Work

The Semantic Mapping module is based on the DCR and CMD framework and is being developed as a separate service on the side of CLARIN Metadata Service, its primary consuming service, but shall be equally usable by other applications.

Further work is needed on more complex types of response (similarity ratio, relation types) and also on the interaction with Metadata Service to find the optimal way of providing the features of semantic mapping and query expansion as semantic search within the search user-interface.

12 Questions, Remarks

- How does this relate to federated search?
- ontologicky vs. semaziologicky (Semanticke priznaky: kategoriálne/archysémy, difernciacne, specifikacne)
- "controlled vocabularies"

References

[1] D. Broeder, O. Schonefeld, T. Trippel, D. Van Uytvanck, and A. Witt, "A pragmatic approach to XML interoperability - the Component Metadata Infrastructure (CMDI)," in *Balisage: The Markup Conference 2011*, vol. 7, 2011. citeulike:9861691.

[2] M. Kemps-Snijders, M. Windhouwer, and P. Wittenburg, "Isocat: Remodeling metadata for language resources," in *International Journal of Metadata, Semantics and Ontologies (IJMSO)*, vol. 4 (4), pp. 261–276, 2009.

[3] D. Broeder, M. Kemps-Snijders, D. V. Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, and C. Zinn, "A data category registry- and component-based metadata framework," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)* (N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, and D. Tapias, eds.), (Valletta, Malta), European Language Resources Association (ELRA), May 2010.

[4] E. Hinrichs, P. Banski, K. Beck, G. Budin, T. Caselli, K. Eckart, K. Elenius, G. Faaß, M. Gavrilidou, V. Henrich, V. Quochi, L. Lemnitzer, W. Maier, M. Monachini, J. Odijk, M. Ogrodniczuk, P. Osenova, P. Pajas, M. Piasecki, A. Przepiórkowski, D. V. Uytvanck, T. Schmidt, I. Schuurman, K. Simov, C. Soria, I. Skadina, J. Stepanek, P. Stranak, P. Trilsbeek, T. Trippel, and I. Vogel, "Interoperability and standards," deliverable, CLARIN, March 2011.

[5] B. Jörg, H. Uszkoreit, and A. Burt, "Lt world: Ontology and reference information portal," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)* (N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, and D. Tapias, eds.), (Valletta, Malta), European Language Resources Association (ELRA), May 2010.

[6] Y. Kalfoglou and M. Schorlemmer, "Ontology mapping: the state of the art," *The Knowledge Engineering Review*, vol. 18, pp. 1–31, Jan. 2003.

[7] P. Shvaiko and J. Euzenat, "Ten challenges for ontology matching," in *On the Move to Meaningful Internet Systems: OTM 2008* (R. Meersman and Z. Tari, eds.), vol. 5332 of *Lecture Notes in Computer Science*, pp. 1164–1182, Springer Berlin / Heidelberg, 2008. 10.1007/978-3-540-88873-4_18.

[8] M. Ehrig and Y. Sure, "Ontology mapping – an integrated approach," in *The Semantic Web: Research and Applications* (C. Bussler, J. Davies, D. Fensel, and R. Studer, eds.), vol. 3053 of *Lecture Notes in Computer Science*, pp. 76–91, Springer Berlin / Heidelberg, 2004. 10.1007/978-3-540-25956-5_6.

[9] S. Noah, N. Alias, N. Osman, Z. Abdullah, N. Omar, Y. Yahya, and M. Yusof, "Ontology-driven semantic digital library," in *Information Retrieval Technology* (P.-J. Cheng, M.-Y. Kan, W. Lam, and P. Nakov, eds.), vol. 6458 of *Lecture Notes in Computer Science*, pp. 141–150, Springer Berlin / Heidelberg, 2010. 10.1007/978-3-642-17187-1_13.

[10] T. Berners-Lee, "Linked data." online: http://www.w3.org/DesignIssues/LinkedData.html, 07 2006. Status: personal view only. Editing status: imperfect but published. Last visited: 2011-04-13.

[11] T. Heath and C. Bizer, "Linked data: Evolving the web into a global data space," *Synthesis Lectures on the Semantic Web: Theory and Technology*, vol. 1, pp. 1–136, Feb 2011.

[12] G. Hirst, "Ontology and the lexicon," in *Handbook on Ontologies* (P. Bernus, J. Błażewics, G. Schmidt, M. Shaw, S. Staab, and R. Studer, eds.), International Handbooks on Information Systems, pp. 269–292, Springer Berlin Heidelberg, 2009. 10.1007/978-3-540-92673-3_12.

[13] P. Buitelaar, P. Cimiano, P. Haase, and M. Sintek, "Towards linguistically grounded ontologies," in *The Semantic Web: Research and Applications* (L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvönen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, eds.), vol. 5554 of *Lecture Notes in Computer Science*, pp. 111–125, Springer Berlin / Heidelberg, 2009. 10.1007/978-3-642-02121-3_12.

[14] ISO12620:2009, "Computer applications in terminology – data categories – specification of data categories and management of a data category registry for language resources," 2009.

[15] A. Powell, M. Nilsson, A. Naeve, and P. Johnston, "DCMI Abstract Model," tech. rep., Mar. 2005.

[16] M. Kemps-Snijders, M. Windhouwer, and S. E. Wright, "Putting data categories in their semantic context," in *Proceedings of the IEEE e-Humanities Workshop (e-Humanities)*, (Indianapolis, Indiana, USA), December 2008.

[17] M. Windhouwer, "Relcat and friends," in *Presentation at CLARIN-NL ISOcat workshop*, (Nijmegen), MPI for Psycholinguistics, 05 2011.

[18] I. Schuurman and M. Windhouwer., "Explicit semantics for enriched documents. what do isocat, relcat and schemacat have to offer?," in *2nd Supporting Digital Humanities conference (SDH 2011), 17-18 November 2011, Copenhagen, Denmark*, (Copenhagen, Denmark), 2011.

[19] D. V. Uytvanck, C. Zinn, D. Broeder, P. Wittenburg, and M. Gardellini, "Virtual language observatory: The portal to the language resources and technology universe," in *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)* (N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, M. Rosner, and D. Tapias, eds.), (Valletta, Malta), European Language Resources Association (ELRA), May 2010.

[20] M. Ďurčo and L.-J. Olsson, "CMDRSB - CLARIN Metadata Repository/Service/Browser," in *Presentation at CMDI Workshop, Nijmegen*, (Nijmegen), MPI for Psycholinguistics, 01 2011.