SMC4LRT - Master Outline

Matej Durco

March 15, 2013

Contents

List of Figures

Todo list

Chapter 1

Introduction

- motivation
- problem statement (which problem should be solved?)
- aim of the work
- methodological approach
- structure of the work

install older python (2.5?) to be able to install dot2tex - transforming dot files to nicer pgf formatted graphs

12

1.0.1 Problem statement

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.

1.0.2 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.

[1] http://dot2tex.googlecode.com/files/dot2tex-2.8.7.zip
[2] file:/C:/Users/m/2kb/tex/dot2tex-2.8.7/

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.

1.0.3 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 **??**). 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 **??**).

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)

1.0.4 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*[3] 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

1.0.5 Keywords

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

Language Resources and Technology, LRT/NLP/HLT

Ontology Visualization

[3] http://linkeddata.org/

Chapter 2

State of the Art

This work is guided by

two (or three? + Infrastructure

main dimensions: the data - in broad, Language Resource and Technology and the method - Semantic Web technologies. This division is reflected in the following chapter:

2.1 (Infrastructure for) Language Resources and Technology

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.

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* [1]. 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`[1], the ontology-based portal covering primarily Language Technology being developed at DFKI[2], is a prominent resource providing information about the entities (Institutions, Persons, Projects, Tools, etc.) in this field of study. [2] Chapter 4 examines the field of LRT in more detail.

2.1.1 Metadata

A comprehensive architecture for harmonized handling of metadata – the Component Metadata Infrastructure (CMDI)[3] [3] – is being implemented within the CLARIN project[4]. 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 [**?**, 4].

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

[1] `http://www.lt-world.org/`
[2] *Deutsches Forschungszentrum für Künstliche Intelligenz* - `http://www.dfki.de`
[3] `http://www.clarin.eu/cmdi`
[4] `http://clarin.eu`

2.1.2 Content Repositories

Metadata is only one aspect of the availability of resources. It is the first step to announce and describe the resources. However it is of little value, if the resources themselves are not equally well accessible. Thus another pillar of the CLARIN infrastructure are Content Repositories - centres to ensure availability of resources. In the following a few well established repositories are mentioned and described, as well as some of the new repositories being set up in the context of CLARIN.

PHAIDRA Permanent Hosting, Archiving and Indexing of Digital Resources and Assets, provided by Vienna University [5]

eSciDoc provided by MPG + FIZ Karlsruhe [6]

TextGrid install: TextGrid2 - check: TG-search
[7]

DRIVER pan-European infrastructure of Digital Repositories [8]

OpenAIRE - Open Acces Infrastructure for Research in Europe [9]

2.1.3 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)

2.1.4 FederatedSearch

How to relate Federated Search to SMC?

2.2 Semantic Web

cite TimBL

RDF/OWL

SKOS

2.2.1 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 [5] and being pursuit across many discplines. (This topic is supported also by the EU Commission within the FP7.[10])

[5] https://phaidra.univie.ac.at/
[6] https://www.escidoc.org/
[7] http:/textgrid.de
[8] http://www.driver-repository.eu/
[9] http://www.openaire.eu/
[10] http://cordis.europa.eu/fetch?CALLER=PROJ_ICT&ACTION=D&CAT=PROJ&RCN=95562

A very recent comprehensive overview of the principles of Linked Data and current applications is the book by Heath and Bizer [6], that shall serve as a practical guide for this specific task.

2.2.2 Schema / 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 [7] and a more recent summary of the key challenges by Shvaiko and Euzenat [8].

In their rather theoretical work Ehrig and Sure [9] elaborate on the various similarity measures which are at the core of the mapping task. On the dedicated platform OAEI[11] 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 [10] 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.

check if relevant: http://schema.org/

2.2.3 Ontology Visualization

2.2.4 Linguistic Ontologies

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 [11]

LingInfo/Lemon [12]

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.

2.3 Summary

This chapter concentrated on the current affairs/developments regarding the infrastructures for Language Resources and Technology and on the other hand gave an overview of the state of the art regarding methods to be applied in this work: Semantic Web Technologies, Ontology Mapping and Ontology Visualization.

[11] Ontology Alignment Evalution Intiative - http://oaei.ontologymatching.org/

Chapter 3

Definitions

3.1 Namespaces

Namespaces mentioned through this document listed:

dcif

skos

3.2 Abbreviations

3.3 Terms

In the following, the terms used in this work are explained.

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)

Is it synonymous to value domain, range

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

Lexicon vs. Ontology Lexicon is a linguistic object an ontology is not.[11] 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.

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 .

ontologicky vs. semaziologicky (Semanticke priznaky: kategoriálne/archysémy, difernciacne, specifikacne)

Chapter 4

Analysis of the data landscape

This section gives an overview of existing standards and formats for metadata and content annotations in the field of Language Resources and Technology together with a description of their characteristics and their respective usage in the projects and initiatives.

4.1 Metadata Formats

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

Collect number about CMD-Framework (profiles, datcats) + historical development

Collect numbers about CMD records (collections, used profiles, ...) in historical perspective

4.1.2 Dublin Core + OLAC

DC, OLAC
DublinCore Resource Types[1]

4.1.3 TEI / teiHeader

TEI/teiHeader/ODD,

[1] http://dublincore.org/documents/resource-typelist/

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 Content/Annotation Formats

CHILDES, TEI, EAF! (CES/XCES) Open Annotation Collaboration (OAC)[2]
[LAF] Linguistic Annotation Framework

4.3 Ontologies, Controlled Vocabularies, Reference Data, Authority Files

Based on popular demand, the work on reference data for the SSH-community should cover at least the following dimensions (with tentative denominations of corresponding existing vocabularies):

- Data Categories / Concepts - ISOcat
- Languages - ISO-639
- Countries - country codes
- Persons - GND, VIAF
- Organizations - GND, VIAF
- Schlagwörter/Subjects - GND, LCSH
- Resource Typology -

AAT - international Architecture and Arts Thesaurus GND - Gemeinsame Norm Datei GTAA - Gemeenschappelijke Thesaurus Audiovisuele Archieven (Common Thesaurus [for] Audiovisual Archives) VIAF - Virtual International Authority File

Other related relevant activities and initiatives

A broader collection of related initiatives can be found at the German National Library website: [3] FRBR - Functional Requirements for Bibliographic Records RDA - Resource Description and Access http://metadaten-twr.org/ - Technology Watch Report: Standards in Metadata and Interoperability (last entry from 2011) At MPDL, within the escidoc publication platform there seems to be (work on) a service (since 2009 !) for controlled vocabularies: [4] Entity Authority Tool Set - a web application for recording, editing, using and displaying authority information about entities – developed at the New Zealand Electronic Text Centre (NZETC). http://eats.readthedocs.org/en/latest/

4.3.1 ISOcat - Data Category Registry

ISO12620

[2] http://openannotation.org/
[3] http://www.dnb.de/DE/Standardisierung/LinksAFS/linksafs_node.html
[4] http://colab.mpdl.mpg.de/mediawiki/Control_of_Named_Entities

4.3.2 Classification Schemes, Taxonomies

LCSH, DDC

4.3.3 Other controlled Vocabularies

Tagsets: STTS Language codes ISO-639-1

4.3.4 Domain Ontologies, Vocabularies

Organization-Lists LT-World !?

4.4 LRT Metadata Catalogs/Collections

[DFKI/LT-World] - collection or ontology

4.4.1 CMDI

collections, profiles/Terms, ResourceTypes!

4.4.2 OLAC

4.4.3 LAT, TLA

Language Archiving Technology, now The Language Archive - provided by Max Planck Insitute for Psycholinguistics [5]

4.4.4 META-NET

4.4.5 ELRA

4.4.6 Other

LDC Linguistic Data Consortium

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

4.5 Other Metadata Catalogs/Collections

Digital Libraries

(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/

[5] http://www.mpi.nl/research/research-projects/language-archiving-technology

4.6 Summary

In this chapter, we gave an overview of the existing formats and dataset in the broad context of Language Resources and Technology

Chapter 5

Underlying infrastructure

5.1 CLARIN / CMDI

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

> 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.

The infrastructure foresees 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 accommodate existing schemas.

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

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

?MDBrowser ?MDService

5.1.1 CMDI - DCR/CR/RR

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 [13], and is implemented in *ISOcat*[1].

[1] http://www.isocat.org/

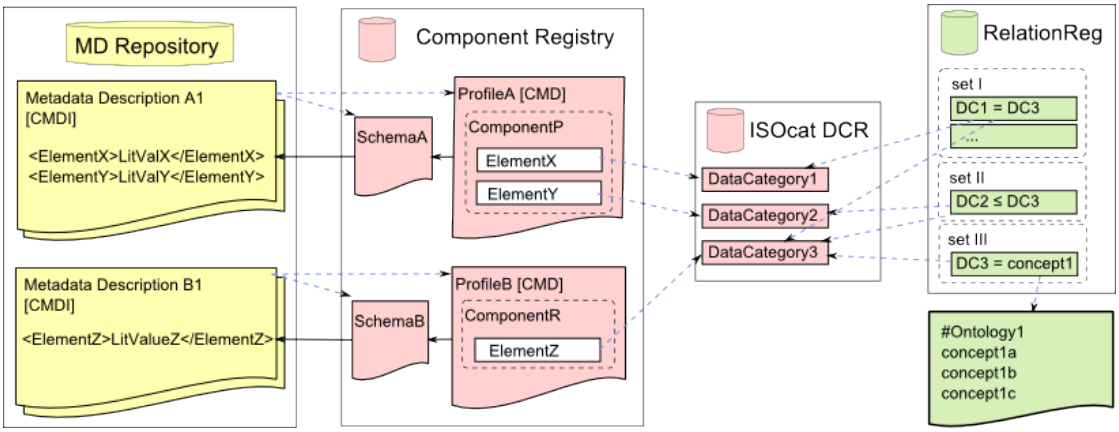

Figure 5.1: The diagram depicts the links between pieces of data in the individual registries that serve as basis for semantic mapping

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" [14]. 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].

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[2]. 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

Describe SCHEMAcat

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.

[2]sample relation set: http://lux13.mpi.nl/relcat/rest/set/cmdi

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 **??**.

5.1.2 Vocabulary Service / Reference Data Registry

Motivation & related activities in the community

The urgent need for reliable community-shared registry services for concepts, controlled vocabularies and reference data for both the LRT and Digital Humanities community has been discussed on many occasions in various contexts. Applications and tasks requiring or profiting from this kind of service comprise Data-Enrichment / Annotation, Metadata Generation, Curation, Data Analysis, etc. As there is a substantial overlap in the vocabularies relevant for the various communities and even more so a high potential for reusability on the technical level, there is a strong case for tight cooperation between different initiatives.

In the context of the CLARIN initiative, one activity to tackle this issue – mainly driven by CLARIN-NL – is the project/taskforce *CLAVAS - Vocabulary Alignment Service for CLARIN* where the plan is to reuse and enhance for CLARIN needs a SKOS-based vocabulary repository and editor OpenSKOS[3], developed and run within the dutch program CATCHplus[4]. See below for a more detailed description of this system. As of spring 2013, the Standing Committee on CLARIN Technical Centers (SCCTC) adopted the issue of Controlled Vocabularies and Concept Registries as one of the infrastructural (A-center) services to be dealt with.

In parallel, within the sister ESFRI project DARIAH a taskforce with the same goal has been set up : *Service for Reference Data and Controlled Vocabularies.* This taskforce was introduced at the 2nd VCC Meeting in Vienna in November 2012. It is conceived as a collaborative endeavor between VCC1/Task 5: Data federation and interoperability and VCC3/Task3: Reference Data Registries (and external partners). The main goal is to *establish a service providing controlled vocabularies and reference data* for the DARIAH (and CLARIN) community.

Regarding the responsibilities of the DARIAH working groups: VCC3/Task 3 identifies and recommends vocabularies relevant for the community. VCC1/Task 5 provides basic/generic services relevant for whole community. Especially, the Schema Registry, that allows to express mappings between different schemas seems to be one starting point. In accordance with the VCC1 strategy, concentrate on pulling together (pooling) existing resources and only implement necessary "glue" to put the pieces together (data conversion, service-wrappers...)

Thus there is a momentum and a high potential for a collaborative approach in at least these two big initiatives CLARIN and DARIAH, that serve a very wide-spread and diverse community.

Abstract service description

As to the service itself it is primarily meant to serve other applications, rather than being used directly by end users, but a basic user interface is still necessary for administration etc. By using global semantic identifiers instead of strings, such a service enables the harmonization of metadata descriptions and annotations and is an indispensable step towards semantic data and LOD. Besides providing vocabularies, the service should also

[3] http://openskos.org

[4] *Continuous Access To Cultural Heritage* - http://www.catchplus.nl/en/

hold and expose equivalencies (and other relationships) between concepts from different vocabularies (concept schemes). These relationships come primarily from existing mappings, but can (and hopefully will) be subsequently generated (manually) for specific subsets on demand in a community process. An example for equivalencies from Wikipedia[5]:

```
GND: 118540238 | LCCN: n79003362 | NDL: 00441109 | VIAF: 24602065 | Wikipedia-Persone
```

Vocabulary Service - CLAVAS

As described in previous section (5.1.1), a solid pilar for defining and maintaining data categories is the ISOcat data category registry. However, while ISOcat has been in productive use for some time, it is – by design – not usable for all kinds of reference data. In general, it suits well for defining concepts/data categories (with closed or open concept domains), but its complex data model and standardization workflow does not lend itself well to maintain "semi-closed" concept domains, controlled vocabularies, like lists of entities (e.g. organizations or authors). In such cases, the concept domain is not closed (new entities need to be added), but it is also not open (not any string is a valid entity). Besides, the domain may be very large (millions of entities) and has to be presumed changing (especially new entities being added).

This shortcoming leads to a need for an additional registry/repository service for this kind of data (controlled vocabularies). Within the CLARIN project mainly the abovementioned taskforce *CLAVAS* is concerned with this challenge. The foundation is the vocabulary repository and editor OpenSKOS[6].

This repository can serve as a project independent manager and provider of controlled vocabularies. One important feature of the OpenSKOS system is its distributed nature. It allows individual instances to synchronize the maintained vocabularies among each other via OAI-PMH protocol. This caters for a reliable redundant system, as multiple instances would provide identical synchronized data, while the primary responsibility for individual vocabularies could lie with different instances/organizations based on their specialization, field of expertise.

Currently, the Meertens Institute[7] of the Dutch Royal Academy of Sciences (KNAW), as well as Netherlands Institute for Sound and Vision[8] are running an instance of OpenSKOS. As the work on this vocabulary repository started in the context of a cultural heritage program, originally it served vocabularies not directly relevant for the LRT-community *GTAA - Gemeenschappelijke Thesaurus Audiovisuele Archieven* or *AAT - Art & Architecture Thesaurus*[9]. As part of the process of adaptation to the needs of CLARIN and LRT-community data categories from ISOcat have been converted into SKOS-format and ingested into the system. CLARIN Centre Vienna is also running a prototypical instance of the OpenSKOS system with ISOcat data.

A plan has been developed/adopted to support further vocabularies relevant for the community. Following are those to be handled in short-term, in order of urgency/relevance/prirority:

- the list of language codes

  url: ISO-639

- country codes

[5] page for J. W. Goethe
[6] http://openskos.org
[7] http://meertens.knaw.nl/
[8] http://www.beeldengeluid.nl/
[9] http://openskos.org/api/collections

- organization names for the domain of language resources

See 4.3 for a more complete list of required reference data together with candidate existing vocabularies and 5.1.3 for discussion on mapping the information about data categories from ISOcat to SKOS.

5.1.3 Interaction between DCR, VAS and client applications

DCR recognizes following types of data categories (Figure 5.2): simple, complex: closed, open, constrained, (container)?

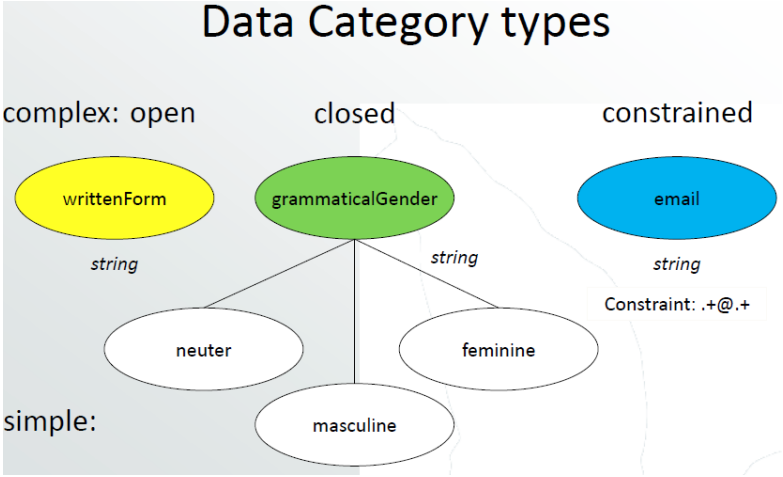

Figure 5.2: Data Category types

DC types - ISOcat introduction at CLARIN-NL Workshop

See A.1 for full DCR data model.

Export DCR to SKOS

Menzo2013-03-12 mail

The semantic proximity of a /data category/ to a /concept/ may mislead to a na"ive approach to mapping DCR to SKOS, namely mapping every data category (from one profile) to a concept all of them belonging to the ISOcat-profile:ConceptScheme. However this is not practical/useful, ISOcat as whole is too disparate, and so would be the resulting vocabulary.

A more sensible approach is to export only closed DCs as separate ConceptSchemes and their respective simple DCs as Concepts within that scheme. The rationale is, that if we see a vocabulary as a set of possible values for a field/element/attribute, complex DCs in ISOcat are the users of such vocabularies and simple DCs the DCR equivalence of values in such a vocabulary.

Menzo

Another aspect is, that a simple DC can be in valuedomains of multiple closed DCs. Also a skos:Concept can belong to multiple ConceptSchemes[10]. So there could a 1:1 one mapping [complex closed DCs] to [skos:ConceptSchemes] and [simple DCS] to [skos:Concepts]. That would automatically convey also the possibly multiplicate membership of simple DCs / skos:Concepts in closed DCs / skos:ConceptSchemes.

[10] http://www.w3.org/TR/skos-primer/#secscheme

Alternatively, for each value domain a SKOS concept scheme with SKOS concepts can be created, i.e., a SKOS concept always belongs to one concept schema, but multiple SKOS concepts refer to the same simple DC using ¡dcr:datcat/¿ (and ¡dcterms:source/¿). This is, how the export for CLAVAS currently works.[11][12]

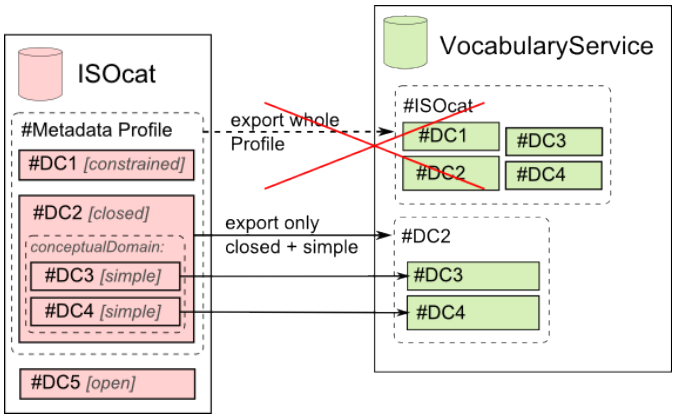

Figure 5.3: The data flow and linking between schema, data categories and vocabularies

Open or constrained DCs are not exported as they don't provide anything to a vocabulary. There is no need to express the relationship between this constrained DC and the vocabulary in CLAVAS itself. Indeed it is not possible to express the conceptualDomain/range of a data category within SKOS.

However, they can refer to a CLAVAS vocabulary. Indeed, providing vocabularies for constrained but large and complex conceptual domains is the main motivation for the vocabulary repository.

However it needs to be yet assessed how useful this approach is. In the metadata profile there are many closed DCs with small value domains. How useful are those in CLAVAS?

Originally, the vocabulary repository has been conceived to manage rather large and complex value domains, that do not fit easily in the DCR data-model. Where the value domains are big (ISO 639-3) or can only be partially enumerated (organization names) ISOcat can't/shouldn't contain the value domains but just refer to CLAVAS, i.e., ISOcat wouldn't be a provider. Still there are some closed DCs which might be good vocabulary providers, e.g., /linguistic subject/ (DC-2527/), and still also need to stay in ISOcat. I think at some point we should create a smaller set of metadata DCs to be harvested by CLAVAS. Therefore a threshold seems sensible, where only value domains with more then 20, 50 or 100 values are exported.

Vocabulary linking and use

Currently (before integration of VAS and DCR), the only possibility to constrain the value domain of a data category is by the means a XML Schema provides

check xml schema possibilities to restrict values

, like a regular expression. So for the data category languageID DC-2482 the rule looks like:

```
<dcif:conceptualDomain type="constrained">
```

[11] http://www.isocat.org/rest/profile/5.clavas

[12] https://trac.clarin.eu/browser/cats/ISOcat/trunk/mod-ISOcat-interface-rest/representations/dcs2/clavas.xsl

```
        <dcif:dataType>string</dcif:dataType>
        <dcif:ruleType>XML Schema regular expression</dcif:ruleType>
        <dcif:rule>[a-z]{3}</dcif:rule>
    </dcif:conceptualDomain>
```

A current proposal by Windhouwer

Menzo2013-03-12 mail

for integration with CLAVAS foresees following extension:

```
<clavas:vocabulary href="http://my.openskos.org/vocab/ISO-639" type="closed"/>
```

`@href` points to the vocabulary. Actually a PID should be used in the context of ISOcat, but it is not clear how persistent are the vocabularies. This may pose a problem as part of DC specification may now have a different persistency then the core.

`@type` could be `closed` or `open`. `closed`: only values in the vocabulary are valid. `open`: the values in the vocabulary are hints/preferred values. Basically the DC itself is then open.

This would yield a definition of the conceptualDomain for the data category as follows:

```
<dcif:conceptualDomain type="constrained">
   <dcif:dataType>string</dcif:dataType>
   <dcif:ruleType>XML Schema regular expression</dcif:ruleType>
   <dcif:rule>[a-z]{3}</dcif:rule>
</dcif:conceptualDomain>
<dcif:conceptualDomain type="constrained">
   <dcif:dataType>string</dcif:dataType>
   <dcif:ruleType>CLAVAS vocabulary</dcif:ruleType>
    <dcif:rule>
       <clavas:vocabulary href="http://my.openskos.org/vocab/ISO-639" type="closed"/>
    </dcif:rule>
</dcif:conceptualDomain>
```

I.e. the new rule pointing to the vocabulary would be *added*, so that tools that don't support CLAVAS lookup but are capable of XSD/RNG validation, can still use the regular expression based definition.

Integrate:

ISOcat refers to CLAVAS as a hint, the metadata schema is the final one that has the real CLAVAS vocabulary reference, i.e., no reference to CLAVAS via ISOcat.

Note though, that anything stated in the DC specification is not binding, but rather a generic hint or recommendation,

check: it is not "normative"

. (Even if the DC is closed.) The authoritative/normative information is in the schema. A schema modeler, (concept)linking an element in the schema to a DC can decide to have another restriction for the values allowed in that element. The information from DCR serves as recommendation or default.

Modelling the vocabulary reference in the schema It needs to be yet defined how the information about the vocabulary can be translated into a valid schema representation. One brute-force approach would be to explicitely enumerate all the values from the vocabulary. This is being currently done within the CMD-framework with the language-codes

cmd-component ISO-639

. However there is clearly a limit to this approach both in terms of size of the vocabulary (ISO-639 contains 7.679 items (language codes) adding some 2MB to each schema referencing it) and its stability/change rate — ISO-639 is a standard with a fixed list, however most other vocabularies are more volatile (think organization).

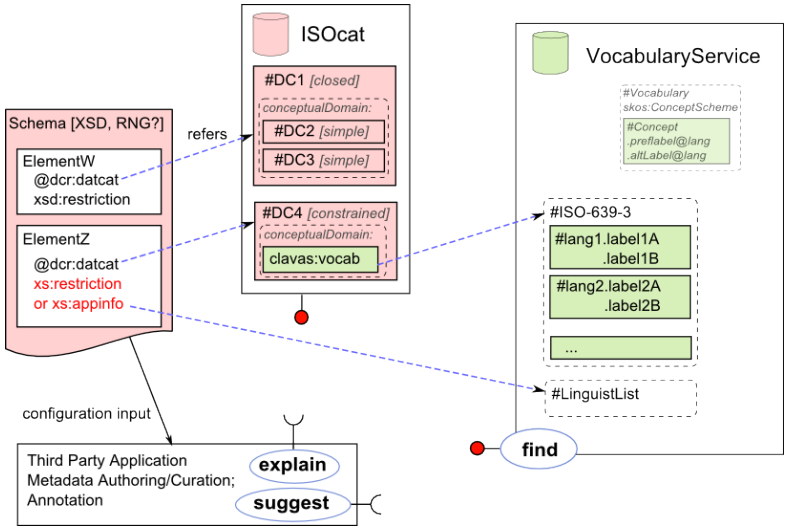

Figure 5.4: The data flow and linking between schema, data categories and vocabularies

Most of these vocabularies also cannot be seen as closed-constrained, i.e. the list that is provided, provides a recommended orthography variant for a given entity, still allowing other values for given field rather than resricting the values to only the items from the vocabulary (think organizations).

So this has to be solved in "soft" way. Most schema languages allow to annotate the schema. This is already used with DCR, adding the `@dcr:datcat` into schema elements. Also CMDI (ComponentRegistry when generating schemas) puts information in ¡xs:appinfo/¿.

Tools like Arbil can get access to these annotations, e.g., a reference to a CLAVAS vocabulary, and act upon it, i.e., use OpenSKOSs autocomplete API. Normal XSD validation then wouldn't validate if a value actually is part of the vocabulary. This isn't a problem if the vocabulary is open, e.g., organisation names, but it is when the value domain is closed, e.g., ISO 639-3. In the latter case the XSD generation might have two modes: a lax (smaller) version which doesn't contain the closed vocabulary as an enumeration and leaves it to the tool, and a strict version which does contain the vocabulary as an enumeration. Probably the latter should stay the default, but Arbil could request the lax version leading to smaller and quicker XSD validation inside the tool.

With this proposal, ISOcat constrained DCs can refer to a CLAVAS vocabulary as a way to constrain (we stretch this a bit if a vocabulary is 'open', e.g., like organization names where it provides the preferred spelling of known organizations but still has to be possible to add new organization names, not in the vocabulary).

In ISOcat such constraints have the same status as, for example, the data type, which is that ISOcat just provides hints it has no way to enforce this. Look at CMDI where the CMDI elements refer to a ISOcat DC via a concept link but they may have a completely different data type. In an ideal world the Component Editor would take over the data type and the CLAVAS vocabulary from the linked DC specification. This way the reference to the CLAVAS vocabulary ends up in the CMD component/profile specification and the derived XSD, and can be used by tools that support CLAVAS, e.g., Arbil (well its in the planning).

something similar for the link to an EBNF grammar in SCHEMAcat:

```
<scr:valueSchema
```

```
xmlns:scr="http://www.isocat.org/ns/scr"
pid="http://hdl.handle.net/1839/00-SCHM-0000-0000-004A-A"
type="ISO 14977:1996 EBNF"/>
```

Finally, the client application (e.g. a metadata editor) is configured/guided by the schema. It can use the reference to the DC to fetch explanations (semantic information) (and translations) from ISOcat, but it is bound to the value range as restricted by the schema.

Could the application use the the vocabulary indication in DC-spec as default or fallback?

5.1.4 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[13]. 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.

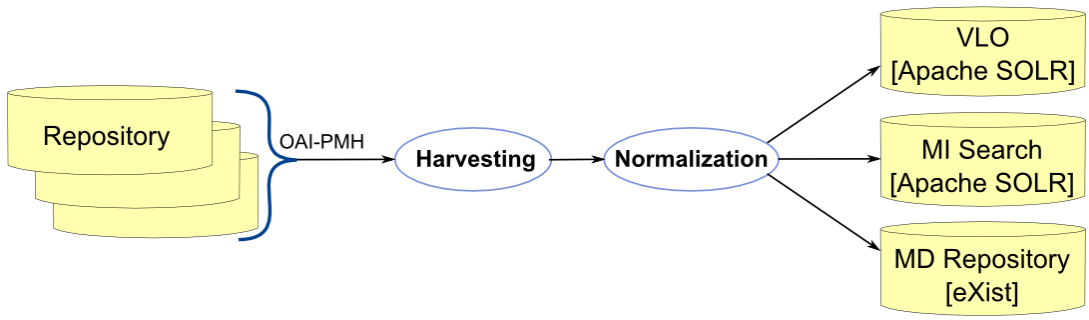

Figure 5.5: Within CMDI, metadata is harvested from content providers via OAI-PMH and made available to consumers/users by exploitation side components

The first stable and publicly available application providing access to the collected metadata of CMDI has been the `VLO - Virtual Language Observatory`[14][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. Underlying search engine is the widely used full-text search engine Apache Solr[15]. 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`[16]. 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.

[13] http://catalog.clarin.eu/oai-harvester/
[14] http://www.clarin.eu/vlo/
[15] http://lucene.apache.org/solr/
[16] http://www.meertens.knaw.nl/cmdi/search/

describe indexing and search

MI Search Engine

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]

5.2 Content Repositories

Metadata is only one aspect of the availability of resources. It is the first step to announce and describe the resources. However it is of little value, if the resources themselves are not equally well accessible. Thus another pillar of the CLARIN infrastructure are Content Repositories - centres to ensure availability of resources.

The requirements for these repositories: PIDs, CMD, OAI-PMH

center-B paper

5.3 Distrbuted system - federated search

Metadata -¿ harvesting via OAI-PMH but Content search has to be really distributed.
?

Z39.50/SRU/SRW/CQL LoC

OAI-PMH

Chapter 6

Semantic Mapping Component

6.1 Data Model?

Terms ? move to SKOS ?
RDF

6.1.1 CMD namespace

Describe the CMD-format?

6.1.2 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[1] (`dc.title`) and on the dot-notation used in IMDI-browser[2] (`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 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.

6.1.3 Query language

CQL?

6.2 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.

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. [3]

In the operation mode, the application accepts any index (*smcIndex*, cf. 6.1.2) 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]
```

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

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

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

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

```
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[4] 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.*

6.3 Semantic Mapping on instance level

6.3.1 Mapping from strings to Entities

Find matching entities in selected Ontologies based on the textual values in the metadata records.

Identify related ontologies: LT-World [2]
task:

1. express MDRecords in RDF
2. identify related ontologies/vocabularies (category → vocabulary)

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

3. use a lookup/mapping function (Vocabulary Alignement Service? CATCH-PLUS?)

$$lookup(Category, Literal) \rightarrow ConceptualDomain??$$

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

6.3.2 Linked Data - Express dataset in RDF

I do think that ISOcat, CLAVAS, RELcat, an actual language resource all provide a part of the semantic network.

And if you can express these all in RDF, which we can for almost all of them (maybe except the actual language resource ... unless it has a schema adorned with ISOcat DC references ... <insert a SCHEMAcat plug ;-) >, but for metadata we have that in the CMDI profiles ...) you could load all the relevant parts in a triple store and do your SPARQL/reasoning on it. Well that's where I'm ultimately heading with all these registries related to semantic interoperability ... I hope ;-)

Menzo

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.

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

check/install: raptor for generating dot out of rdf

5

defining the Mapping:

1. convert to RDF translate: MDRecord → [#mdrecord #property literal]
2. map: #mdrecord #property literal → [#mdrecord #property #entity]

6.4 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))

[5] http://librdf.org/raptor/

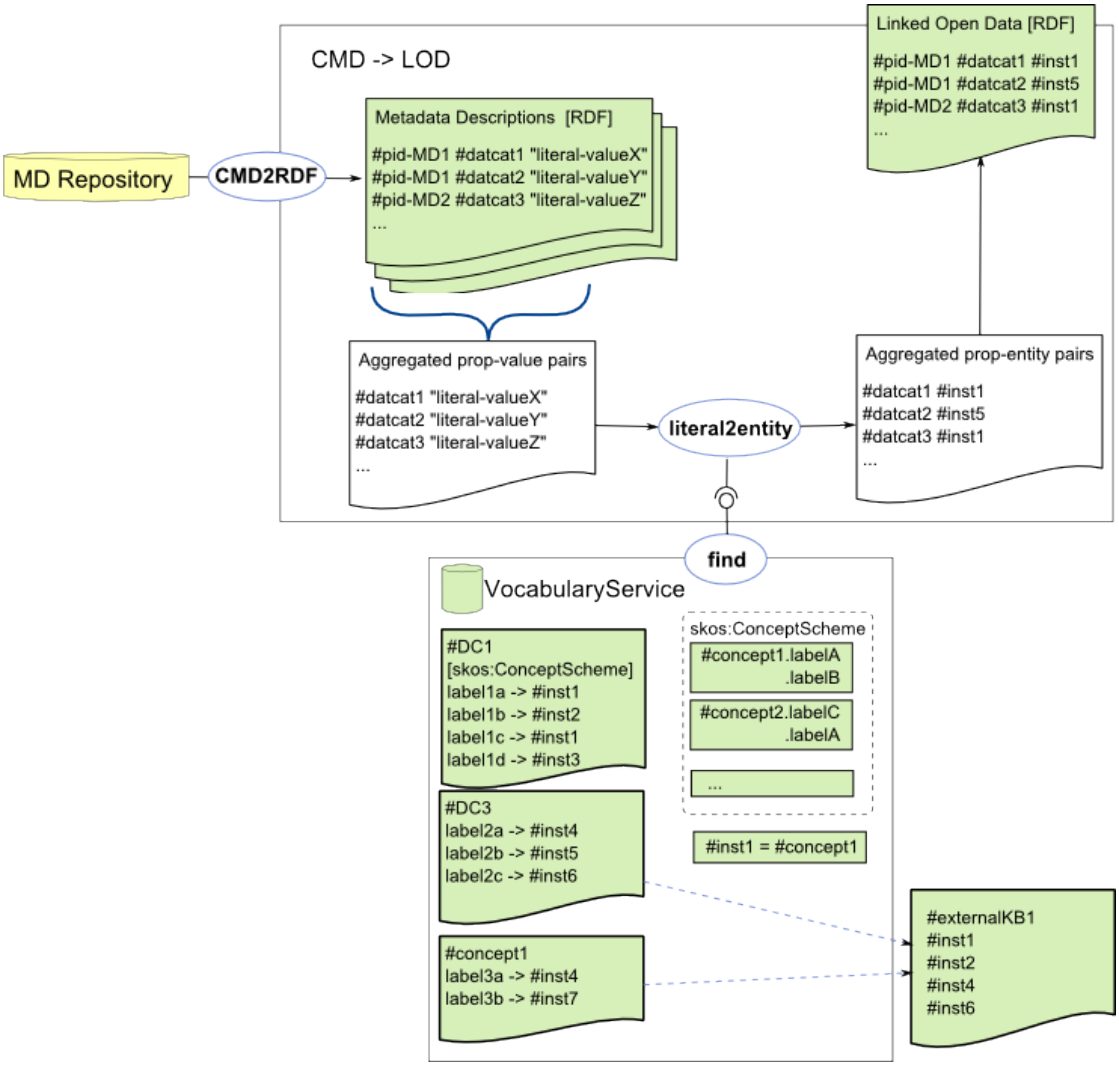

Figure 6.1: The process of transforming the CMD metadata records to and RDF representation

6.4.1 Query Expansion

6.5 Semantic Mapping in Metadata vs. Content/Annotation

AF + DCR + RR

Chapter 7

Implementation

The core function of the SMC is 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[1].

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.

7.1 Initialization

First there is an initialization phase, in which the application fetches the information from the source modules (cf. **??**). 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.

7.2 SMC as module for Metadata Repository

(MD)search frameworks:

Zebra/Z39.50 JZKit

Lucene/Solr

eXist - xml DB

[1] http://svn.clarin.eu/SMC

7.3 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.

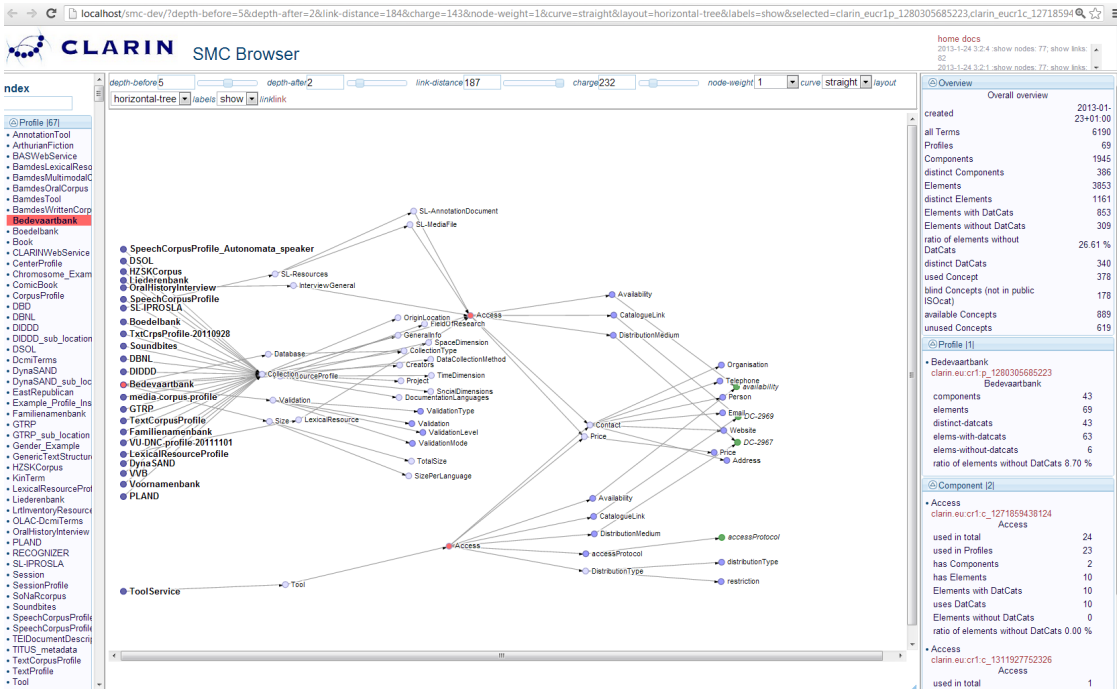

Figure 7.1: Screenshot of the SMC browser

7.4 SMC LOD

read: Europeana RDF Store Report

install Jena + fuseki

234

Load data: relcat, clavas, olac-and-dc-providers cmd, lt-world?

7.5 User Interface?

7.5.1 Query Input

7.5.2 Columns

7.5.3 Summaries

7.5.4 Differential Views

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

7.5.5 Visualization

Landscape, Treemap, SOM

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

[2] http://jena.apache.org

[3] http://jena.apache.org/documentation/serving_data/index.html

[4] http://csarven.ca/how-to-create-a-linked-data-site

Chapter 8

Evaluation

8.1 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.

8.2 Research Questions

8.3 Sample Queries

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

8.4 Usability

Chapter 9

Conclusions and Future Work

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.

Bibliography

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[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.

Appendix A

Data model ?

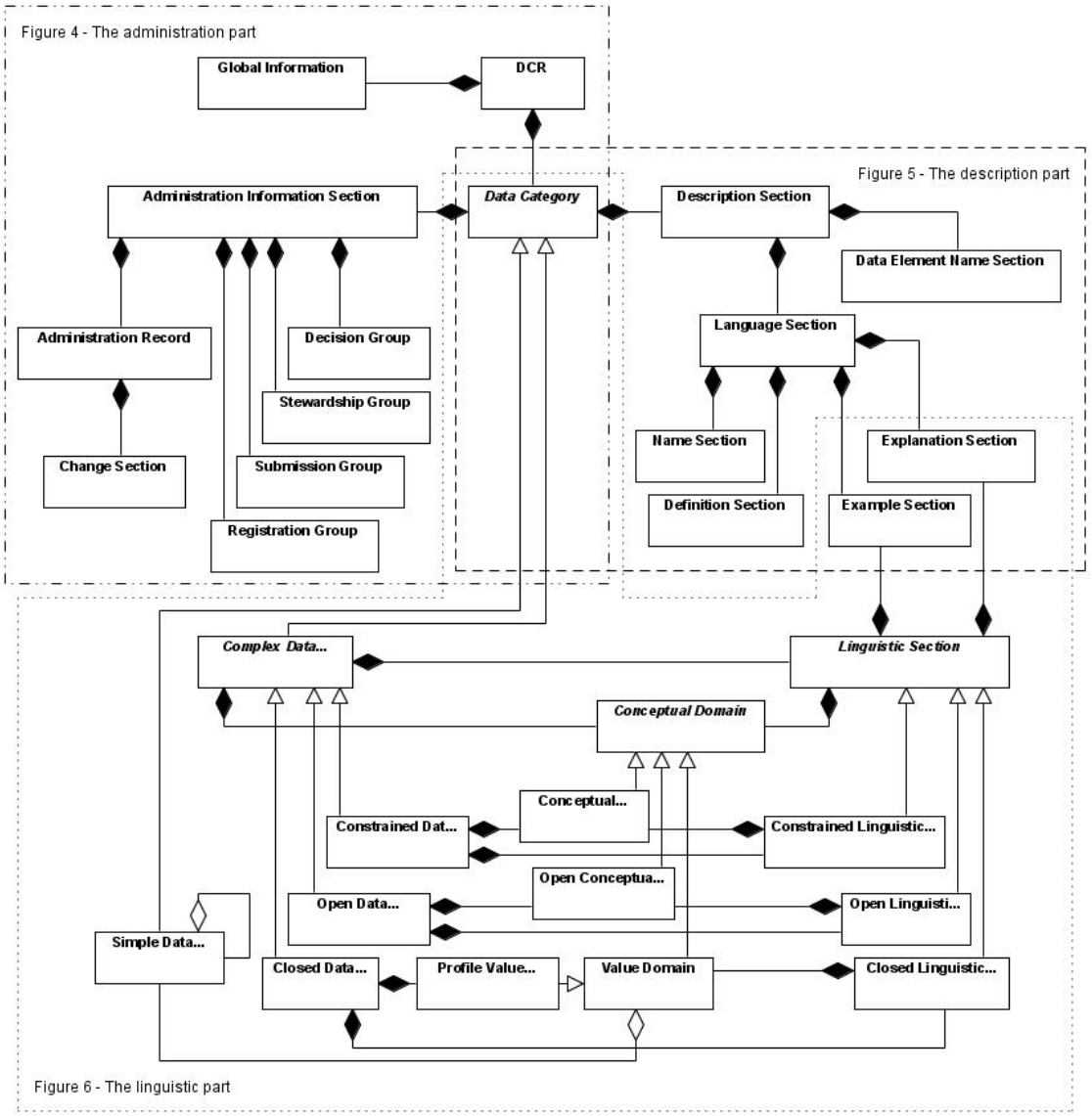

Figure A.1: DCR data model

DCR data model