# SMC4LRT - Master Outline

Matej Durco

September 13, 2011

## Contents

# 1 Introduction

Title: Semantic Mapping (Component) for Language Resources

## 1.1 Main Goal

We propose 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/knowledge-base via ontology-driven browsing.

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.

All these scenarios require a preprocessing step, that would produce the underlying linkage, both between categories/concepts and between instances (mapping literal values to entities). We refer to this task as semantic mapping, that shall be accomplished by coresponding "Semantic Mapping Component". In this work the focus lies on the process/method, i.e. on the specification and (prototypical) implementation of the component rather than trying to establish some final/accomplished mapping. Although a tentative/naive alignement 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 aiming at creating the actual sensible mappings usable for real tasks.

Actually due to the great diversity of resources and research tasks such a "final" complete mapping/alignement does not seem achievable at all. Therefore also the focus shall be on "soft", dynamic mapping, investigating the possibilities/methods to enable the users to adapt the mapping or apply different mapping with respect to their current task or research question, essentially being able to actively manipulate the recall/precision ratio of their searches. This entails the examination of user interaction with and visualization of the relevant information in the user interface and enabling the user to act upon it.

## 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 task/function of Semantic Search on concept and on individuals level and the underlying Semantic Mapping and the requirements within the defined context, followed by a design proposal for an appropriate component fitting within the infrastructure. especially with focus on the feasibility of employing ontology mapping and alignement techniques and tools for the creation of mappings.

In a prototype we want to deliver a proof of the concept, combined with an evaluation to verify the claims of fitness for the purpose. This evaluation is twofold. It shall verify the ability of the system to support dynamic mapping based on a set of test queries and secondly the usability of the ui-controls.

+? Identify hooks into LOD?

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 main result of this work will be a specification of the pair of components the Semantic Search and the underlying Semantic Mapping. This propositions will be supported by a proof-of-concept implementation of these components and an evaluation of querying the dataset comparing traditional search and semantic search.

One important by-product of the work will be the original dataset expressed as RDF with links into existing datasets/ontologies/knowledgebases, building a base for another nucleus of Linked Open Data.

Specification definition of a 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

### 1.4 State of the Art

- VLO - Virtual Language Observatory `http://www.clarin.eu/vlo/`, [?]
- LT-World ontology-based `http://www.lt-world.org/`, [?]
- VAS - Catch Plus
- OAEI

### 1.5 Keywords

Metadata interoperability, Ontology Mapping, Schema mapping, Crosswalk, Similarity measures, LinkedData Fuzzy Search, Visual Search?
Language Resources and Technology, LRT/NLP/HLT
Ontology Visualization
Federated Search, Distributed Content Search (ILS - Integrated Library Systems)

## 2 Related Work

### 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

**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.[**?**] 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 [**?**]

LingInfo/Lemon [**?**]

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 dowhn 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** "a 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 Analysis

## 4.1 Data landscape

Describe situation regarding the datasets and formats
collections, profiles/Terms, ResourceTypes!
DC, OLAC, ISLE/IMDI, CHILDES, TEI, EAF! (CES/XCES)

## 4.2 Infrastructure

CMDI [**?**]

## 4.3 Ontologies, Controlled Vocabularies, Knowledge Organizing Systems

### 4.3.1 Classification Schemes, Taxonomies

LCSH, DDC

### 4.3.2 Other controlled Vocabularies

Tagsets: STTS Language codes ISO-639-1

### 4.3.3 Domain Ontologies, Vocabularies

Organization-Lists LT-World !?

## 4.4 Use Cases

- MD Search employing Semantic Mapping
- MD Search employing Fuzzy Search
- Content Search
- Combined MEtadata Content Search
- Visualization of the Results - charts on facets/dimensions
- Create and publish Virtual Collection based on complex Search (intensional/extensional)
- Let Create ad-hoc corpus

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.

# 5 Semantic Mapping

## 5.1 Profiles to Data Categories

CMD:Profile.Comp.Elem -¿ DatCat

## 5.2 Semantic Relations between (Data)Categories

Relation Registry

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

## 5.3 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 [**?**]
task:

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

   | function lookup: Category x String -¿ ConceptualDomain |
   |---|

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

## 5.4 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))

### 5.5 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]

### 5.6 Content/Annotation

AF + DCR + RR

### 5.7 Visualization

Landscape, Treemap, SOM

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

## 6 System Design

SOA

### 6.1 Architecture

Makes use of mulitple Components of the established infrastructure (CLARIN ) [**?**], [**?**]:

- Data Category REgistry,
- Relation Registry
- Component Registry
- Vocabulary Alignement Service

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

### 6.2 CMDI

MDBrowser MDService

### 6.3 Query Language

CQL?

### 6.4 User Interface

#### 6.4.1 Query Input

#### 6.4.2 Columns

#### 6.4.3 Summaries

#### 6.4.4 Differential Views

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

## 7 Evaluation

### 7.1 Research Questions

### 7.2 Sample Queries

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

### 7.3 Usability

## 8 Conclusions and Futur Work

## 9 Questions, Remarks

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