SMC4LRT

Semantic Mapping Component for
Language Resources and Technology

2011-06-06

Matej Ďurčo, ICLTT, Vienna;

Context

ICLTT

on **Language Resource and Technology**

- CLARIN – Common Language Resources and Technology Infrastructure
- CMDI - CLARIN Metadata Infrastructure
  heterogeneous collection of (Metadata about) Resources
- ISOcat (ISO 12620) - a framework within ISO TC 37 for defining:
- Data Categories – Definitions of widely accepted linguistic concepts

apply **Semantic Technologies**

- Ontology Mappping / Schema Mapping
- Ontology Browsing / Visualization
- Linked Open Data

Main Goal/s

ICLTT

- Enhance Metadata Search → **Semantic Search**

  Basic Idea

  query:

  ```
  Actor.Name any Peter
  ```

  \+ relations:
  (#DatCat)

  ```
  #sameAs (#Actor, #Person)
  #sameAs (#Name, #FullName)
  ```

  = expanded query:

  ```
  Actor.Name any Peter
  OR Actor.FullName any Peter
  OR Person.Name any Peter
  OR Person.FullName any Peter
  ```

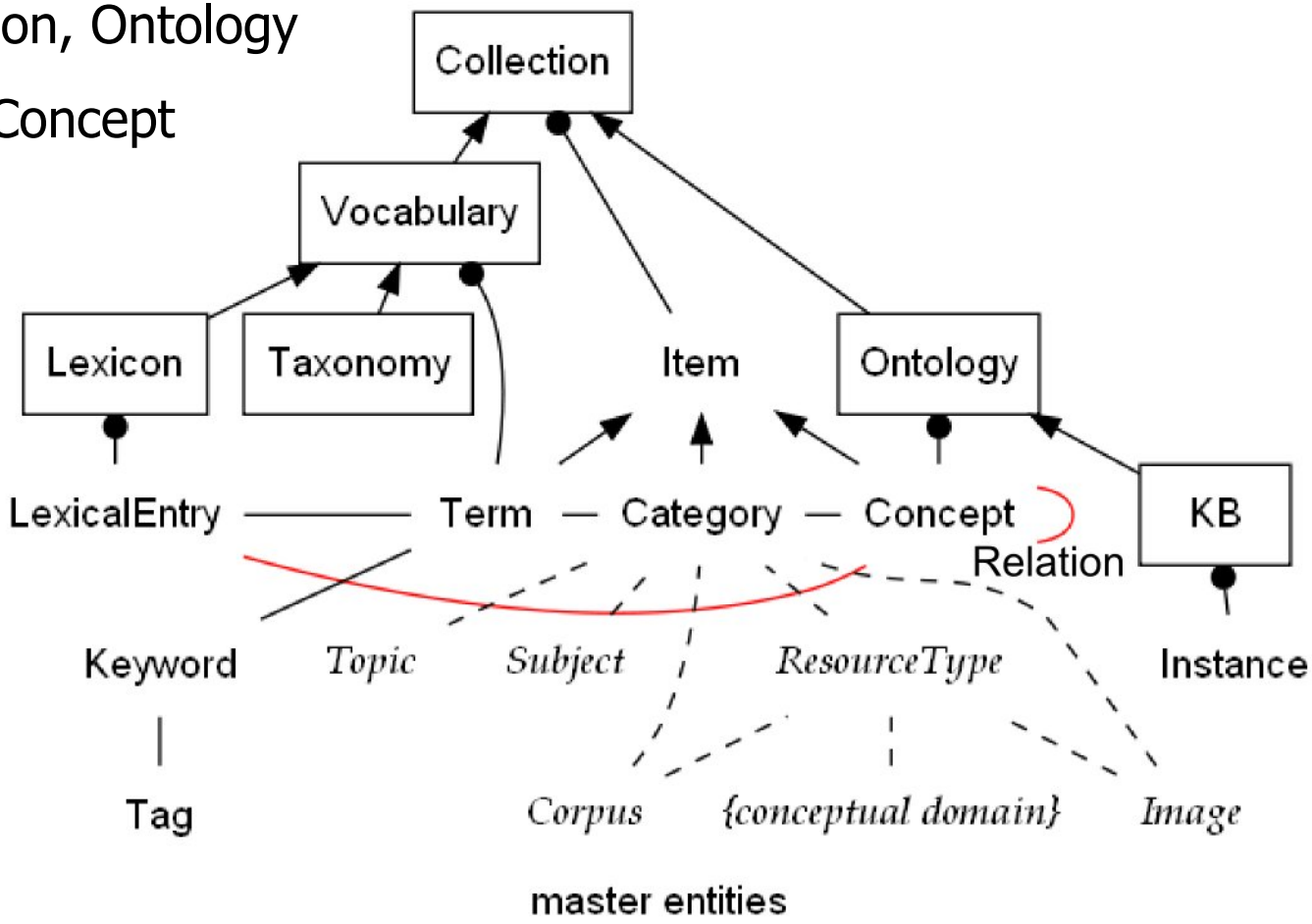
  (Class level)

- Semantic Browsing
  - Browse Metadata/Resources via ontologies (LT-World) (Instance-Level)
- Interoperability / Reuses
  - Connect dataset to Linked Open Data

Definitions

ICLTT

- Vocabulary, Lexicon, Ontology
- Term, Category, Concept



?

- MD Profile / Schema
- MD Description

Components

ICLTT

- DataCategoryRegistry - isocat DCR (ISO/TC37)
  Define/Standardize a reusable set of (basic) data categories
- CMDI - ComponentRegistry
  define profiles/schemas at will, but reference DatCats!
- CMDRSB - Repository/Service/Browser
  CMDI exploitation-side trinity http://clarin.aac.ac.at/MDService2/
- RelationRegistry
  allows defining relations between DatCats
- VLO - Virtual Language Observatory
  faceted browser for CLARIN Metadata, maps all hetergeneous information from all profiles to 10 facets!
- VAS – Vocabulary Alignment Service (CATCHPlus.nl)
  find concept to literal, find aligned concepts
- LT-World - Domain ontology

Components - CMDI

ICLTT

Perform search/browsing on the metadata catalog using the ISO DCR and other concept registries and CLARIN relation registry

Create metadata schema from selection of existing components. Allow creation of new components if they have references to ISOcat

Search Service

Semantic Mapping

Relation Registry

ISOcat Concept Registry

DCMI Concept Registry

other Concept Registry

CLARIN Component Registry

Joint Metadata Repository

Metadata harvesting by OAI protocol

Metadata Repository

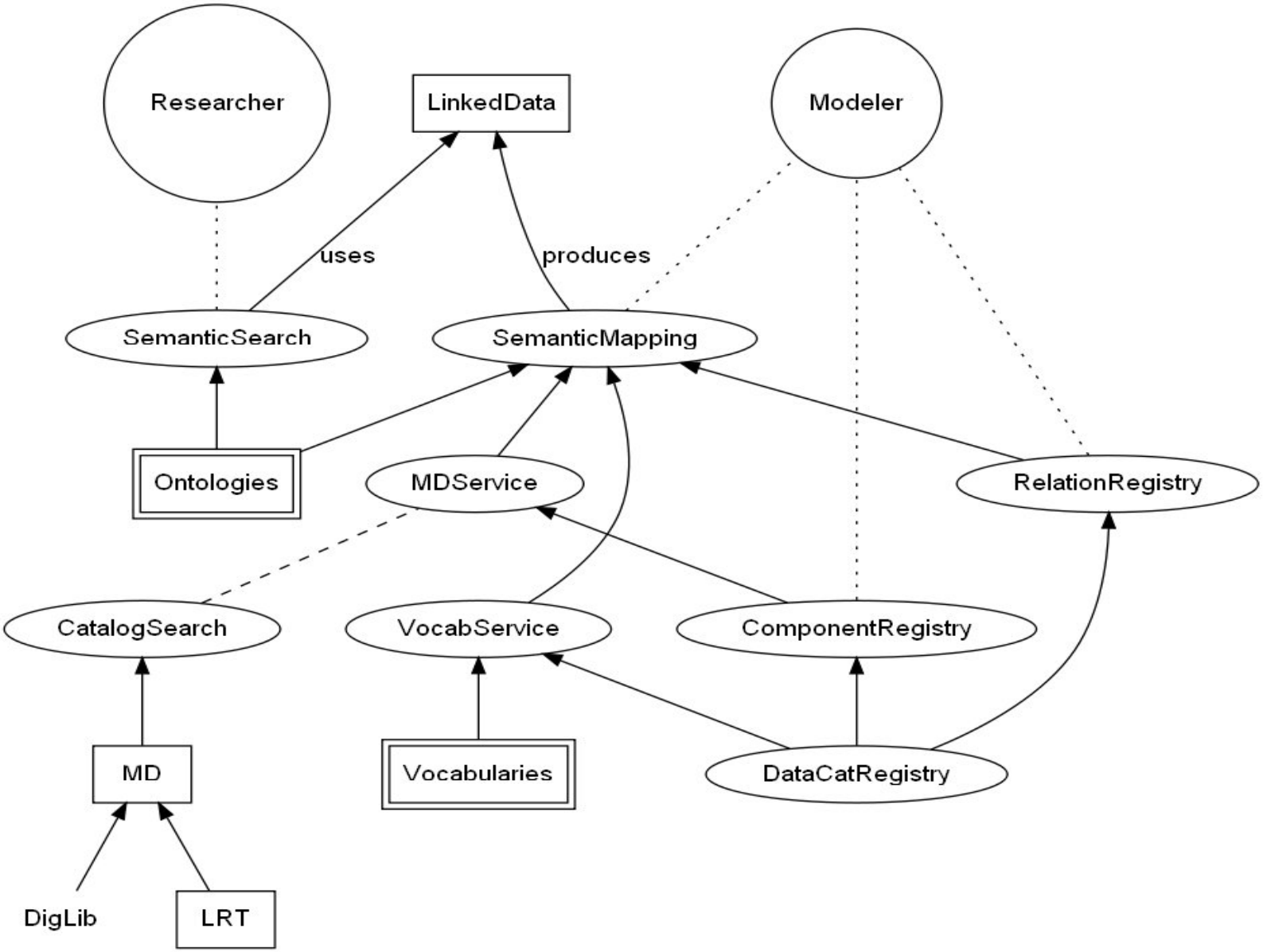
Metadata Repository

Metadata descriptions created

Metadata component profile was selected from metadata component registry

Components - dependencies

ICLTT



SemMap dependencies

Approach – Class/Concept level

ICLTT

- Use linkage: Profiles → Data Categories ← Relation Registry

MDRecord

```
<CMD>
<Header>
 <MdProfile>{profileID}</MdProfile>
<Components><{profileName}>
      <{component}>
        <{element}>
```

CMD-Profile-Specification

```
<CMD_ComponentSpec>
 <Header><ID>{profileID}</ID>...</Header>
 <CMD_Component name="{profileName}">
  <CMD_Component name="{component}">
   <CMD_Element name="{element}"
          ConceptLink="{datcat-uri}">
```

Data Category Registry

```
<dcif:dataCategorySelection>
  <dcif:dataCategory
     pid="{datcat-uri}" >
        {detail-information}
```

Relation Registry

```
<rdf:RDF>
<rdf:Description rdf:about="{datcatX-uri}">
<sameAs rdf:resource="{datcatY-uri}"/>
</rdf:Description>
```

1. just mapping based on the ConceptLink resolvable via ComponentRegistry
   different Profile/Elements pointing to the same DatCat

2. use Information from Relation Registry:
   a) equivalence relation between DatCats
   b) equivalence relation also between Component DatCats (yet to come)
   c) use also other relations in Relation Registry (`subClassOf`, synonymy?, …)

3. Apply selected (user-defined) relation-sets from Relation Registry

Approach – Individuals/Instance Level

ICLTT

One step when (pre)processing incoming new MD-sets

1. Express MD-Records as RDF-triples:

```
<#mdrecord #property "string-value">
```

2. Identify potential target Domain Ontologies/Vocabularies

3. Create inverted Index:

```
label → entity
```

Category	Label	Entity
dc:Organization	„MPI“	#MPI
	„Max-Planck...“	#MPI
	„DFKI“	#DFKI
	„De Fo Kü In“	#DFKI
skos:LCSH	„19th Poetry“	lcsh:19thPoetry
skos:DDC	„19th Poetry“	ddc:19thPoetry

4. Define lookup function:

```
lookup(category, string-value)
  → <external-entity, measure>
```

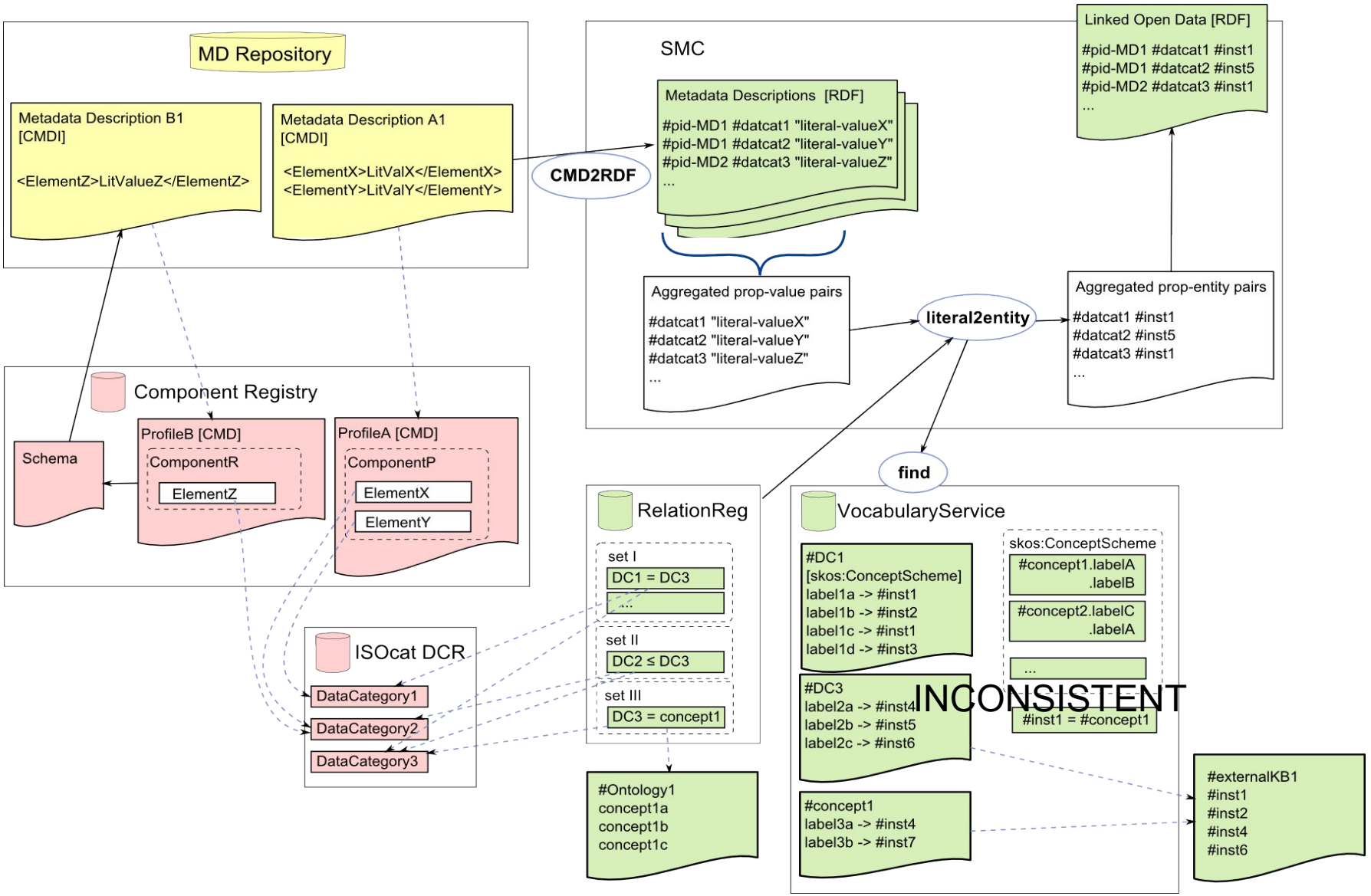
5. Enrich dataset with new facts:

```
<#mdrecord #property #external-entity>
```

Property-values of Metadata-Records are linked to instances of domain-ontologies

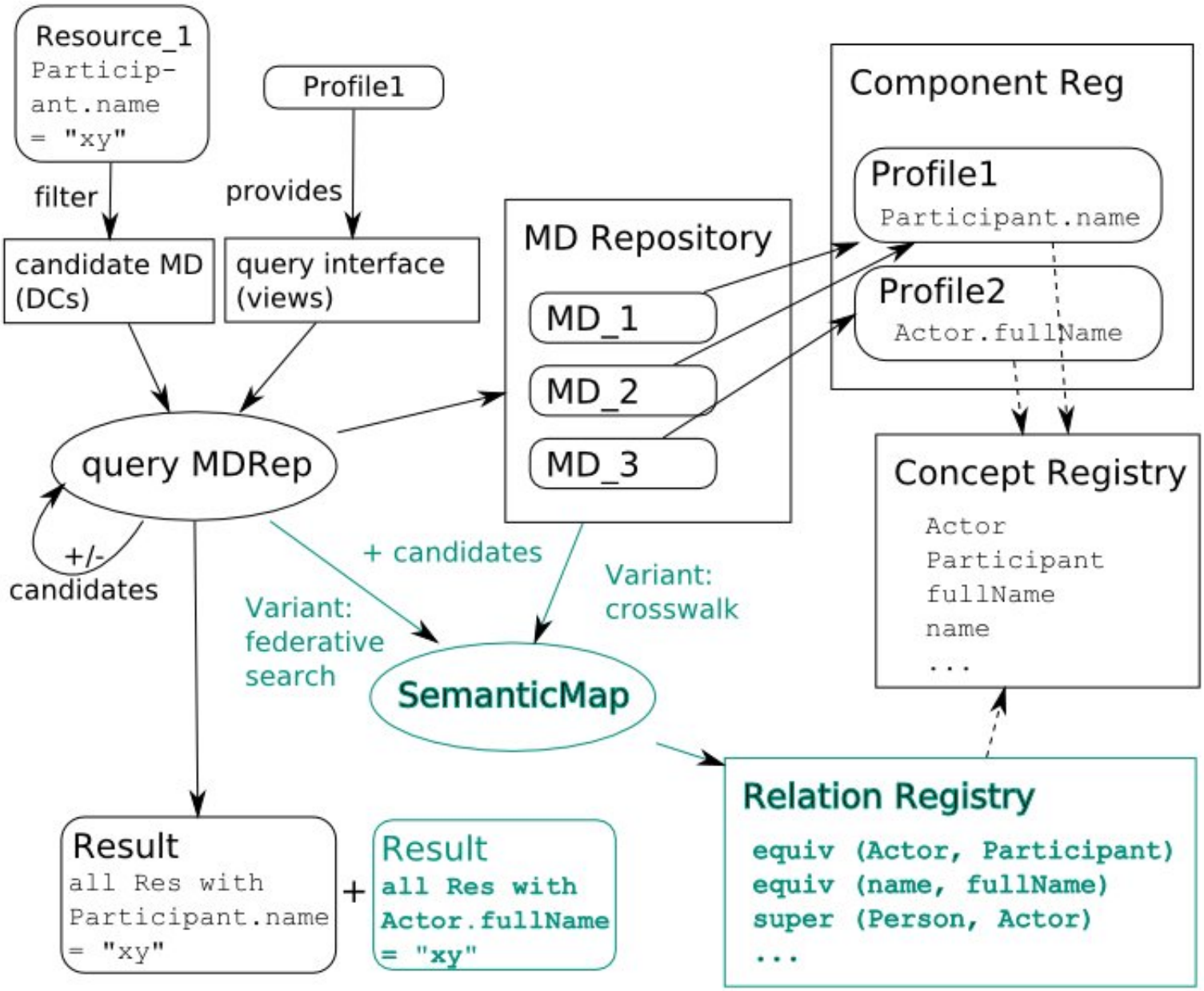
Semantic Mapping - Linking and Data Flow

ICLTT



Semantic Search - Query sequence

ICLTT



Candidate Categories/Properties

ICLTT

- ResourceType, Format, AnnotationLevelType
  → *map to:* isocat-DataCategories
  (Thematic Views: Metadata, Morphosyntax, ...)
- Genre, Topic, Subject
  → *map to:* Taxonomies, Library Classification systems
  (LCSH, DDC, Dornseiff,...)
- Project, Institution, Person, Publisher
  open controlled vocabularies (real entities)
  → *map to:* LT-World (perhaps others: LCCN, DBPedia?)

Expected Results

ICLTT

- Specification + Prototype of a Semantic Mapping Component
  allowing to transform CMD-Metadata into RDF

- Specification + Prototype of a Semantic Search Component
  REST-WebService enriching the MD-Search, allowing query expansion and ontology/concept-based search

- CLARIN Metadata expressed as RDF/LOD-Dataset

Next Steps

ICLTT

- Literature → Related Work
  - Linked Open Data
  - Ontology Mapping
  - Ontology Browsing/Visualization

- Analyze Data
  - Existing MD-Schemas (DC, OLAC, MODS, TEI, IMDI, CMD, ...)
  - LT-World Ontology
  - SKOS-Data available via Vocabulary Alignement Service
  - LCSH, LCCN
  - DBPedia

References - LRT

ICLTT

[1] 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.

[2] 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.

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

[4] E. Hinrichs, P. Banski, K. Beck, G. Budin, T. Caselli, K. Eckart, K. Elenius, G. Faa, M. Gavrilidou, V. Henrich, V. Quochi, L. Lemnitzer, W. Maier, M. Monachini, J. Odijk, M. Ogrodniczuk, P. Osenova, P. Pajas, M. Piasecki, A. Przepiorkowski, 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.

References – Semantic Technologies

ICLTT

[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.