The CMD Cloud

Matej Ďurčo, Menzo Windhouwer
Institute for Corpus Linguistics and Text Technology (ICLTT), The Language Archive - DANS
Vienna, Austria, The Hague, The Netherlands
matej.durco@oeaw.ac.at, menzo.windhouwer@dans.knaw.nl

Abstract

The CLARIN Component Metadata Infrastructure (CMDI) established means for flexible resource descriptions for the domain of language resources with sound provisions for semantic interoperability weaved deeply into the meta model and the infrastructure. Based on this solid grounding, the infrastructure accommodates a growing collection of metadata records. In this paper, we give a short overview of the current status in the CMD data domain on the schema and instance level and harness the installed mechanisms for semantic interoperability to explore the similarity relations between individual profiles/schemas. We propose a method to use the semantic links shared among the profiles to generate/compile a similarity graph. This information is further rendered in an interactive graph viewer the SMC Browser. The resulting interactive graph offers an intuitive view on the complex interrelations of the discussed dataset revealing clusters of more similar profiles. This information is useful both for metdata modellers, for metadata curation tasks as well as for general audience seeking for a 'big picture' of the complex CMD data domain.

Keywords: semantic mapping, metadata, research infrastructure

1. Introduction

The Component Metadata Infrastructure (CMDI, (Broeder et al., 2010)) conceived within the CLARIN project is now 5 years old and thriving. By allowing a flexible yet harmonized definition of metadata schemas, it has offered a robust common framework for consolidating the scattered landscape of resource descriptions in the LRT community, without trying to impose/prescribe one schema to cover all the resources (which seems futile in the light of the variety of resources to be described).

A look into the data domain shows that the basic concept of a flexible metamodel with integrated semantic layer is being taken up by the community. Metadata modellers are increasingly making use not only of the infrastructure, but are also reusing the modelling work done so far.

In this paper, we first – for methodical foundation – briefly summarize previous work, then give a short overview of the current status of the infrastructure both on the schema and instance level. As the main contribution – grounded in the semantic mapping mechanisms of CMDI – we propose a mechanism to compute and explore the relation/similarity among the profiles defined in CMD, delivering a bigger overall picture of the domain.

2. Previous work

Our task of determining similarity between schemas can be formulated as the schema/ontology matching problem. There is a plethora of work on methods and technology in the field of *schema and ontology matching* as witnessed by a sizable number of publications providing overviews, surveys and classifications of existing work ((Kalfoglou and Schorlemmer, 2003; Noy and Stuckenschmidt, 2005; Shvaiko and Euzenat, 2012; Amrouch and Mostefai, 2012) and more).

Although the semantic layer of the CMD Infrastructure, which integrates the task of identifying semantic correspondences directly into the process of schema creation, makes to a high degree obsolete the need for complex a posteriori schema matching/mapping techniques, still, for the discussed task of schema similarity some of the techniques are relevant. In particular, we would like to point out the work by Ehrig (Ehrig and Sure, 2004; Ehrig, 2006) who defines *ontology mapping* as a function on individual ontology entities based on a *similarity* function, that for a pair of entities from two ontologies computes a ratio indicating their semantic proximity. This ratio is further used to derive the *ontology similarity*, operationalized as a weighted aggregation function (Ehrig and Staab, 2004), combining individual similarity measures.

One inspiration for this work was also the well-known LOD cloud[1] (Cyganiak and Jentzsch, 2010).

3. The Component Metadata Infrastructure

Naturally the core of CMDI consists of components. These components group metadata elements and possibly other components. The reusable components are managed by the Component Registry (CR). To describe a resource types a metadata modeller combines existing and, when needed, new components from the CR into a metadata profile. Due to the flexibility of this model the metadata structures can be very specific to an organization, project or resource type. Although structures can thus vary considerably they are still within the domain of metadata for linguistic resources and thus share many key semantics. To deal with the variety general CMDI tools, e.g., the Virtual Language Observatory[2] which is a facetted browser/search for CMD records, operate on a shared semantics layer. To establish these shared semantics CMD components, elements and values can be linked to so-called data categories (DC) defined in separate concept registries. The major concept registries currently in use by CMDI are the Dublin Core metadata elements and terms (Powell et al., 2005) and the ISOcat Data Category Registry (DCR) (Windhouwer and Wright, 2012).

[1] http://lod-cloud.net/
[2] http://www.clarin.eu/vlo/

Table 1: The development of defined profiles and DCs over time.

	2011-01	2012-06	2013-01	2013-06	2014-01
Profiles	40	53	87	124	158
Components	164	298	542	828	1110
Elements	511	893	1505	2399	3101
Distinct data categories	203	266	436	499	737
Ratio of elements without DCs	24,7%	17,6%	21,5%	26,5%	24,2%

While the Dublin Core set of elements and terms is closed the ISOcat DCR is an open registry, which means that any metadata modeller can register the concepts it needs. Due to both the use of several concept registries and the open nature of some of these, multiple equivalent concepts can be created. CMDI uses the RELcat Relation Registry (RR) to create near sameness groups of these concepts.

4. Current status of the joint CMD Domain

In the following section, we give an overview of the current status in the CMD domain, both on the schema level, i.e. with regard to the defined profiles and data categories used, as well as on the instance level, the actual CMD records.

4.1. CMD Profiles

In the CR 153[3] public[4] Profiles and 859 Components are defined. Table 1 shows the development of the CR and DCR population over time.

Next to the 'native' CMD profiles a number of profiles have been created that implement existing metadata formats, like OLAC/DCMI-terms, TEI Header or the META-SHARE schema. The resulting profiles proof the flexibility/expressivity of the CMD metamodel. The individual profiles differ also very much in their structure – next to flat profiles with just one level of components or elements with 5 to 20 fields (*dublincore*, *collection*, the set of *Bamdes*-profiles) there are complex profiles with up to 10 levels (*ExperimentProfile*, profiles for describing Web Services) and a few hundred elements, e.g., the maximum schema from the META-SHARE project (Gavrilidou et al., 2012) for describing corpora has 117 components and 337 elements.

4.2. Instance Data

The main CLARIN OAI-PMH harvester[5] collects records from 57 providers on a daily basis. The complete dataset amounts to around 600,000 records. 20 of the providers offer CMDI records, the other 37 provide OLAC/DC records, that are being converted into the corresponding CMD profile after harvesting, amounting to round 44.000 records. On the other hand, some of the comparatively few providers of 'native' CMD records expose multiple profiles (e.g. Meertens Institute uses 12 different profiles). So we encounter both situations: one profile being used by many providers and one provider using many profiles.

We can also observe a large disparity on the amount of records between individual providers and profiles. Almost 250,000 records are provided by the Meertens Institute (*Liederenbank* and *Soundbites* collections), another 25% by MPI for Psycholinguistics (*corpus* + *Session* records from the *The Language Archive*). On the other hand there are 25 profiles that have less than 10 instances. This can be owing both to the state of the respective project (resources and records still being prepared) and the modelled granularity level (collection vs. individual resource). There is ongoing work to make the various granularity levels more explicit.

5. CMD cloud

As the data set keeps growing both in numbers and in complexity, there is a rising need for advanced ways to explore it. In this work, we present a method to analyze and visualize the relations among defined CMD profiles, with the *schema matching* – in particular, the mapping and similarity function proposed by (Ehrig and Sure, 2004; Ehrig, 2006) – serving as methodical basis.

5.1. SMC browser

The technological base for the presented method is the *SMC browser*[6], a web application being developed by the CMDI team, that lets the metadata modeller explore the information about profiles, components, elements and the usage of DCs as an interactive graph. This allows for example to examine the reuse of components or DCs in different profiles. The graph is accompanied by statistical information about individual 'nodes', e.g., counting how many elements a profiles contains, or in how many profiles a DC is used.

5.2. Basic approach

The basic idea for constructing the CMD cloud is to 1) collect the size of each profile (as the number of components and elements, or number of distinct data categories used); 2) compute the pairwise similarity ratio between the profiles based on some similarity measure; 3) generate a graph with profiles as nodes and the pairwise similarity relation expressed as weighted edges between them. When rendered, the size of the nodes in the graph reflects the size of the profile as computed before. The absolute number of matching identities is expressed as edge weight and the similarity ratio as *link strength* (inversely proportional

[3] All numbers are as of 2014-03 if not stated otherwise

[4] Users of the CR create components and profiles in their private workspace, and they can make them public when the components or profiles are ready for production.

[5] http://catalog.clarin.eu/oai-harvester/

[6] http://clarin.oeaw.ac.at/smc-browser

to link distance), drawing more similar profiles nearer together. Additionally, a variable threshold governs the level of similarity to be rendered as link.

5.3. Similarity ratio

At the core of the discussed method is the concept of similarity between entities and the challenge how to operationalize it. In the initial step, the similarity ratio is based on the most reliable information, the reuse of data categories, computed as the average of the quotients of matching distinct data categories for each of the two profiles.

$$
\begin{aligned}
sim_{p1} &:= \frac{count(distinct(Datcats_{match}))}{count(distinct(Datcats_{p1}))} \\
sim_{p2} &:= \frac{count(distinct(Datcats_{match}))}{count(distinct(Datcats_{p2}))} \\
sim &:= \frac{(sim_{p1} + sim_{p2})}{2}
\end{aligned}
\tag{1}
$$

Note though, that there is a number of other features and formulas that can be used to assess the similarity of two schemas (structures) (cf. 5.6.).

5.4. Results

The basic result is the graph of profiles with links based on their similarity. There are various ways to render this information. As SMC browser allows to select different subgraphs and adapt layout options, figure 1 depicts just one possible visual output of the analysis. This view shows nicely the clusters of strongly related profiles in contrast to the greater distances between more loosely connected profiles. SMC Browser also features alternative more detailed views that allow to detect visually which components and data categories are shared by which profiles. In a way a zoom in on the links between the nodes in the CMD cloud. The generated graph manifests a very high degree of interconnectedness in the generated graph (There are 7.835 links between the 157 profiles. A fully connected graph would have 12.403 edges.) resulting from the fact, that every profile shares at least one or two data categories with many other profiles. However, besides making the rendered graph illegible and difficult to lay out, such a result is also not a good answer to the question of similarity. Therefore a threshold was introduced to only consider links above a certain similarity ratio.

5.5. Applications

The SMC Browser and CMD cloud were developed primarily for assisting the task of metadata modelling. A modeller can get a quick overview of the existing profiles, their structure and their interrelations, allowing her to choose the most suitable one for describing the resources at hand.

When enriched with statistical information about instance data it can also serve as an alternative advanced interface for exploring the joint CLARIN metadata domain. It will offer the much needed 'big picture' for this huge heterogeneous collection of resources, an intuitively comprehensible visualization of its complex interrelations. This makes the tool also applicable for the metadata curation task, allowing to easily recognize structures and values that are being reused often ('hot spots') in contrast to outliers ('weak links'). With appropriate linking established the user can get from the structural overview (graph) directly to the corresponding records.

5.6. Planned extensions

There are a number of further factors, that can be taken into account, when computing the profiles similarity. The obvious next step is to consider the component reuse. Applying the relations between data categories as defined in Relation Registry would further raise the similarity ratios. Also, we need to cater for profiles with little data categories coverage. This can be resolved by including the data-category-coverage-ratio into the calculation.

We also plan to adopt more sophisticated approaches to compute entity and aggregated schema similarity as proposed in (Ehrig and Staab, 2004; Ehrig, 2006), like string or structural similarity between 'nodes'.

A very important planned addition opening a whole new field of applications is to integrate statistical information about instance data into the generation of the graph. In the 'instance'-mode node size would represent the number of instances for given profile and edge width the amount of data in the shared data categories. On instance level, also the ratio of shared values between fields/elements could be computed and used as another similarity indicator (though computationally very demanding).

6. Conclusions

In this paper, we gave a short overview of the current status of the CMD data domain as basis for the main contribution: an analysis of the semantic similarity between the profiles. This work offering a bird's eye view on the CMD data domain can serve as alternative starting point for exploring the dataset and provides valuable input for metadata modellers and the metadata curation task.

7. References

Amrouch, S. and Mostefai, S. (2012). Survey on the literature of ontology mapping, alignment and merging. In *Information Technology and e-Services (ICITeS), 2012 International Conference on*, pages 1–5. IEEE.

Broeder, D., Kemps-Snijders, M., et al. (2010). A data category registry- and component-based metadata framework. In Calzolari, N., Choukri, K., et al., editors, *LREC*, Valletta, May. ELRA.

Cyganiak, R. and Jentzsch, A. (2010). The linking open data cloud diagram. online, 09.

Ehrig, M. and Staab, S. (2004). Qom–quick ontology mapping. In *The Semantic Web–ISWC 2004*, pages 683–697. Springer.

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

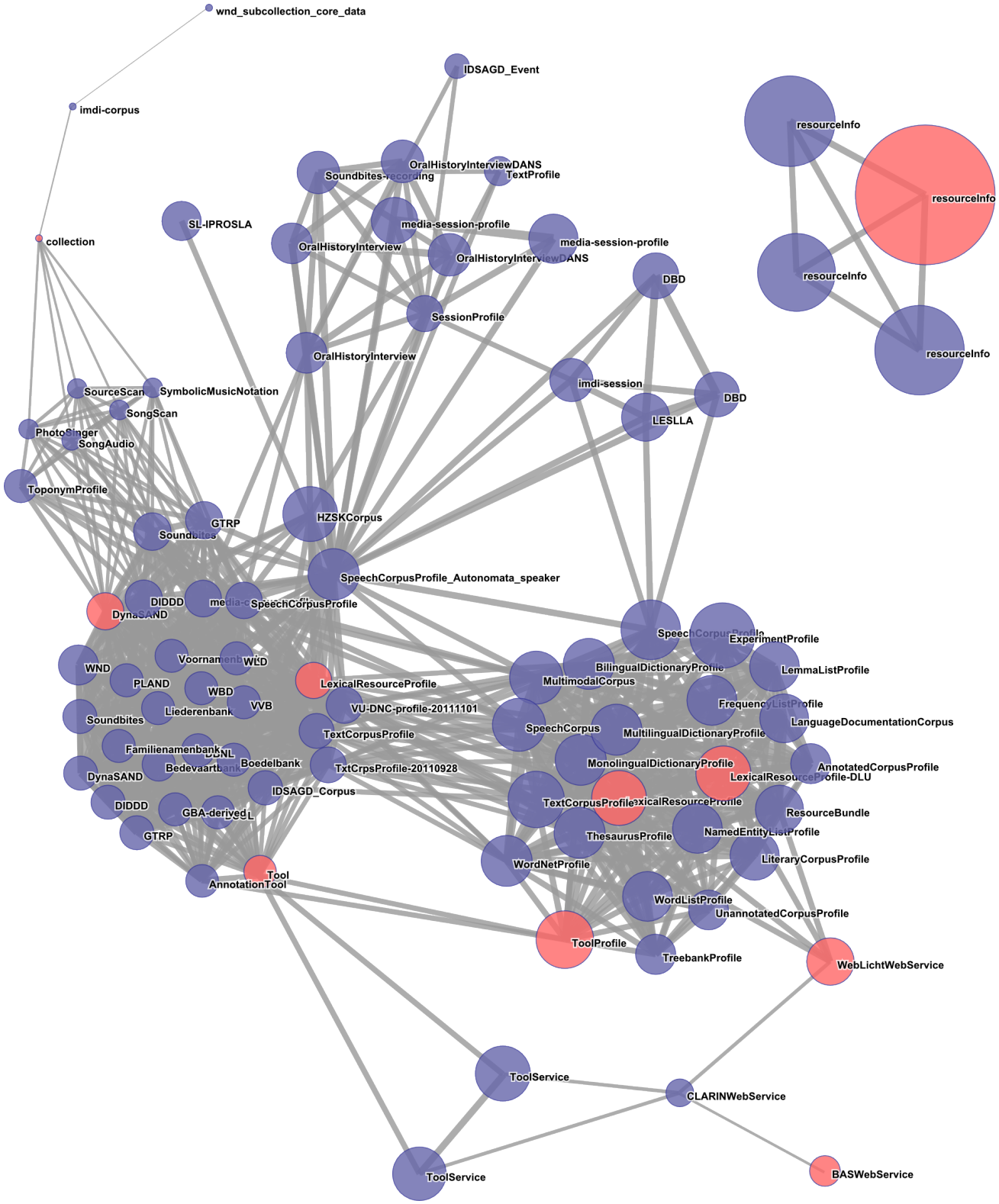

Figure 1: A graph view of the similarity relations between CMD profiles (*threshold=0.6*)

Ehrig, M. (2006). *Ontology Alignment: Bridging the Semantic Gap (Semantic Web and Beyond)*. Springer. PhD thesis.

Gavrilidou, M., Labropoulou, P., et al. (2012). The META-SHARE metadata schema for the description of language resources. In Calzolari, N., Choukri, K., et al., editors, *LREC*, Istanbul, May. ELRA.

Kalfoglou, Y. and Schorlemmer, M. (2003). Ontology mapping: the state of the art. *The Knowledge Engineering Review*, 18(1):1–31, January.

Noy, N. and Stuckenschmidt, H. (2005). Ontology alignment: An annotated bibliography. In *Semantic Interop-*

erability and Integration - Schloss Dagstuhl.

Powell, A., Nilsson, M., Naeve, A., and Johnston, P. (2005). DCMI Abstract Model. Technical report, March.

Shvaiko, P. and Euzenat, J. (2012). Ontology matching: state of the art and future challenges.

Windhouwer, M. and Wright, S. E. (2012). Linking to linguistic data categories in isocat. In *Linked Data in Linguistics*, pages 99–107, Frankfurt, Germany, March. Springer.