Generic versus Specific Concepts: Experiments around the CCR and the VLO

Menzo Windhouwer
HuC Digital Infrastructure
menzo.windhouwer@di.huc.knaw.nl

Lars Ahrenberg
Linköping University
lars.ahrenberg@liu.se

Susanne Haaf
BBAW
haaf@bbaw.de

Penny Labropoulou
ILSP
penny@ilsp.gr

Oddrun Ohren
National Library of Norway
oddrun.ohren@nb.no

Ineke Schuurman
KU Leuven
ineke.schuurman@ccl.kuleuven.be

1 Introduction

The CLARIN Concept Registry (CCR; Schuurman et al (2016)) is a repository of concepts aiming to enable semantic interoperability in the CLARIN infrastructure. In the metadata context, they are meant to be linked to elements and/or components of metadata profiles. As such, a concept linked to some element indicates its meaning and also that it is semantically close to other elements linked to the same concept. For this to work, it is evident that the level of specificity of the concepts must be taken into serious consideration. This paper explores two different approaches to this, as will be described in the following chapters.

2 Problem description

In the CCR metadata domain, all concepts should – one way or another – represent entities or properties related to some resource. However, these concepts are not necessarily atomic, nor independent of each other. On the contrary, they are often semantically composite and related in various ways. Two typical cases of composite concepts are:

1. The concept essentially relates the described resource to a separate entity with properties of its own. For example, a resource might be related to a *country*, and the country has both a *name* and an *identifier*. In a Component Metadata (CMD; Broeder et al (2012)) profile this may either be represented as a flat list of semantically composite, but internally independent elements *Country name* and *Country identifier*, or by a component *Country* containing the elements *Name* and *Identifier*. Both these representations are semantically equivalent, and it is important that the concepts in CCR allow for expressing that.
2. The concept relates to some part or aspect of the described resource, rather than of the resource as such. For example, the *country* of the *source material* of the resource indicates that the resource has source material, which in turn is related to *country* (this time disregarding the possible inner structure of *country*). In CMD this may be represented in a variety of ways, e.g. by the semantically complex but technically simple element *Country of the source material* (of the resource), by a CMD component *Source material* containing *Country* as an element or component, or by a CMD component *Associated countries* containing an element or component *Source material*. Again, we want the concepts in CCR to allow for expressing the equivalence of these representations.

The diversity of CMD structures and concepts already created for them lead to a recurring discussion among the CCR coordinators: Should we recommend metadata modellers to use in their profiles and components concepts as generic as possible (but as specific as necessary), or rather propose a move to a set of even more generic building block style concepts? To put an end to this discussion it was decided to run an experiment in the context of the Virtual Language Observatory (VLO; Van Uytvanck et al (2010)) facets. The population of these facets from a CMD record would benefit from well matching and recommended concepts in the CCR.

For the experiment, the CCR coordinators split up in two teams to follow two approaches. In the first approach ("team specific"), one creates concepts for the VLO facets that will be associated with elements in a CMD profile – this reflects the current most common approach. In the second approach ("team generic"), one creates concepts that could be associated with elements and components, and a combination of those would map to a specific VLO facet. Five facets were selected and their tooltips used as the primary description of their intent. In this paper, the *Country* facet (VLO tooltip: "the country of origin of the source material of the resource") will be used as the running example. Before describing the experiments done by both teams, the process currently used by the VLO to populate its facets with values from a CMD record will be described, as it shows how the VLO and CCR currently cooperate.

3 How the VLO and the CCR cooperate

For each facet, the VLO maintains a list of CCR concepts (considered semantically close) to look for in a CMD profile.[1] For example, the concepts for the *Country* VLO facet are

1. *location country* (CCR_C-2532_d004b0a6-fd1d-3ca3-abf1-1e6aeb3e37b2) defined as "The country where the resource was created or originated"
2. *country name* (CCR_C-3792_68c770a4-d58c-46dd-d429-5609ce5f81c3) defined as "Indication of the name of a country."
3. *country coding* (CCR_C-2092_36cd7ca8-e412-9f29-7ea7-4a3ba4ba2c91) defined as "Designation of the standard used to code the country."

When the VLO importer, the tool that imports CMD records into the VLO, encounters a CMD profile it searches for elements that refer to these concepts via their concept links. If it finds one, it determines the path to this element, i.e., the XPath consisting of all the components one has to visit to reach this element. A symmetric semantic path, also known as the semantic context, can also be created: it consists of the concept links of these components. Although not often used, the direct context, i.e., the first concept link encountered when following the path from the selected element via its superordinate components up to the root component, can be marked as acceptable or unacceptable. If the direct context is unacceptable, the path is disregarded. A path might be unacceptable if it points to a property of some aspect of the resource instead of the resource itself, e.g., the mother language of a speaker instead of the language used in the audio recording. An acceptable path can be the basis for retrieving element contents suitable for the VLO facet under consideration from a specific record. For example, for the *media-session-profile* profile (clarin.eu:cr1:p_1336550377513) three paths are found containing *Country* information:

1. /media-session-profile/media-session/media-session-actors/media-session-actor/BirthCountry/Country/Code
2. /media-session-profile/media-session/Location/Country/Code
3. /media-session-profile/media-session/media-annotation-bundle/media-file/Location/Country/Code

At the technical level these paths are basically XPaths, which can be directly resolved in a specific CMD record in order to populate the values of the *Country* facet.

4 Experiments with Specific Concepts

This experiment was conducted as specified in the original CCR manual (2016), i.e., the meaning of a concept is described in a text definition which must be *reusable*,[2] *unambiguous* and *concise*. For our experiments, this meant that concepts used in a definition that are relevant in the context of CLARIN, will get definitions of their own when necessary, while links are provided to those definitions that are already available.

[1] This list can be inspected at cmdi.clarin.eu/mapping/, and also allows to see acceptable paths found for a profile.

[2] Definitions being as generic as possible, while as specific as necessary

The tooltip serves as point of departure for the definitions to be provided, as team specific could not use CMD:

country: The country of origin of the source material of the resource

This contains links to four more generic concepts, which are defined in turn:

1. *country*: a current or former (independent) national state, city state, country, …
2. *origin*: the (geographical) location where a tool or resource is constructed
3. *source material*: the content (written, spoken, ...) to be researched
4. *resource*: entity containing material to be researched (e.g. corpus, etc.)

In this case the VLO concept *country* comes along with a more generic definition of *country* as well, as the VLO reading is a very specific one.

5 Experiments with Generic Concepts

Team Generic started with some experiments around the VLO mapping process to see if generic concepts would work in, and hopefully improve, the mapping process. As not many CMD profiles contain concept links for components this was remedied by some manual workarounds, e.g., interpreting component names and manually mapping them to candidate generic concepts.

5.1 Using Generic Concepts to Specify the Semantic Context

The key concepts, i.e., *resource*, *source*, *origin*, and *country*, from the *Country* VLO facet tooltip were put in a path that describes the semantic context, i.e., *resource* // *source* // *origin* // *country* (where // denotes that there might be intermediate concepts). Next, this semantic path was manually matched with the paths found by the VLO importer (see the end of section 3). This turned out to be hard; in many cases no component to which the generic concepts like *resource*, *source* and *origin* could be attached, existed. Hence, the semantic path was adapted to provide a better match with the existing components, e.g., leaving out *resource* as we can assume all metadata describe resources. The same can be done with *source* and *origin*, assuming they can be implicit. This leaves the most minimal path, i.e., only the *country* concept to match (// *country*). However, this will also accept paths that are not semantically suitable, e.g., an actor's country of birth.

5.2 Using Generic Concepts to White- or Blacklist Semantic Contexts

In this experiment the idea of black- and white-listing, as partially already implemented in the VLO importer (see section 3), is used to deal with the problem identified in section 5.1. Again, the paths used by the VLO importer are the starting point. For now, the salient components are grouped in a white- or blacklist. For the *country* concept this results in the following pseudo code for rejecting or accepting a path:

```
if      context is empty
then    ACCEPT
elif    context in (OriginLocation, resourceCreationInfo, GeneralInfo, Creation)
then    ACCEPT
elif    context in (ccr:fccc56dde24d, media-file, mediaFile, fileDesc, Project, ExperimentContext,
        personInfo, organizationInfo, Author_DiscAn, publisher, PersonalBackground,
        ProfessionalBackground)
then    REJECT
else    ACCEPT
```

When evaluating this for the *media-session-profile* paths the following results are achieved:

a. REJECT: /media-session-profile/media-session/media-session-actors/media-session-actor/BirthCountry/Country/Code
b. ACCEPT: /media-session-profile/media-session/Location/Country/Code

c. REJECT: /media-session-profile/media-session/media-annotation-bundle/media-file/Location/ Country/Code
The above means that the paths to a birth country or the storage location of a resource are rejected. The final step is to move from the components (e.g., resourceCreationInfo, GeneralInfo) to the generic concepts, e.g., *country*, *origin*, *creation*, *general*, to which they could be linked.

5.3 Defining Generic Concepts

The resulting set of generic concepts will have to be defined with generic statements in the CCR. For instance, some of the generic concept definitions might be:

- *country*: a region constituting an independent state, nation, province, etc., which was or is independent or distinct from others in terms of institutions, language, etc. (definition based on: oed.com/view/Entry/43085)
- *creation*: the action or process of bringing something into existence. (source: en.oxforddictionaries.com/definition/creation)

Notice the alternative definition for the generic *country* concept (see section 4 for the definition of the specific approach). These are all working definitions, i.e., still under discussion and will be aligned and voted upon before inclusion in the recommended concept set in the CCR.

6 Conclusions and Future Work

The experiments, as described in this paper, showed that the specific and generic approach are actually coming nicely together, i.e., they both identified an almost identical set of generic concepts to define. They differ in where these generic concepts are linked: either in a concept definition, or in the metadata profile. Also, the generic approach spares the need to define the specific concepts. Unfortunately, due to the need of blacklisting certain contexts, the number of generic concepts needed is greater than initially expected. Still the impression is that the generic approach will be more flexible and will in the end need a smaller number of recommended (generic) concepts. These can be combined into many different semantic contexts in the CMD profiles, as opposed to representing each semantic context in the CCR itself. To come to full power, the generic approach will require some extensions in the CMD Infrastructure: The VLO importer will have to be modified to use the whole semantic context instead of the direct context only. Also, the CMD will need to be extended to allow for multiple concept links to be added to only one component, element, etc., so a composite element CountryName, can refer to the same generic concepts as the equivalent structure: Country/Name.
Finally, it was noted that the VLO facet tooltips include ambiguities that make their interpretation hard. Thus, the CCR coordinators will work on some suggestions to improve them and share those improved tooltips with the VLO development team and the Metadata Curation taskforce.

References

D. Broeder, M. Windhouwer, D. van Uytvanck, T. Goosen, T. Trippel. CMDI: a Component Metadata Infrastructure. In the *Proceedings of the Metadata 2012 Workshop on Describing Language Resources with Metadata: Towards Flexibility and Interoperability in the Documentation of Language Resources*. 2012.

I. Schuurman, M. Windhouwer. The CLARIN Concept Registry (manual for editors): Principles behind the CCR. June, 2016. (draft)

I. Schuurman, M. Windhouwer, O. Ohren, D. Zeman. CLARIN Concept Registry: The New Semantic Registry. In *Selected Papers from the CLARIN 2015 Conference*, Linköping Electronic Conference Proceedings. 2016.

D. van Uytvanck, C. Zinn, D. Broeder, P. Wittenburg, M. Gardellini. Virtual Language Observatory: The Portal to the Language Resources and Technology Universe. In the *Proceedings of the International Conference on Language Resources and Evaluation* (LREC). 2010.