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

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

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

1 Introduction

In the CLARIN Concept Registry (CCR; Schuurman et al (2016)) metadata domain, all concepts should - one way or another - represent properties of some resource. However, the expressed properties are not necessarily atomic, nor independent of each other. On the contrary, they are often semantically composites and linked in various ways. Two typical cases of composite properties are:

1. When the indicated property 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*. As resource descriptions, these two representations are semantically equivalent, and it is important that the concepts in CCR allows for expressing that.
2. When the indicated property is a property of 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 has property *country* (this time disregarding the possible inner structure of country). In CMD this may be represented in a variety of ways and degree of structuring, 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.

This diversity of CMD structures and concepts already created for them lead to a recurring discussion among the CCR coordinators: Should we recommend concepts of arbitrary specificity, or strive towards a recommended set of generic, building block style concepts? To put an end to this discussion it was decided to run an experiment in the context of the VLO 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 created two teams: Team Generic and Team Specific. Team Specific would create concepts for the VLO facets that could be associated with the elements, which is the current common approach. Team Generic would create concepts that could be associated with elements and components, where a combination of those would map to a specific VLO facet. Five VLO facets were selected and their tooltips are used as the primary description of their intent.

In this paper the Country facet ("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 together.

2 How the VLO and the CCR cooperate together

For each facet the VLO maintains a list of CCR concepts to look for in a CMD profile.[2] For example, the concepts for the Country VLO facet are

1. location country (CCR_C-2532_d004b0a6-fd1d-3ca3-abf1-1e6aeb3e37b2)
2. country name (CCR_C-3792_68c770a4-d58c-46dd-d429-5609ce5f81c3)
3. country coding (CCR_C-2092_36cd7ca8-e412-9f29-7ea7-4a3ba4ba2c91)

When the VLO importer 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., this path consists of all the components one has to visit to reach this element. An symmetric semantic path, also known as the semantic context, can also be created: it consists of the concept links of these components. Although this functionality is not used often,[3] the direct context, i.e., the first concept link encountered to when visiting the components from the element on its way to the root component, can be marked as acceptable or unacceptable. If the direct context is unacceptable the path is disregarded. An acceptable path can be used to find the element's value in a specific record, and this value is than the VLO facet value for that record.[4] For example, for the *media-session-profile* profile (clarin.eu:cr1:p_1336550377513) 3 paths are found:[5]

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

On the technical level these paths are basically XPaths, which can be directly resolved in a specific CMD record resulting in the values for the Country facet.

3 Experiments with Specific Concepts

Team Specific works as defined in the original CCR manual (2016), i.e., the meaning of a concept is described in text, while the definition should be *reusable* and *unambiguous*. The definition should also be 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, while links are provided to those definitions.
For the running example this means a definition for the concept specifically related to the VLO facet Country as below

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

In this definition 4 more generic concepts appear, which are to be defined in turn:

- country: a current or former national state
- origin: the (geographical) location where a tool or resource is constructed
- source material: the content (written, spoken, ...) to be researched
- resource: entity containing material to be researched (corpus, etc)

---

[2] This list can be inspected here: cmdi.clarin.eu/mapping/
[3] Currently this functionality is mainly used to disregard the languages an actor speaks to be taken as the language the resource is about, e.g., to prevent an Aweti resource to be associated with Latin, a language the informant also learned at school.
[4] The page mentioned in footnote 2 can also be used to inspect which acceptable paths are found for a specific CMDI profile.
[5] As of December 2017.

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.

4 Experiments with Generic Concepts

Team Generic started with some experiments around the VLO mapping process to see if generic concepts would function 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 map them to candidate generic concepts.

4.1 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 2). This turned out to be hard. There is often not a component to which the generic resource concept can naturally be attached, and the same can be true for source or origin. The semantic path was adapted to match better, e.g., leaving out resource as we can assume all metadata describes resources. Ultimately this leaves only the country concept to match (//country), but this will also accept semantically not well matching paths, e.g., an actor’s country of birth.

4.2 Using Generic Concepts to White- or Blacklist Semantic Contexts

In this experiment the idea of black- and whitelisting, as partially already implemented in the VLO importer (see section 2), is used to deal with the problems found in section 4.1. Again the paths used by the VLO importer are the starting point. For now the salient components are grouped in a white- or blacklist. This results in the following pseudo code for a rejecting or accepting a path:

if context 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 final step is to move from the component names to the generic concepts, e.g., country, origin, creation, general, which should become the concept links for these components.

4.3 Defining Generic Concepts

TODO

5 Conclusions and Future Work

TODO
The VLO facet tooltips are hard to interpret, the CCR coordinators will work on suggestions to improve them.
The specific and generic approach actually meet very closely regarding the actual generic concepts identified. However, in the generic approach we prevent the need to define the specific concept. Unfortunately, due to the need of blacklisting certain contexts some more concepts are needed. Still the impression is that the generic approach will be more flexible and needs a smaller number of recommended concepts.
The generic approach will require some extensions. In the VLO importer the use of the whole semantic context and not only the direct context. And CMD will need to be extended to allow multiple concept links, so a composite element, e.g., CountryName, can refer to the same generic concepts as a finer grained, but equivalent, structure, e.g., Country / Name.

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*. At LREC 2012, Istanbul, Turkey, May 22, 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 K. De Smedt (ed.), *Selected Papers from the CLARIN 2015 Conference*, Linköping Electronic Conference Proceedings, April, 2016.