CLARIN Concept Registry: The New Semantic Registry

Ineke Schuurman
KU Leuven, Belgium
and Utrecht University, The Netherlands
ineke@ccl.kuleuven.be

Menzo Windhouwer
Meertens Institute
Amsterdam, The Netherlands
menzo.windhouwer@meertens.knaw.nl

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

Daniel Zeman
Faculty of Mathematics and Physics
Charles University in Prague
Czech Republic
zeman@ufal.mff.cuni.cz

Abstract

The CLARIN Concept Registry (clarin.eu/conceptregistry) is the place in the CLARIN Infrastructure where common and shared semantics of, but not limited to, linguistic concepts are defined. This is important to achieve semantic interoperability, and to overcome to a degree the diversity in data structures, either in metadata or linguistic resources, encountered within the infrastructure. Whereas in the past, CLARIN has been using the ISOcat registry for these purposes, nowadays this new registry is being used, as ISOcat turned out to have some serious drawbacks as far as its use in the CLARIN community is concerned. The main difference between the two semantic registries is that the CCR is a concept registry whereas ISOcat is a data category registry. In this paper we describe why the decision to switch to a concept registry has been made. We also describe the most important other characteristics of the new (Open)SKOS-based registry, as well as the management procedures used to prevent a recurrent proliferation of entries, as was the case with ISOcat.

1 Introduction

One of the foundations of the CLARIN Component Metadata Infrastructure (CMDI; Broeder et al. 2012; clarin.eu/cmdi) is a semantic layer (Durco and Windhouwer, 2013) formed by references from CMDI components or elements to entries in various semantic registries. Popular have been references to the metadata terms provided by the Dublin Core Metadata Initiative (DCMI; dublincore.org) and the data categories provided by ISO Technical Committee 37's Data Category Registry (DCR; ISO 12620, 2009) ISOcat (isocat.org). For describing more content-related data, like describing the components of a morpho-syntactic tagset, ISOcat was also used. Using entries either based on (ISO) standards, *de facto* standards (for example generally used in the Netherlands or, broader, in the Dutch-speaking regions) or even made by users themselves (esp. in legacy data).

Although using ISOcat has been encouraged by CLARIN, it has certain drawbacks. As pointed out by Broeder et al. (2014) and Wright et al. (2014), ISOcat, with its rich data model combined with a very open update strategy, has proved too demanding, at least for use in the CLARIN context. Among other things, confusion on how to judge whether a candidate ISOcat entry adequately represents the semantics of some CMDI component or element, has led to proliferation far beyond the real need. This resulted in a semantic layer of questionable quality. Therefore, when ISOcat, due to strategic choices made by its Registration Authority, had to be migrated and became, for the time being, static, CLARIN decided to

look for other solutions to satisfy the needs of the infrastructure. As a result the Meertens Institute is now hosting and maintaining the CLARIN Concept Registry (CCR; `clarin.eu/conceptregistry`).

This paper motivates and describes this new semantic registry, its model, content and access regime, indicating the differences from ISOcat where appropriate. Proposed management procedures for the CCR are also outlined, although not in detail.

2 Issues in ISOcat

One of the principles behind ISOcat was that it was an open registry, so it was very easy to get a login and to get the rights to enter new data categories. Such entries could remain private, only you yourself could read and edit them, plus the persons you had given permission to do so. Or they could be made public, i.e. everybody can read them (but not edit). Still the content of the registry was out of control. People were, for example, urged not to provide entries that were more or less copies of already existing ones, but a) there was no way to prohibit it, and b) people sometimes copied an entry, just in order to make sure that the original owner would not change the entry without them knowing it. So the first issue to be addressed was:

- **Proliferation**, due to
  - access: too many people having write access;
  - quality: ignorance or negligence of ISOcat/CLARIN requirements (esp. with regard to definitions);
  - reliability: (public) entries being changed in a semantically meaningful way, i.e. changing their meaning.

The proliferation issue can, to a large extent, be solved by giving far less people permission to contribute new entries and to change existing ones in the new environment, i.e. the CCR should not be an open registry.

Another issue, already referred to in Section 1, concerns the complexity of a Data Category Registry like ISOcat. For example, in ISOcat different entries were needed for the data category /*genre*/, depending on how the needed value domain is defined, even when the definition as such would be the same (cf. Section 3). In other cases one had to specify in which format the 'data type' could be described, with over 40 options to select from. In most cases the default was chosen, as most users did not understand all options. Strictly speaking several parts of the DCR data model are not really needed for CLARIN-purposes, like the 'data type' mentioned. So the second issue to be addressed was:

- **Complexity**, due to
  - forced duplicates (data category type);
  - options only useful for experts in a specific field (data type);
  - obligatory parts of the data model not really necessary for CLARIN purposes.

These issues have been taken care of in the CCR (cf. the next sections). We will first discuss the issue last mentioned (Complexity), as the choice made strongly influences the design of the new registry: dealing with concepts instead of data categories. Avoidance of the causes of proliferation described under the first bullet is mainly related to procedural measures (cf. Section 7).

3 From data categories to concepts

As indicated above, the shift from ISOcat to a new format also involved changing the main entity of the registry. Instead of focusing on data categories as before, the new registry contains concepts. The transition is illustrated in Figure 1 depicting core features of the data categories and their abstraction into concepts.

A data category as modeled in ISOcat is an elementary descriptor in a linguistic structure or an annotation scheme (ISO 12620, 2009), implying it is a descriptor of something. In the CMDI universe this

something is most often a language resource or related objects. Moreover, there are several types of data categories, notably

- Data categories representing attributes or properties of something, in the sense that they are to be assigned values. In ISOcat these were called complex data categories. Their value sets may either be closed (controlled value vocabularies), open (to be specified freely by the user) or constrained (must follow specific rules). Their domains (the set of objects to which they may be applied) are not formally specified, but are often implied by the definition.

- Data categories representing atomic elements to be included in the value set of some complex data category. In ISOcat these were called simple data categories.

This means that a data category is defined not only by its meaning, but also by how it may be used. For example, consider a data category */genre/*, defined as a complex category (may be assigned a value) with value to be chosen from a finite set of genres. Any annotation/metadata scheme needing a */genre/* data category must agree to the same set of values. In cases where additional values or a completely different value set is called for, or where the user should be able to specify genre freely, we are in effect talking about different data categories, - the existing */genre/* cannot be reused. The same is true if genre is needed as a value of another complex data category. Consider for example the data category */subject type/* defined with value set including topic, genre, temporal coverage a.o. In this case yet another */genre/* data category will have to be defined, this time as a simple data category.

While such a model is very rich in expressive power, it is notoriously hard to maintain consistency and requires a high degree of understanding and alertness from users to be successful.

In the CCR we want to take advantage of the fact that groups of data categories, although applicable in different contexts have the same meaning in terms of a more or less similar definition. Following the genre example, the core idea of the concept *genre* is the same, whether it is to be used as an attribute with values (and irrespective of value set) or itself as a value of some other attribute, e.g. subject type. By disregarding information on application domain and value range and focusing only on definition and conceptual relations, the registry should be leaner and easier to maintain. On the other hand, the resulting semantic layer spanning the collective set of CMDI records inevitably will be coarser and thereby less informative.

4 An OpenSKOS registry

In CLARIN-NL the Meertens Institute had already developed (and continues hosting) the CLAVAS vocabulary service (`openskos.meertens.knaw.nl`) based on the open source OpenSKOS software package (Brugman and Lindeman, 2012; `openskos.org`), which was originally created in the Dutch CATCHPlus project. The OpenSKOS software provides an API to access, create and share thesauri and/or vocabularies, and also provides a web-based editor for most of these tasks. The software is used by various Dutch cultural heritage institutes. The Meertens Institute joined them to collectively maintain and further develop the software.

Based on the experiences with ISOcat OpenSKOS was evaluated to see if it would meet the needs of the CLARIN community and infrastructure. The major aim was to improve the quality of the concepts by having a) a much simpler data model and b) a less open, but also less complicated, procedure for adding new concepts or changing existing ones and recommending them to the community. In addition, certain technological requirements of the CLARIN infrastructure had to be met. Based on this evaluation the Meertens Institute extended the OpenSKOS software in various ways:

- Concepts in the CCR get a handle as their Persistent IDentifier (PID);

- The CCR can easily be accessed by the CLARIN community via a faceted browser (cf. Figures 2 and 3; `clarin.eu/conceptregistry`);

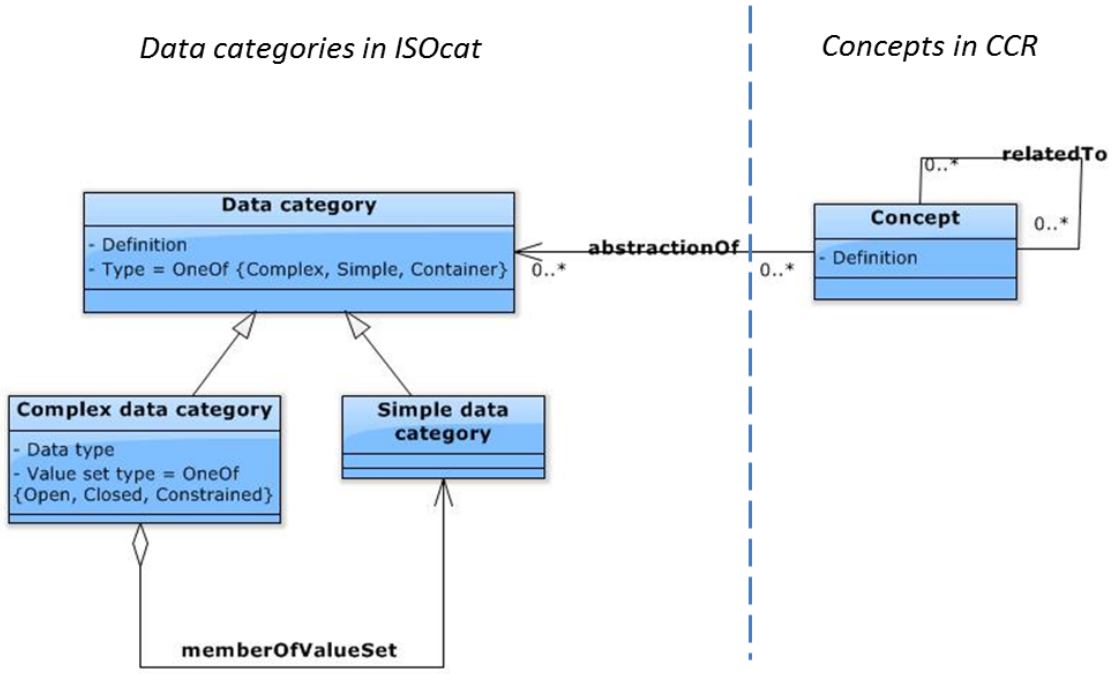

Figure 1: Core features of data categories and their abstraction into concepts

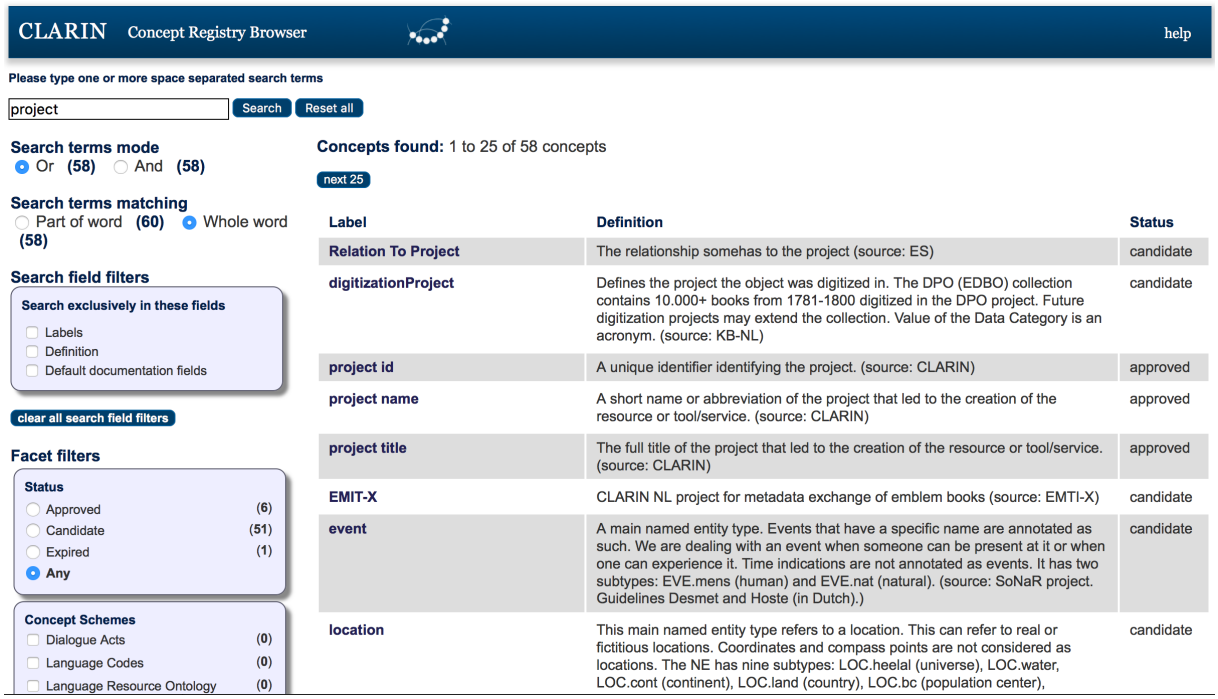

Figure 2: The CCR faceted browser

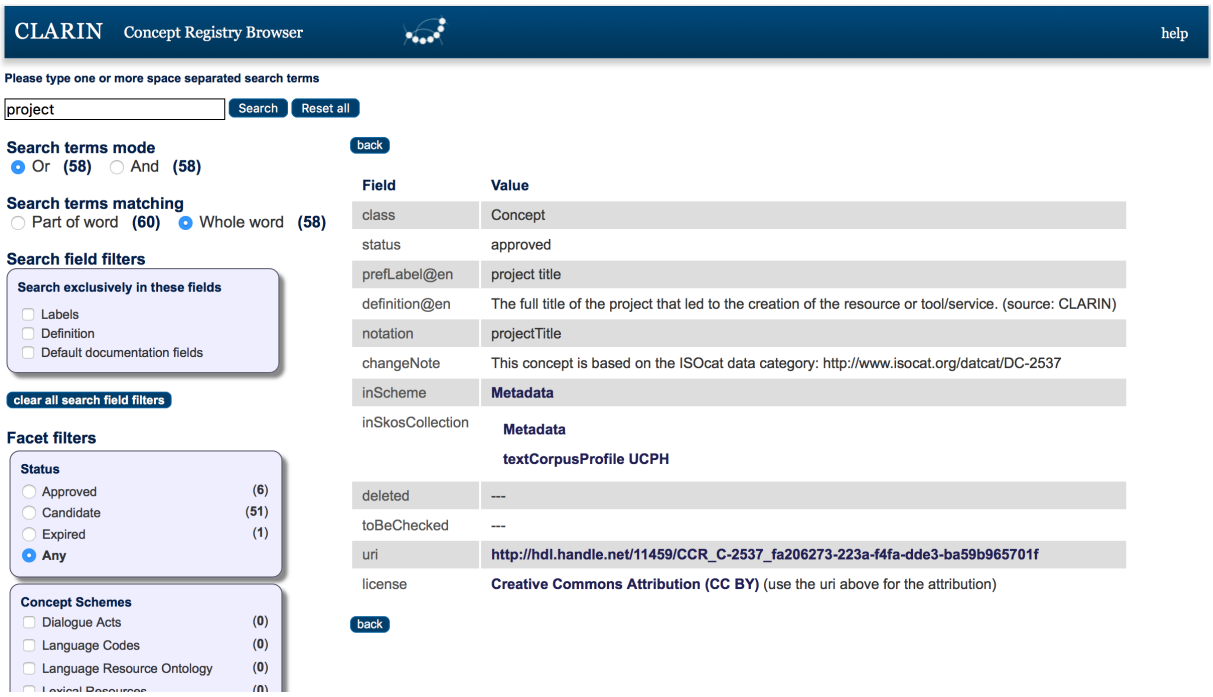

Figure 3: The CCR faceted browser - concept view

- Support for SKOS collections;
- Shibboleth-based access to the CCR.

Currently these extensions reside in a private Meertens Institute source code repository, but as part of the CLARIN-PLUS project (`clarin.eu/content/factsheet-clarin-plus`) these extensions (and more) will be integrated with the next version of OpenSKOS now under development.

5 Representing CCR concepts in the SKOS model

The data model supported by OpenSKOS is a substantial part of the Simple Knowledge Organization Scheme (SKOS) recommendation by W3C (`w3.org/skos`). SKOS is typically used to represent thesauri, taxonomies and other knowledge organization systems. At the Meertens Institute support for collections was added and currently Picturae, a Dutch service provider within the cultural heritage domain and the original developer of OpenSKOS, works on supporting the extended labels of SKOS-XL.

The work done by the CLARIN community in ISOcat was made available in the CCR by importing selected sets of data categories as new concepts (cf. Section 6). This made it possible to start a round of clean-up and creating a coherent set of recommended concepts (cf. Section 7). This import is not lossless as data category specific properties like the data category type and data type are lost. However, these properties have turned out to be one of the main causes of confusion and proliferation in the use of ISOcat (Broeder et al., 2014; Wright et al., 2014). In general SKOS appears to be a suitable model for the CCR. Each CCR concept may be assigned preferred labels (at most one per language,[1] alternative labels, definitions, examples and various kinds of notes. Moreover, the ISOcat thematic domains and data category selections could be maintained by importing them to SKOS concept schemes and collections, respectively. Only one import decision turned out to be problematic: converting the data category

[1]For the moment all entries are in English. Only when the entries have been approved by the national CCR coordinators other languages may be added. This is a lesson learned from ISOcat where translations often were not in sync.

identifier into a concept notation. SKOS notations are required to be unique within their concept scheme, whereas this constraint did not apply to data category identifiers in the DCR data model. A first clean-up round to remedy this has been finished successfully.

The SKOS model provides the possibility to express semantic relationships between concepts, e.g. broader than, narrower than and related to. In contrast, the DCR data model only contained relationships based on the data category types, e.g. a simple data category belonged to the value domain of one or more complex data categories. These domain-range relationships do not correspond well to any of the SKOS relationship types. Careful manual inspection would be needed to determine if any mapping can be made. Hence, for now these relationships have not been imported into the CCR. At a later date these facilities of the SKOS model and OpenSKOS can be exploited and could eventually take over the role originally envisioned for RELcat (Windhouwer, 2012). For now the initial focus is on the concepts themselves.

Neither SKOS itself nor OpenSKOS yet provides an extensive versioning model, i.e. concepts can be expired but there is no explicit link to a superseding concept. This is now on the wishlist for the next version of OpenSKOS as developed in CLARIN-PLUS.

Being RDF-based SKOS also brings the potential to more easily join forces with the linked data and semantic web communities. However, our current focus is on cleaning up the registry, thus gradually obtaining a coherent hub to be offered to the linked data cloud.

6 The CCR content

In the past few years, many national CLARIN teams made an effort to enter their data in ISOcat. This work has not been useless as all entries of relevance to a specific CLARIN group have been imported into the CCR. Leaving out redundant entries already means a considerable reduction in number of entries (from over 5000 in ISOcat (Broeder et al., 2014) to 3139 in CCR (June 2015). Although the imported concepts received new handles, care was taken to retain a link with their ISOcat origin. Automated mapping is thus possible and can be used to convert references to ISOcat data categories into references to CCR concepts. A mapping tool[2] for this has been developed and especially used for the existing CMDI components and profiles. But the tool is generic and can be used for other types of resources.

7 Maintaining the CCR: procedures and actors

Just like ISOcat the CCR can be browsed and searched by anyone, member of the CLARIN community or not, and anyone can refer to the concepts. However, contrary to ISOcat, only specifically appointed users, namely the national CCR content coordinators[3] are given rights to update the CCR (cf. Figure 4). These coordinators were appointed by their respective CLARIN national consortia when the problems with the usage of ISOcat became apparent. Their mission is to improve the quality of the data categories (now concepts) used within CLARIN. With the CCR in place the national CCR content coordinators have teamed up more actively and established procedures around the CCR to fulfill this mission.

To deal with the ISOcat legacy the coordinators are doing a round of clean-up with the aim to deprecate[4] low quality concepts and recommend high quality concepts. Notice that, just like in ISOcat, deprecated concepts remain accessible, i.e. their semantic descriptions are not lost, but their active usage is discouraged. The main focus is on providing good definitions. A good definition should be ”as general as possible, as specific as necessary” and should therefore be:

1. Unique, i.e. not a duplicate of another concept definition in the CCR;
2. Meaningful;
3. Reusable, i.e. refrain from mentioning specific languages, theories, annotation schemes or projects;
4. Concise, i.e. one or two lines should do;

[2] `github.com/TheLanguageArchive/ISOcat2CCR`

[3] Cf. `clarin.eu/content/concept-registry-coordinators`

[4] In the OpenSKOS status model deprecation means a concept gets the status expired.

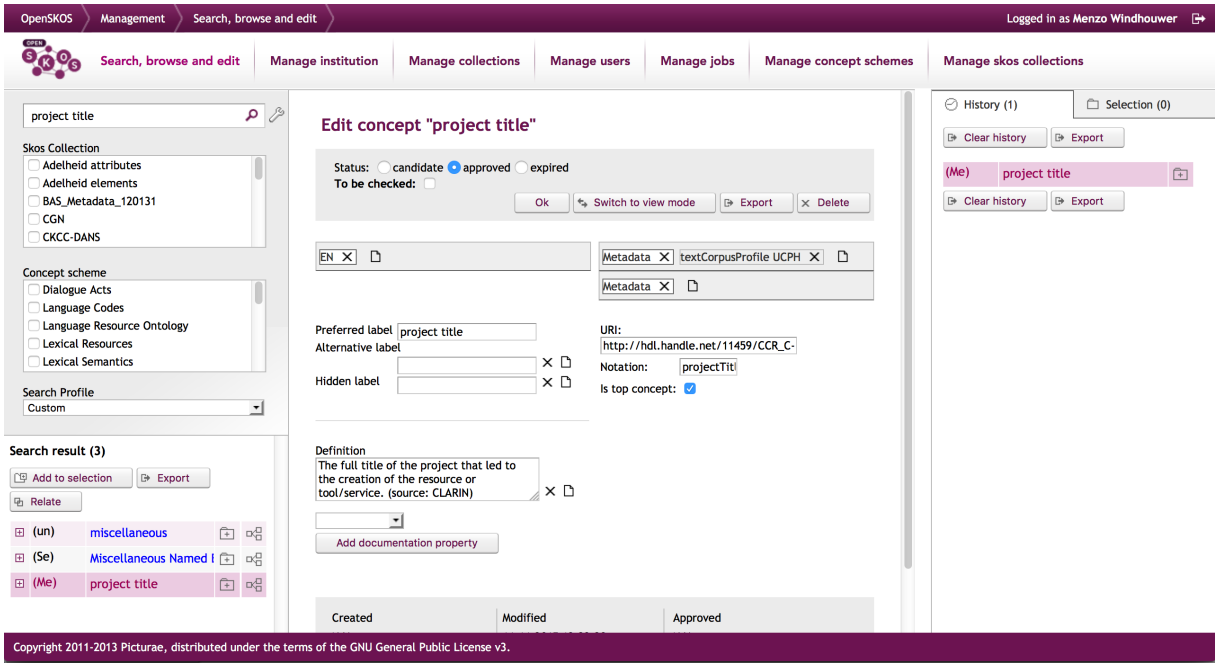

Figure 4: The CCR editor

5. Unambiguous.

As far as point 5 is concerned, a concept used in the entry of another concept should be referred to by its handle. Detailed guidelines are under development by the coordinators and will become generally available in due course. Apart from defining best practice for the coordinator group, such guidelines will benefit users directly, enabling them to issue informed requests to the CCR coordinators (see below).

The changes the coordinators can do to existing concepts are limited, i.e. they should not change the meaning. Only typos, unclear formulations, etc. can be remedied. Otherwise a new concept has to be created, and the original one may be deprecated.

All the coordinators or their deputies are involved in these changes. In cases where they do not agree a vote might take place and the change will be performed if 70% or more of the coordinators agree. A book keeping of the results of votes is maintained at the CCR section of the CLARIN intranet. The time frame within which the discussions and possibly a vote have to reach a decision is 2 weeks. In the holiday seasons and during the initial start-up phase a longer time period can be agreed upon by the coordinators.

Members of the CLARIN community wanting new concepts or changes to existing ones need to contact their national CCR content coordinator (`clarin.eu/content/concept-registry-coordinators`). Users from countries with no content coordinator should use the general CCR email address (ccr@clarin.eu) to file their requests. These requests will then be discussed within the national CCR content coordinators forum as described above. Note that in OpenSKOS any changes made to concepts are directly public. Therefore new entries or changes will only be entered after their content has been approved by the content coordinator forum. This procedure will take some time, but should result in a registry containing concepts with a better quality and with less proliferation. And therefore the CCR content should eventually deserve a higher level of trust as was the case for the ISOcat content. One can also expect that the need for new concepts will diminish over time due to the CCR covering more and more of the domains relevant to CLARIN.

8 Future work

In the Netherlands two other linguistic projects are also using concept registries. We want to investigate what the possibilities are with respect to interoperability, e.g. migrate concepts from these registries into the CCR. But keeping in mind that the maintenance of the collections of concepts should be kept strictly separate in order not to run into troubles à la ISOcat, i.e. it should remain clear which concepts are recommended by CLARIN.

Furthermore, as mentioned before the CLARIN-PLUS project, which started in the second half of 2015, aims to strengthen the CLARIN infrastructure on various fronts, including the CCR. Although the focus is mainly technical the improvements will also help the CCR coordinators with their task and improve the expressiveness of the CLARIN semantic interoperability layer. The CLARIN-PLUS CCR work is done by the Meertens Institute and is split into 4 phases (CE-2015-0688):

Phase 1 Monitor and test the stability of the new OpenSKOS version currently under construction by the OpenSKOS user community, especially Sound & Vision and Picturae;

Phase 2 Merge the currently existing various OpenSKOS forks into one trunk, so everyone in the user community can benefit from new features and stability improvements;

Phase 3 Implement features that fully support the concept life cycle, e.g. referring to a succeeding concept from a deprecated concept;

Phase 4 Implement features to support internal and external relationships of any type, i.e. not only SKOS relations and not only between concepts, and attribution thereof.

Currently, early 2016, phases 1 and 2 are ongoing and focus on strengthening the technical basis of OpenSKOS, and thus the CCR. The upcoming phases 3 and 4 will add new functionality needed by the CLARIN, and OpenSKOS, community.

9 Conclusions

Although CLARIN just started working on the new OpenSKOS-based CLARIN Concept Registry and there is still a lot of ISOcat legacy to deal with, the new registry looks promising. Our feeling is that it will be able to provide a more sustainable and higher quality semantic layer for CMDI. An important lesson from the ISOcat experience is that technology is not always the main problem, although a complicated data model or interface never helps. What we do believe in, is establishing robust, yet simple management procedures, as outlined in Section 7. These rely on teamwork in the national CCR content coordinators forum, together with active involvement of the user community.

Acknowledgments

The authors like to thank the national CCR content coordinators forum for the fruitful discussions on the procedures around the CCR. They also like to thank the Max Planck Institute for Psycholinguistics, CLARIN-NL and the Meertens Institute for their support to realize a smooth transition from ISOcat to the CCR.

References

Daan Broeder, Menzo Windhouwer, Dieter van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. CMDI: a Component Metadata Infrastructure. Proceedings of LREC Workshop *Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR*. Istanbul, Turkey.

Daan Broeder, Ineke Schuurman, and Menzo Windhouwer. 2014. Experiences with the ISOcat Data Category Registry. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland.

Hennie Brugman and Mark Lindeman. 2012. Publishing and Exploiting Vocabularies using the OpenSKOS Repository Service. *Proceedings of the Describing Language Resources with Metadata workshop (LREC 2012)*, Istanbul, Turkey.

CE-2015-0688. *CLARIN-PLUS CCR analysis*. CLARIN ERIC, Utrecht.

Matej Durco and Menzo Windhouwer. 2013. Semantic Mapping in CLARIN Component Metadata. In E. Garoufallou and J. Greenberg (eds.), *Metadata and Semantics Research (MTSR 2013)*, CCIS Vol. 390, Springer.

ISO 12620:2009. *Specification of data categories and management of a Data Category Registry for language resources*. International Organization for Standardization, Geneve.

Menzo Windhouwer. 2012. RELcat: a Relation Registry for ISOcat data categories. *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey.

Sue Ellen Wright, Menzo Windhouwer, Ineke Schuurman and Daan Broeder. 2014. Segueing from a Data Category Registry to a Data Concept Registry. *Proceedings of the 11th international conference on Terminology and Knowledge Engineering (TKE 2014)*, Berlin, Germany.