CHAPTER 5

CLAVAS: A CLARIN Vocabulary and Alignment Service

Hennie Brugman
Meertens Institute, Amsterdam, Netherlands
hennie.brugman@meertens.knaw.nl

ABSTRACT

The CLARIN Vocabulary and Alignment Service (CLAVAS) project started as an attempt to address problems that occurred with the use of the CLARIN ISOcat repository to deal with vocabularies used for value ranges in CMDI metadata elements and attributes. CLAVAS objectives were easier maintenance of vocabularies and new ways to exploit these vocabularies in end user tools; the underlying aims were the improvement of metadata quality and the reduction of concept proliferation in the ISOcat repository.

Also, during the project three specific vocabularies were processed and made available via the newly created CLAVAS service platform. We present and evaluate CLAVAS in relation to the family of semantic registries that together played a role in CLARIN: ISOcat, OpenSKOS and the CLARIN Concept Registry. We discuss developments that occurred after the end of the CLAVAS project period, especially the discontinuation of ISOcat as a CLARIN service and the subsequent roles of OpenSKOS and CLAVAS.

5.1 Introduction

Central in CLARIN is work on the standardisation, creation and exploitation of metadata for linguistic resources. Many metadata records were produced, collected and published in IMDI, and have been subsequently in CMDI (Broeder et al., 2012). However, over the course of CLARIN more and more problems with the quality of this metadata became apparent (Broeder et al., 2014). One category of problems had to do with the interpretation of metadata fields and values. Although

How to cite this book chapter:
Brugman, H. 2017. CLAVAS: A CLARIN Vocabulary and Alignment Service. In: Odijk, J and van Hessen, A. (eds.) *CLARIN in the Low Countries*, Pp. 61–69. London: Ubiquity Press. DOI: https://doi.org/10.5334/bbi.5.

guidelines and technical means were provided to document, publish and share the semantics of resource descriptions, in practice this did not lead to standardisation and reuse, but to proliferation of often underspecified semantic descriptions.

CLARIN provided the ISOcat Data Category Service as a central registry for the meaning of resource descriptions (Windhouwer et al., 2010). ISOcat combined two purposes: on the one hand it aimed for complex and precise semantic specifications of data categories, with community standardisation as the long-term purpose; on the other hand it supported registration of more volatile and project-specific ranges of values (e.g. lists of genres). For both types of data categories it used the same open update policy, in which everybody is allowed to add their own data categories.

In the meantime, the CATCHPlus project[1] developed OpenSKOS (Brugman and Lindeman, 2012). OpenSKOS is a repository service platform that offers uniform and standardised ways to publish, manage and retrieve vocabulary data in forms that can be used for various usage scenarios, e.g. during metadata creation (e.g. pick lists with autocompletion) or during metadata curation (e.g. search for preferred labels, given some alternative label).

The name 'OpenSKOS' refers to:

- A *standard* format. The OpenSKOS service uses the W3C SKOS[2] recommendation (Miles and Bechhofer, 2009) as its data model.
- An *architecture*. The platform supports multiple nodes that can exchange vocabulary data using the OAI-PMH harvesting protocol.
- An *Application Programming Interface*. This API offers functionality for creating and maintaining SKOS vocabularies.
- A Linked Data endpoint. Vocabulary data in an OpenSKOS node can be addressed with resolvable HTTP links or queried using the SPARQL query language. Links both within and across vocabularies are supported.
- A *publication* platform.
- A platform for *creation and maintenance* of vocabularies.
- A user community promoting the use of *Open licences* for vocabulary data.

Unlike ISOcat, OpenSKOS uses a simple data model with deliberately imprecise semantics. Its main purpose is easy and flexible management, publication and reuse of vocabularies used for resource description. It has no ambition to standardise concepts within communities, other than de-facto, by shared usage. OpenSKOS itself allows an open update strategy, like ISOcat does, in the sense that if no specific restrictions are applied OpenSKOS vocabularies can be freely uploaded and modified.

Where the ISOcat approach seemed especially suitable to define and publish core domain concepts, such as often-used fields in metadata specifications, OpenSKOS seemed the better choice for dealing with the second type of concepts, that is the ranges of values used to fill those fields. The CLARIN-NL CLARIN Vocabulary and Alignment Service (CLAVAS) project was started as a best-of-both-worlds effort to combine the two concept registry systems, where each system dealt with the type of concepts that was best supported by its design.

The CLAVAS[3] project extends CATCHPlus' OpenSKOS platform with CLARIN-specific components: first, it publishes three vocabularies required by the CLARIN community and it offers tools to keep those vocabularies synchronised with their sources. Second, it offers an interactive web application that can be used to curate existing vocabularies or to create new ones from scratch

[1] www.catchplus.nl
[2] http://www.w3.org/TR/skos-primer/
[3] https://openskos.meertens.knaw.nl/clavas/

(the OpenSKOS Editor). And third, it aims for implementation of usage scenarios where mainstream CLARIN tools (like Arbil[4]) use vocabularies directly via the OpenSKOS RESTful API.

This chapter starts with a presentation of the history, design objectives, description and comparison of both ISOcat and OpenSKOS semantic registries. In subsequent sections we present the CLAVAS project, its objectives, components and architecture, we evaluate its results and we present lessons learned.

Currently, CLAVAS usage is the proposed solution for utilising external vocabularies at the operational level in the CMDI 1.2 specification (CLARIN CMDI Taskforce, 2014) and therefore still is part of the CLARIN landscape of semantic registries: we thus also describe the state of this landscape and recent developments.

5.2 ISOcat and OpenSKOS

As said in the introduction, CLAVAS started as an attempt to combine two existing semantic registry systems, ISOcat and OpenSKOS, thereby exploiting the strengths of each of the two approaches. In Table 5.1 a direct comparison between the two initial systems on a number of specific aspects is given.

	ISOcat	**OpenSKOS**
data model	• (ISO) data category-based • accurate semantics	• (SKOS) concept-based • simplified semantics
relations between concepts in vocabularies included	no	yes
linking of concepts across vocabularies supported	no	yes
update strategy	open	open
main objectives	• community standards, reference vocabularies • documentation • reuse of vocabularies	• publish and reuse vocabularies • cross-link/align vocabularies • easy in, easy out • many alternatives for input and output • practical applications by humans and machines
input	• manual via GUI	• manual via GUI • RDF file upload • harvesting using OAI-PMH • via RESTful API
output	• user interface • RESTful API	• user interface • RESTful API • OAI-PMH data provider • browsable html pages
standards	ISO TC37 DCR; 12620	W3C SKOS

Table 5.1: Comparison of ISOcat and OpenSKOS.

[4] https://tla.mpi.nl/tools/tla-tools/arbil/

Concerning the data models used, ISOcat uses data categories, with Complex Data Categories representing properties of items and Simple Data Categories representing atomic elements to be included in the value set of Complex Data Categories (Schuurman et al., 2015). ISOcat does not contain relations between Data Categories. OpenSKOS has a simpler model, and uses SKOS as its data model. The SKOS (and OpenSKOS) model includes relations between concepts, as well as properties of these concepts to model alignment between concepts in different vocabularies (SKOS ConceptSchemes).

Concerning input and output options, the OpenSKOS design objectives have a stronger focus on supporting a wide range of practical applications, both by humans and by machines, such as online term suggestion using autocompletion in the context of (metadata editing) tools.

5.3 The CLAVAS Project

In this section we first present the CLAVAS components and architecture and then discuss the vocabulary data involved in the project.

5.3.1 Components and Architecture

Figure 5.1 shows an architectural overview of the CLAVAS building blocks. The figure represents a data flow, starting from sources of vocabulary data (left) and ending with the CLAVAS semantic repository (right). The blue parts are the parts developed by the CLAVAS project. We chose three different source vocabularies to import into this repository: ISO 639-3 language codes,[5]

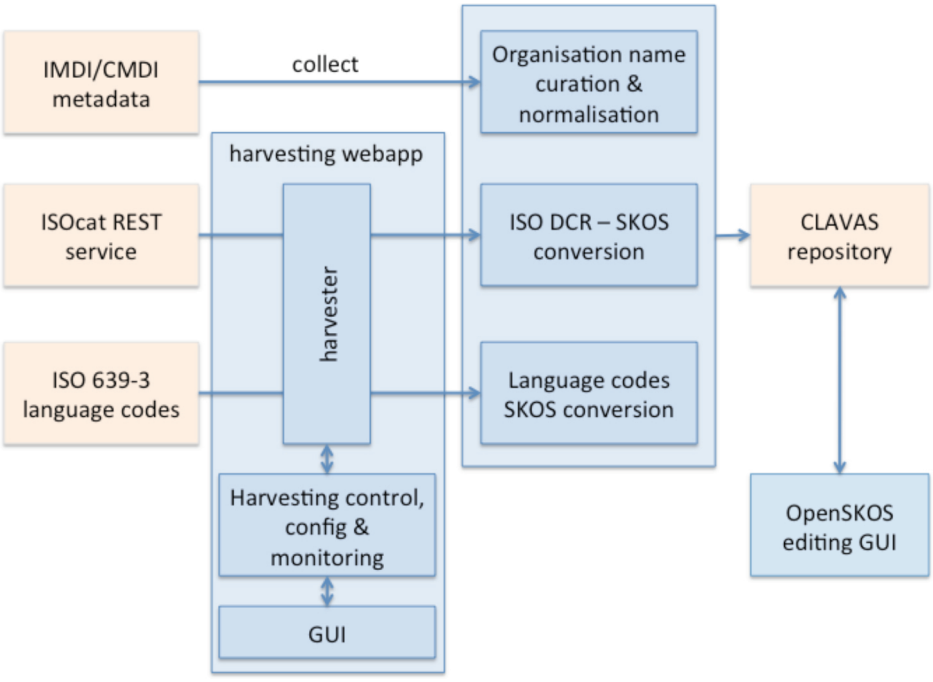

Figure 5.1: CLAVAS architectural overview.

[5] http://www-01.sil.org/iso639-3/

organisation names, and ISOcat closed and Simple Data Categories.[6] These vocabularies differed with respect to quality at the start of the project, transport protocol and import pipeline. Two of the import pipelines were fully automated; the third involved manual curation. Import pipelines can be controlled via a dedicated harvesting web application. Once imported in CLAVAS, all vocabularies can be manually inspected and curated with the help of a specialised vocabulary curation web application, the OpenSKOS Editor.

5.3.1.1 Updating Vocabularies

A simple web application was built and integrated in the CLAVAS OpenSKOS site. This web application has three tabs, one for each of the CLAVAS vocabularies. The two that can be periodically updated have a facility to modify download paths for the source information and a download button; the RDF file that is created and downloaded is suitable to upload to an OpenSKOS instance. Sources for the web application as well as for all converters are available from the CLAVAS GitHub repository.

5.3.1.2 OpenSKOS Vocabulary Editor

CLAVAS requirements with respect to the vocabulary editor were relatively simple: the presence of the functionality to curate simple, unstructured lists of concepts was sufficient. However, the Netherlands Institute for Sound and Vision contributed to the OpenSKOS project by providing a full-blown thesaurus editing environment, including support for simple workflows. The OpenSKOS Editor was built by Picturae as an integrated extension of the OpenSKOS software stack and was made available under open source licence via the CATCHPlus OpenSKOS GitHub repository. The OpenSKOS Editor was tested and applied during the CLAVAS project, and is currently part of the CLAVAS service.

5.3.2 Three Vocabularies

5.3.2.1 ISO 639-3

ISO 639-3 language codes can be downloaded/harvested by the CLAVAS harvesting web application. The download location (at SIL) of the source files can be modified if it changes in the future. The source files are parsed and converted to an SKOS RDF/XML file that is uploaded and subsequently published by the CLAVAS OpenSKOS instance. The core of the conversion module is a script that can also be used as a standalone tool. The script is available as part of the CLAVAS open source distribution on GitHub.[7]

5.3.2.2 ISOcat

Closed and simple data categories can be harvested directly from ISOcat as RDF/XML; this function was made available for CLAVAS by MPI/TLA. The CLAVAS harvesting web app does some necessary post-processing on the ISOcat SKOS RDF/XML data. The resulting SKOS is uploaded to and made available from the CLAVAS OpenSKOS site.

5.3.2.3 Organisation Names

Organisation names used in CLARIN IMDI and CMDI metadata are diverse and are very domain-specific (for example, several linguistics departments at universities all over the world). Because no

[6] http://www.isocat.org/
[7] https://github.com/hennie/CLAVAS

existing vocabulary resource was found that covered the full range of organisation names actually used in the CLARIN VLO, it was decided to start a manual curation project. This project was undertaken by the Nijmegen-based Data Curation Service, in close collaboration with CLAVAS. The process was as follows: all organisation name instances were extracted from the VLO. Two manual passes bundled spelling variations of organisation names and identified the preferred spelling. A tabular text format was used to store this information;remarks and editorial notes were also recorded. The intermediate results were automatically processed and converted to SKOS. The conversion process also extracted hierarchical structure where possible. Conversion errors were kept separate in the form of a set of SKOS Concepts that needed further curation. The converted result was evaluated using the search and browse facilities of the OpenSKOS Editor. In a final manual pass the OpenSKOS Editor was used to fix the Concepts. In the process, the OpenSKOS Editor was also evaluated with respect to usability for CLARIN vocabulary curation tasks. As with previous vocabularies, the Organization Names vocabulary was published on the CLAVAS OpenSKOS site.

5.4 Evaluation of Results

5.4.1 Evaluation Criteria

At the beginning of CLAVAS we explicitly formulated success factors from the perspectives of all types of stakeholders for the CLAVAS service to be developed. We first present our initial evaluation criteria and subsequently discuss how far they were met in the results of the CLAVAS project.

Our initial evaluation criteria were as follows:

Researchers creating and maintaining metadata:

- Will produce better-quality metadata with less effort.
- Will be able to use vocabularies that are outside the scope of ISO DCR (e.g. long lists of names, or vocabularies from the cultural heritage domain).
- Will have seamless access to vocabulary information in the context of their metadata editing environment.
- Will have easy ways (such as autocompletion) to select from available terms.
- Will be able to find terms on the basis of alternative labels or closest match.
- Will be able to contribute terms that are missing from a vocabulary (if that is allowed for the vocabulary in question).

Collection users

- Will experience better precision and recall when searching metadata.
- Will be able to involve vocabularies when formulating queries.

Collection managers curating metadata

- Will be able to check existing metadata values against some vocabulary. The system should be able to provide them with matching terms or closest matching candidates.
- Will be able to do vocabulary curation when they are working on metadata curation.

Vocabulary curators

- Will be able to easily check if a suitable term already exists in some vocabulary (e.g. by searching for it on the basis of one of its alternative labels and inspecting notes or definitions).

CLAVAS content managers

- Will be able to set parameters of harvesting or update processes and have control over these processes. The harvesting process will be so clear, reliable and stable that it will be easy to (re)run it, even if this (re)run is done only once a year.

Builders of tools and services

- Will be able to integrate vocabulary support in their tools and services easily.

5.4.2 Actual Realisation of These Criteria

Although the CLAVAS service is up-and-running and all planned components are in place, not much use is made of CLAVAS: the most probable reason for this is that CLAVAS benefits will become apparent only after integration of CLAVAS services into other tools, which has not happened yet. Another reason may be related to the potential lack of usefulness (for ISO language codes and the – somewhat ad hoc – republished version of ISOcat data categories) and quality (for organisation names) of the vocabularies currently published by CLAVAS. In more detail, these are the evaluations of the service per group of stakeholders:

For metadata editors:

- It is too early to make any claims about improved metadata quality. A necessary step to take is the integration of the use of CLAVAS/OpenSKOS in tools like Arbil. In general, CLAVAS supports open vocabularies that are not suited for ISOcat, it offers autocompletion and easy, fine-grained search on the basis of all SKOS attributes, and it contains the basic functionality to allow suggestion of missing concepts by metadata editors.

For collection users and metadata managers:

- It is too early to make any claims relevant to these stakeholders here.

For vocabulary curators:

- The criteria are met. Curation with OpenSKOS was tested by the CLARIN-NL Data Curation Service and found satisfactory.

For CLAVAS content managers:

- Periodic updates of ISO 639-3 and ISOcat vocabularies are possible and easy to perform. The Organization Names vocabulary is bootstrapped from the VLO and can be curated manually using the OpenSKOS Editor.

For tool builders:

- The RESTful API is freely available for builders of tools and services.

5.5 Current State and Future Work

Although CLAVAS was not widely adopted by the CLARIN community in terms of actual usage, this is different for the underlying OpenSKOS platform, which is used by several organisations and projects across Europe, both in the domain of linguistic resources and in the description of cultural heritage collections.

Schuurman et al. (2015) describes why and how the ISOcat semantic registry was phased out and replaced by a new CLARIN Concept Registry (CCR) based on OpenSKOS. The main aims of this migration were to reduce complexity and the proliferation of concepts by adopting a simpler data model, new tooling and revised, stricter maintenance procedures that include editorial supervision.

For the new CCR several new components were developed by the Meertens Institute: support for persistent identifiers (handles), Shibboleth-based authentication, support for skos:Collections in the OpenSKOS data model and interactive web access via a new faceted browser.

Both CCR and CLAVAS are now based on the same data model (SKOS) and the same service platform (OpenSKOS). Furthermore, usage of CLAVAS as a service platform to provide open or closed vocabularies as value domains for CMDI elements and attributes is part of the CMDI 1.2 specification (CLARIN CMDI Taskforce, 2014). Thus it can be expected that CLAVAS will play a bigger role in the CLARIN community in the near future.

Currently, there are new developments going on in the OpenSKOS community, which, in addition to CLARIN, includes a number of large cultural heritage institutions and companies in the Netherlands and Europe. One of them, the Netherlands Institute for Sound and Vision, together with Picturae, is developing a new version of OpenSKOS. The new version contains an RDF triple store. It supports SKOS XL as data model and provides a SPARQL endpoint to its users. For this updated OpenSKOS the Meertens Institute is currently extending the API to cover the complete OpenSKOS data model, and is building a new faceted browser on top of that.

5.6 Conclusions

CLAVAS succeeded in delivering a service platform for the provision and manual curation of vocabularies for value domains in CMDI. However, the service is currently not widely used. The main reason for this most likely is that CLAVAS, being mostly middle-ware, hardly offers functionality directly targeted at end users. Built-in support from tools (e.g. Arbil) may help improve on this situation.

CMDI and CMDI vocabularies are evolving towards RDF and Linked Open Data. For example, a recent development is CMDI2RDF, a CLARIAH[8] activity that is going to provide CMDI metadata as RDF. CLAVAS and OpenSKOS, being SKOS- and therefore RDF-based, nicely fit into this development.

CLAVAS is part of the wider range of OpenSKOS applications, and is the component where CLARIN may benefit the most from contributions of other OpenSKOS users. Especially, well-curated and extensive vocabularies from the cultural heritage domain may turn out to be very useful for linguistic and digital humanities applications as well.

Acknowledgements

Work on CLAVAS was funded by CLARIN-NL. OpenSKOS was produced by the CATCHPlus project, which was funded by NWO as part of the CATCH programme. We thank the Netherlands Institute for Sound and Vision and Picturae for their collaboration and for funding parts of OpenSKOS.

[8] http://www.clariah.nl/

References

Broeder, D., Windhouwer, M., van Uytvanck, D., Goosen, T., & Trippel, T. (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.

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

Brugman, H. & Lindeman, M. (2012). *Publishing and Exploiting Vocabularies using the OpenSKOS Repository Service.* Istanbul, Describing Language Resources with Metadata workshop at LREC2012.

CLARIN CMDI Taskforce (2014). *CMDI 1.2 changes - executive summary.* Technical Report CE 2014-0318, CLARIN ERIC, Utrecht, The Netherlands, April 2014.

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

Miles, A., Bechhofer, S. (2009). *SKOS Simple Knowledge Organisation System Reference*. W3C Recommendation 18 August 2009.

Schuurman, I., Windhouwer, M., Ohren, O. & Zeman, D. (2015). *CLARIN Concept Registry: the new semantic registry*. At the CLARIN Annual Conference. Wroclaw, Poland, October 15–17, 2015.

Windhouwer, M.A., Wright, S.E., Kemps-Snijders, M. (2010). *Referencing ISOcat data categories*. In proceedings of the LRT standards workshop (LREC 2010), Malta, May 18, 2010.