CLARIN's Virtual Language Observatory (VLO) under scrutiny -- The VLO taskforce of the CLARIN-D centres

Susanne Haaf[1], Peter Fankhauser[2], Thorsten Trippel[3], Kerstin Eckart[4], Thomas Eckart[5], Hanna Hedeland[6], Axel Herold[1], Jörg Knappen[7], Florian Schiel[8], Jens Stegmann[4], Dieter van Uytvanck[9]

CLARIN-D centres: [1]Berlin-Brandenburg Academy of Sciences and Humanities, Berlin; [2]Institute for the German Language, Mannheim; [3]University of Tuebingen; [4]University of Stuttgart; [5]University of Leipzig; [6]University of Hamburg; [7]Saarland University, Saarbruecken; [8]University of Munich; [9]Max-Planck Institute, Nijmegen

Keywords: VLO, CMDI, metadata curation, closed vocabulary

1. Introduction: Challenges for searching for language resources

Finding language resources poses a challenge for researchers from the humanities and social sciences using the CLARIN infrastructure. The challenges are in terms of usability of language resources, completeness of knowledge about existing resources and detail of information about possibly useful resources. The question of detail is related both, to the level of description and in the granularity of resource, i.e. if resources are grouped as one or if the parts are separated and treated as separate resources.

CLARIN provides language resources including corpora, lexical resources, software tools, webservices, etc. The number of resources is huge and getting an overview is virtually impossible without technical assistance. For this reason, a specialised search and discovery service was created, the VLO, which allowed a faceted search for language resources that provide descriptive metadata at the CLARIN centres and other registered institutions providing metadata in accepted formats via OAI-PMH. The content of the facets are based on the content of the metadata files and mappings of data categories onto predefined facets.

Due to the variability of the CMDI metadata framework (see Broeder, et al.,2010), various resource objects and types can be described and though the descriptions may be similar in some detail, they will be different as many other areas as there are types of resources. At present (June 2014) about 650 000 resources are searchable via the VLO, claiming to be about 250 resource types. The huge variety of types - notwithstanding the question if this is justified or not - creates a complexity for searches: individual search terms may not partition the search space significantly if they are too general while at the same time if they are too specific they are useless for faceted search or for guiding users to resources they are not aware of but where they have some characteristic features.

Thus, huge amounts of resource descriptions are put together within the VLO and queries across this stock should be possible not only via string search but also via filtering methods, i.e. via lists of searchable categories provided by the facet browser. In an internal review process, it was apparent that the challenges were not completely met: resources were not easy to find, the facet values were inconsistent and confusing to users, the descriptions were problematic and the usability of the search interface was falling behind expectations.

2. Constitution and Purpose of the VLO Taskforce

Seeing the challenges not being met, CLARIN-D decided to put considerable effort into improving the situation. To accomplish that each CLARIN-D centre was asked to nominate two delegates to a taskforce with the mission of working on the VLO: each centre nominated a technical expert to help on the technical implementation and data provision, and a content expert for curating the content of metadata records without having to find another expert in case changes were devised.

The VLO taskforce (VLO-TF) started its work in October 2013. In regular meetings questions of metadata curation, suitable ways of exploiting CMDI records for the VLO, possible changes to the web platform which to improve the usability of the VLO etc. were discussed.

For the metadata curation it became obvious that a higher degree of standardization would be advisable, i.e. using more uniform values of potentially closed classes of metadata categories, using core data categories more systematically, and providing prose descriptions with a more generic reader in mind. A cross centre evaluation helped to provide feedback on the practices at each centre.

On the more technical side, questions of usability and applicability of the web platform for different possible usage scenarios were addressed. This meant that the facets were required to show specific data categories from the metadata instances, ignore others with the same data categories but in other components, the number and definition of facets and the complexity of list of values presented. During the evaluation in the VLO-TF some technical problems were also identified.

According to the recommendations discussed by the task force, the metadata records of the CLARIN-D centres were adjusted and, where necessary, improved and requirements for the VLO platform were specified.

3. Attended Tasks of the VLO Taskforce

As stated above, the VLO provides a facet browser which allows for the filtering of metadata records according to previously specified categories (facets). The concept of facets for metadata research within the VLO posed three problems, the VLO-TF had to deal with: first, the choice of categories for the first acquisition of the VLO, second, the automatic selection of suitable metadata for the VLO facets, third, the filling of the facets in a homogeneous way, and fourth, the issue of quality assurance and quality control for the metadata harvested by the VLO. Furthermore, apart from the issues concerning the facetted search of the VLO, the VLO-TF attended to the question of adequate representations of relationships between resources within CMDI profiles. Finally, the VLO-TF was concerned with the issue of documentation since the discussions conducted about the VLO and the guidelines resulting from them should be made comprehensible and usable for data providers.

3.1 Selection of Facets

To solve the first issue the given selection of search facets within the VLO was taken under consideration by the taskforce, and a new selection was agreed upon, comprising:

- Resource Type (e.g. text, lexical resource, video data, audio data, etc.)
- Modality (e.g. speech, writing, facial-expressions, etc.)
- Format (e.g. TXT, JPG, TEI-XML, etc.)
- Language(s) of the resource (i.e. primary language the ressource is written/spoken in)
- Organisation (institution currently providing the ressource)
- Country (country which the ressource originates fromas opposed to the country, where the ressource is hosted at the moment)
- National Project (project providing the ressource)
- Collection (superordinate collection of ressources)
- Time Coverage (time span represented by the primary data/time of creation/recording etc.; as opposed to the amount of time put into the preparation and provision of the resource).

From these search facets, only the *Time Coverage* facet could not yet be integrated into the VLO facet browser but is still under preparation.

On the other hand a couple of search facets included in the original selection are still part of the facet browser since their necessity and usability, respectively, is still being investigated. Those facets are: Continent, Genre, Subject, Data Provider, Keyword.

Apart from the search facets the VLO-TF attended to the description facets given at the target page of each resource. Here, the question arises, which facets would be best to quickly describe a resource. The current selection of description facets consists of 15 items, namely collection, continent, country, dataProvider, description, genre, id, languages, metadataSource, name, nationalProject, organisation, projectName, subject, year. This selection is currently under revision.

3.2 Mapping different CMDI metadata specifications on one facet browser

The second issue discussed by the VLO-TF was listing the values of the facets based on the CMDI metadata instances provided by the various repositories. One particular challenge of this task consists is in the variety of profiles used by different data providers and their use of CMDI specifications due to their differing needs for resource description. With this variability it is difficult to automatically extract the information needed for a specific facet from a metadata record. For example, the element `<date>` could refer to the date of creation of a text as well as the date of its first publication or the date, the metadata record was created, if it is interpreted without the context of a component using this element. Therefore, CMDI metadata specifications are recommended to include a mapping with appropriate ISOcat categories (see Broeder et al, 2014) and reuse components. Even though this connection to ISOcat might very well help with the disambiguation of ambiguous CMDI components, ISOcat categories are still imprecise or ambiguous due to the quality of description.

The VLO-TF therefore – as a short-term solution – addressed the problem of incorrect content of facets by providing XPath expressions corresponding with VLO facets for those CMDI specifications which were utilized for the resource descriptions of the CLARIN-D centres. Those collections of XPaths include whitelists (lists of true positives, i.e. those XPaths which lead directly to element contents suitable for a certain facet) and blacklists (lists of false positives, i.e. those XPaths which lead to element content which might be mistaken as suitable for a certain facet).

In the long term it is planned to return to the method of analyzing ISOcat categories automatically for providing values of the VLO facets. For this, members the VLO-TF started to examine the ISOcat Data Categories used by CMDI profiles in terms of the scope of their usage as well as to define sets of ISOcat Data Categories suitable for VLO facets. An important third task in this context is the disambiguation of certain highly ambiguous ISOcat Data Categories. Here, the VLO-TF will propose recommendations to the National Metadata Quality and ISOcat coordinators.

3.3 Controlled Vocabularies

After the correct metadata descriptions for a particular facet are extracted from the available CMDI records the next challenge is to cluster similar information. That is, similar metadata descriptions might differ in their language (e.g. resource type “written” vs. “schriftlich”), serialization (e.g. “written” vs. “Written”) or selection of label (e.g. “written” vs. “written corpus” vs. “writing”). Therefore it is necessary to, wherever applicable, control the vocabularies used for certain metadata categories. The difficulty, however, is to decide, where and how to apply such controlled vocabularies and how to communicate the closed vocabulary to the metadata providers, addressing questions such as: Should the controlled vocabulary be

included in the component definition, so that metadata providers resort to the given vocuabularies for their metadata descriptions? Or should there be algorithmic procedures to map different metadata descriptions to a restricted vocabulary? The VLO-TF decided to follow both ways. On the one hand controlled metadata vocabularies for VLO facets are going to be supplied in order to facilitate the recording of metadata compatible with other VLO resource descriptions. On the other hand, metadata providers should still be allowed to resort to their own vocabularies. Therefore, solutions for the automatic mapping of similar metadata descriptions are being implemented.

3.4 Metadata Quality

Another problem for the correctness of mapping content of the metadata categories to VLO facets as well as the homogeneity of data within facets is the quality of metadata. As stated above, among other things the VLO-TF is working on creating and establishing guidelines for the design of CMDI components (e.g. connection to ISOCat categories), the extensiveness of metadata records (e.g. the information needed for the VLO) or the style of metadata descriptions (e.g. usage of controlled vocabularies). These guidelines, once defined and documented, help estimating metadata quality (see also Trippel, et al., 2014). For this task, the VLO-TF aims at developing and implementing algorithms to check the quality of metadata harvested for the VLO. Based on the results of such checks it will become possible to send feedback to metadata providers or to adjust metadata to the given guidelines before integrating them in the VLO.
Since these software solutions for metadata quality checks are still under preparation, the VLO-TF members mutually reviewed other CLARIN-D center's metadata. This initiative was an important first step towards metadata quality and homogeneity within the VLO.

3.5 Relationships

Some CLARIN centers represent relationships between resources, for example *part-of/hasPart* for corpora that consist of sub-corpora or *version-of* for representing various versions of resources. The CMDI-framework provides a variety of ways to encode such relationships. As a consequence, the representation of relationships is rather heterogeneous, and the VLO cannot easily exploit them, e.g., for improving ranking of results - latest, root resources first - or for facilitating navigation among resources.
To alleviate this problem, the taskforce has systematically analyzed and documented the various relationship representations with the goal of homogenizing them by recommending best practices at least for the most important kinds of relationships.

3.6 Documentation

The main focus of the VLO-TF in the first phase has been on homogenization, in particular, agreeing on a common set of search facets and converging on vocabularies to fill these facets, where appropriate. As these agreements mature, they need to be documented. To this end, the VLO-TF plans to document the facets with a description of their intended semantics and use, a formal semantics by means of ISOCat, and recommended best practices with respect to their value range. Moreover, reusable CMDI components together with example resources will be used to illustrate the actual use of the recommended facets. This documentation shall thus give data providers a set of readily usable building blocks to describe their resources in such a way that they can be readily found in the VLO.

4. Future Work and Prospects

The Virtual Language Observatory has been designed as the central platform for primary access to the diverse resources and tools provided by CLARIN. Here, users are able to filter the wide range of various resources according to their specific needs and research interests. Thus being primarily addressed to CLARIN's users the VLO as a platform as well as the resources represented by this platform have to be subject to continuous usability and quality checks. To address this task the VLO-TF has been constituted. Here, representatives of all CLARIN-D centres are working together to improve the VLO and the metadata it provides in various aspects, such as creating concepts for the presentation of the material, finding appropriate ways for quality assurance, looking for CMDI conformant solutions for difficult metadata modelling tasks, or offering guidelines for data providers and users in order to facilitate their respective work with the VLO. Until now, the VLO-TF has been concentrating on several different tasks. In many cases solutions could already be found which were then implemented by the VLO developers in Nijmegen and Leipzig. However, the work of the VLO-TF has not at all come to an end, yet, but will be continued successively, this way preferably helping the VLO to become an easy to use platform for the whole range of language resources and tools provided by the CLARIN community.

5. References

Broeder, D.; Kemps-Snijders, M.; Van Uytvanck, D.; Windhouwer, M.; Withers, P.; Wittenburg, P. and Zinn, C. (2010). A Data Category Registry- and Component-based Metadata Framework. In Calzolari, N. et al. (eds.). *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 10)*, May 19–21, 2010, Valletta, Malta, pp. 43–47.

Broeder, D.; Schuurman, I. and Windhouwer, M. (2014). Experiences with the ISOcat Data Category Registry. In Calzolari, N. et al. (eds). *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 14)*, May 26–31, 2014, Reykjavik.

Trippel, T.; Broeder, D.; Durco, M. and Ohren, O. (2014). Towards automatic quality assessment of component metadata. In Calzolari, N. et al. (eds). *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 14)*, May 26–31, 2014, Reykjavik, Iceland.