Virtual Language Observatory 3.0: What's New?

Twan Goosen[1], **Thomas Eckart**[2]

[1]CLARIN ERIC [2]University of Leipzig
[1]Trans 10, 3512 JK Utrecht, The Netherlands [2]Augustusplatz 10, 04109 Leipzig, Germany
[1]twan@clarin.eu [2]teckart@informatik.uni-leipzig.de

Keywords: metadata, CMDI, search, faceted browsing

1. Introduction

The Virtual Language Observatory faceted browser (VLO)[1] was conceived as a means to explore the vast and diverse array of language related resources described by means of the Component Metadata Infrastructure (CMDI; see Broeder et al. (2012)) and compatible metadata formats in a uniform and easy way (Van Uytvanck et al., 2012). Within the CLARIN infrastructure, the VLO has the role of a general purpose entry point for discovering and browsing through the resources described within the CLARIN metadata in an easy and intuitive manner. In contrast, it is not intended to provide methods for quickly answering detailed research questions by means of advanced or specialised search options.

Three major versions of the VLO have been developed over the past years. The first incarnation was based on the Flamenco framework (Stoica et al., 2007). It prominently presented the set of facets available for narrowing down the search results. The facets were based on a mapping from concrete metadata fields to a small set of selected salient aspects such as country, language and subject (Van Uytvanck et al., 2010). After this first version, the VLO was reimplemented, using Apache Solr[2] as search backend with a custom web front-end based on Apache Wicket[3]. The user interface was largely based on the original VLO. The "importer" component of the VLO was developed to transform field values into faceted indices on basis of the ISOcat data categories (Kemp-Snijders et al., 2009) and other information available defined in the CMDI metadata profiles. Many advanced mapping functionalities have been incorporated, as described by Van Uytvanck et al. (2012).

The most recent development is an almost complete rewrite of the front-end, which led to the third major release of the VLO. This version uses a continuation of the importer of the previous version of the VLO. The following section describes the changes in both the front end and importer components since the report of Van Uytvanck et al. (2012).

2. Changes

A functional and technical analysis of the second version of the VLO was carried out by Goosen (2014). In this analysis, a number of issues are identified, primarily relating to mapping, usability and quality and maintainability of the code base. Concrete implementations of mapping improvements are being investigated (see Haaf et al., 2014), but many of the recommendations with respect to usability and code base have been adopted in the implementation of VLO version 3. As a result, the stability, performance and maintainability of the application as a whole have been increased as a result of library upgrades, improved modularisation, and increased coverage by automated tests. The following two sections focus on the functional changes in the front end and importer modules respectively.

2.1. Web front end

The front end of the VLO has been almost completely reimplemented and redesigned. It is based on the latest version of the Apache Wicket framework and makes heavy use of its capability to provide partial updates via AJAX[4] with fallback for browsers with JavaScript disabled or unavailable. This makes for a responsive user interface while maximising compatibility with a broad range of clients.

In contrast to version 2 of the VLO, which has most 'screen real estate' reserved for the presentations of the facets, the entry page of the new front end (see Figure 1) prominently shows a search field and, below that, a number of browsing options. Entering a search term or selecting a browsing option will take the user to the main search interface (see Figure 2), which shows a large search field above the paginated list of query result on the left, and a list of expandable facets on the right. Rather than showing available values for all facets by default, the items in this list are 'collapsed' at first. The aim of this is to offer a less cluttered interface while keeping the ability to quickly find the desired value for selection and observe the interaction between separate facets.

Each expanded facet has a text field that allows the user to search for specific values. The list below the field shows the ten matches with the highest number of matching records and gets updated dynamically as the user enters a filter term. Typing 'English' as a filter term, for example, will limit the set of options to "English", "Hawai'i Creole English", "Middle English", etc. If there are more than ten matching options, a *more* link is available that opens a pop-up dialogue allowing the user to browse through all values available for the selected facet and to use the same filtering mechanism on this complete list. In contrast to previous

[1] `http://catalog.clarin.eu/vlo`
[2] `http://lucene.apache.org/solr`
[3] `http://wicket.apache.org`
[4] Asynchronous Javascript and XML

Figure 1: The VLO 3.0 entry page

versions of the VLO, this list is paginated and can be sorted either alphabetically or by number of matching records.
Selecting a value for one or more facets dynamically updates the query result listing, as well as a 'breadcrumbs bar' at the top of the user interface that allows the user to go back to previous points in the selection process or easily remove any of the facet values from the current selection.
Each result item can be expanded, so that the user can obtain more detailed information concerning a specific record (typically a resource, collection or tool) without having to leave the search interface, as was the case with previous versions of the VLO. Clicking the title of a record will take the user to the record page, which provides all available information for that specific record, including links to all individual resources it refers to and optionally to a *landing page* or *search page*, if specified in the record by the provider, or a link to CLARIN's Federated Content Search Aggregator[5] if a search endpoint and handle are available. When multiple values exist for a single field inside a record, these values are now shown separated by a visual divider. Another new feature in the record page is the separate table of technical details, which contains record specific values that are not of interest to the average user but can be valuable to administrators and developers, and is shown as an expandable section on the bottom of the page. Furthermore, the experienced loading time of the resource listing has been improved by a *lazy* resolution of resource references on basis of handles[6].
An advanced 'search options' panel in the search interface allows the user to limit the returned results to items that support CLARIN Federated Content Search (FCS)[7].

[5] `http://weblicht.sfs.uni-tuebingen.de/Aggregator`, see Stehouwer et al. (2012)

[6] Handles are identifiers inside the Handle System, see `http://handle.net/`

[7] The most salient entry point of FCS to end users is the 'Searching multiple corpora' featured offered at `http://www.clarin.eu`

2.2. Importer

Despite the focus on a complete overhaul of the user interface there also have been some significant changes regarding the import process. Based on updated libraries (namely recent versions of the search platform Apache Solr and the VTD-XML[8] parser) several import procedures were revised and recent changes to the CMDI standard had to be taken into account.

This especially includes an extended set of post processing procedures to allow extraction of more values from CMDI files with sometimes heterogenous data types. Most notably more popular vocabularies are supported for several facets and the extension of CMDI ResourceProxy types are reflected in the importer. The later changes formed the basis for a seamless connection with FCS (see Stehouwer et al., 2012) and the support for repository specific search and description pages in the frontend of the VLO.

Driven by extensive feedback by a variety of users (like from the CLARIN-D VLO taskforce) the selection and definition of facets were revised several times. As a consequence new facets were added and most configurations were subject to changes. This was supported by a first clear definition of all facets and a supportive tool that enables metadata creators to compare their CMDI profiles with the current configuration of the VLO[9]. One of the most relevant conclusions of these user requests was the demand for an improved quality of extracted facet values. By implementing a simple blacklisting approach (based on blacklisted XPath expressions) the amount of falsely extracted values could be reduced significantly. This was furthermore supported by several feedback rounds during developer sprints.

[8] `http://vtd-xml.sourceforge.net`

[9] Implemented by Menzo Windhouwer at the Max Planck Institute for Psycholinguistics, available at `http://lux13.mpi.nl/isocat/clarin/vlo/mapping`

Figure 2: The VLO 3.0 search page with 'Turkish Sign Language' selected on the *Language* facet and a free text search for 'narrative'. The second search result has been expanded to show the full description and more details.

3. Future work

The information available in the VLO is the result of a chain starting with the creation and publication of metadata for resources, collections, tools, etc., going through a process of semantic mapping and normalisation, and ending with a specific presentation of the results of this mapping. The end-user experience is determined by a large number of factors distributed over the individual elements in this chain. High quality metadata is the first prerequisite for delivering usable and effective search and presentation facilities. Recently effort has been put into improving the quality of the metadata being made available at the beginning of this chain through a standardised assessment mechanism (Trippel et al., 2014). While strongly related, the topic of metadata quality improvement is outside the direct scope of VLO development.

The second prerequisite is a proper mapping from the heterogeneous metadata realm to a relatively fixed set of facets. The extraction of meaningful and usable values out of the raw CMDI metadata has high priority. Important aspects are further post-processing of the derived values (i.e. providing some robustness with respect to low metadata quality), and incorporating the semantic context of values specified in metadata fields (e.g. a language code can appear in reference to the subject of a record, as its content or in its context). A more in-depth discussion of potential improvements with respect to metadata quality and the mapping process is provided by Haaf et al. (2014)

Recent developments on the VLO front end have focused on stabilising the code base and providing a solid base for future improvements. For example, the user interface can be refactored to make use of a standardised toolkit such as Bootstrap[10] to improve the user experience, visual identity and usability on a variety of form factors including portable devices. Making use of standard features available in Apache Wicket, themes and localisation modules (offering a user interface in the user's preferred language if available) can be added to the VLO relatively easily. Its look and feel could then be adapted to match a common CLARIN identity or that of any other community. Search efficiency on the user's end could be improved by a sophisticated ranking system determining the order in which search results are presented. Priority can be determined on basis of relevance to the query expressed by the user, as well as the nature of the records themselves (e.g. whether they represent individual resources or collections, see Haaf et al. (2014)). A further potential improvement in the front end would be the addition of an 'advanced' search option that allows the user to combine multiple values in a single facet in an *and* or *or* clause. A more dynamic selection of facets for display (e.g. depending on the active selection) is also conceivable.

In the near future, focus will also lie on implementing support for the upcoming CMDI 1.2 standard (Goosen et al., 2014), which primarily requires adaptations to the importer. On a final note, it should be understood that the VLO was not intended or designed to fulfil all search needs across the CLARIN metadata domain. While the VLO can, as described above, be improved in serving its original purpose, it is not likely to benefit from attempts to accommodate too many, potentially conflicting features. Rather, there is a potential for additional exploitation tools focussing on complementing qualities and providing for different needs within the CLARIN infrastructure.

[10] http://getbootstrap.com/

4. Acknowledgements

The authors wish to thank Kees Jan van de Looij and Dieter Van Uytvanck for their contributions to the recent developments on the VLO, and the numerous members of the CLARIN community that provided feedback on alpha and beta versions.

5. References

Broeder, D., Windhouwer, M., van Uytvanck, D., Goosen, T., and Trippel, T. (2012). CMDI: a component metadata infrastructure. In *Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR Workshop Programme*, page 1.

Goosen, T. (2014). VLO analysis. Technical Report CE-2014-0263, CLARIN ERIC, http://www.clarin.eu/content/vlo-analysis.

Goosen, T., Windhouwer, M., Ohren, O., Herold, A., Eckart, T., Ďurčo, M., and Schonefeld, O. (2014). CMDI 1.2: Improvements in the CLARIN component metadata infrastructure. In *CAC*.

Haaf, S., Fankhauser, P., Trippel, T., Eckart, K., Eckart, T., Hedeland, H., Herold, A., Knappen, J., Schiel, F., Stegmann, J., and van Uytvanck, D. (2014). CLARIN's virtual language observatory (VLO) under scrutiny – the VLO taskforce of the CLARIN-D centres. In *CAC*.

Stehouwer, H., Durco, M., Auer, E., and Broeder, D. (2012). Federated search: Towards a common search infrastructure. In *LREC*, pages 3255–3259.

Stoica, E., Hearst, M. A., and Richardson, M. (2007). Automating creation of hierarchical faceted metadata structures. In *HLT-NAACL*, pages 244–251.

Trippel, T., Broeder, D., Durco, M., and Ohren, O. (2014). Towards automatic quality assessment of component metadata. In *LREC*, pages 3851–3856.

Van Uytvanck, D., Stehouwer, H., and Lampen, L. (2012). Semantic metadata mapping in practice: the virtual language observatory. In *LREC*, pages 1029–1034.

Van Uytvanck, D., Zinn, C., Broeder, D., Wittenburg, P., and Gardellini, M. (2010). Virtual language observatory: The portal to the language resources and technology universe. In *LREC*, pages 900–903.