Title	CLARIN-PLUS VLO work plan
Version	1
Author(s)	Twan Goosen (CLARIN ERIC)
Date	2015-11-25
Status	Final
Distribution	Public
ID	CE-2015-0687

1 Planning

First phase

Year	2015			2016				
Month	**Oct**	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**
VLO 3.3								
Specification								
VLO 3.4								
VLO 4.0								

Second phase

Year	2016		2017					
Month	**Nov**	**Dec**	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**
VLO 4.1								
VLO 4.2								
VLO 4.3/Deliverable								

2 Work

The Virtual Language Observatory (VLO) faceted browser consists of two main components that are being developed: the *importer*, which creates indices based on harvested CMDI metadata records, and the *web application,* which presents this index and the underlying metadata to the user through a searchable and browsable interface. Both components communicate with a shared instance of a *Solr*[1] server, which manages the indices and handles queries on these indices, and has a tailor made configuration to accommodate the information required by the VLO components.

The VLO contains information derived from a set of metadata description documents (currently it indexes about 800,000 such documents) from a range of sources that can be considered quite heterogeneous in terms of origin, domain, method of description and quality. This heterogeneity poses particular challenges on the VLO in relation to its main objective of making the described resources discoverable and accessible via a unified interface. The ACDH team[2] has recently analysed[3] the VLO and its "periphery" and has identified a number of such challenges that are not or only partly addressed in the current version of the VLO, together with a number of potential solutions. As a result of this initiative, combined with the administration of concrete issues and potential

[1] http://lucene.apache.org/solr/
[2] Austrian Centre for Digital Humanities: Matej Durco, Go Sugimoto, Davor Ostojic, Meghan King
[3] https://docs.google.com/document/d/1rjNUqwr9KgUY4XLQiuvzpWIvb3zEIvza8PGvX93Bm8Y

enhancements, based on both user feedback and expert insight that has been maintained on the CLARIN Trac over the past couple of years, an extensive set of concrete features and enhancements to be implemented can be compiled. Section 2 of this document specifies the subset of this set to be addressed (implemented or, in some cases, further discussed and specified) in the time frame available in the context of CLARIN-PLUS task 2.4.2 with a final deliverable (in the form of a functioning software product) scheduled for May 2016.

Notice that not all features and enhancement described in this section will necessarily be implemented during the present CLARIN-PLUS task. Section 3.1 details the schedule in which the planned set of changes is listed per release. Section 3.2 lists the changes that we expect to get implemented at a later stage.

2.1 UX: front end style, layout and usability

The style and layout of the current VLO front end has evolved over the years, which has resulted in a sufficiently usable, but not overly user friendly or visually pleasing user interface. Moreover, it visually does not fit in with any other CLARIN application or website. This section lays out a number of changes to the VLO web application component to improve "user experience" and integration into the "CLARIN web".

2.1.1 Common CLARIN style

CLARIN has started to adopt a common visual style for their central web applications, currently implemented to some degree in the Centre Registry[4], FCS Aggregator[5] and Virtual Collection Registry[6]. Although it has to be noted that there are differences between these and the styles have not yet been consolidated into a single specification, a number of design principles have been established:

- Responsive design
- Usage of a limited number of specific colours that provide good contrast
- Usage of a uniform icon set and selection of fonts
- Fixed layout for page header and footer

The VLO can be made to adhere to these principles and the associated design specifics with relatively little effort, especially once a common style implementation has been established in the shape of a collection of CSS rules and/or (client side) scripts.

In addition to the style of the application, there is also the *layout* of the application interface and its constituents, i.e. the placement and structuring of interface components. Layout aspects of various parts of the application are treated in the next couple of sections. Many of the described enhancements are partially or wholly derived from the mockups by the ACDH team[7].

2.1.2 Search page layout and search result presentation

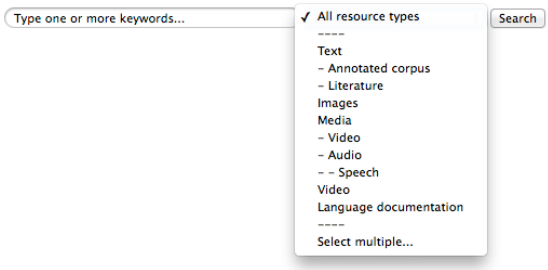
The main search page, with full text search box and facet value selectors, can be made easier to use, especially for the novice user, by following a number of "de facto standards" found in similar search interfaces found in for example web portals but also commercial web shops.

---

[4] https://centres.clarin.eu/
[5] http://weblicht.sfs.uni-tuebingen.de/Aggregator/
[6] http://clarin.ids-mannheim.de/vcr
[7] https://docs.google.com/document/d/1gO2EAzViLdloYrDkwyG6DW01WEo5kd9JRNHISCmxLyI

Full text search

The 'free search' box should be salient and easy to find and therefore appear on the top of the page and aligned centrally or to the right hand side. It could combine a "row" of the user interface with selectors for one (or possibly more) "primary" facet(s), in particular "resource type". The following mockup illustrates such a layout:



Facet value selectors

"Regular" facet value selectors are most commonly displayed on the left hand side of the page in faceted browsers. Therefore it makes sense to do the same thing in the VLO and move the "facet bar" from the right hand side to the left. In addition, these selectors will have to allow for multiple value selection and possibly the selection of a logical operator (see section 2.2.1).

Search results

Items in the search results can be made easier to interpret by visually presenting some of their properties by means of icons. Fields and properties that can be represented (at least in some cases) by icons are:

- Resource type (of the entire record)
- Media type of the referenced resources, with a counter (e.g. 1 audio file, 2 images)
- Whether the record is a collection, i.e. whether it is a leaf in a metadata hierarchy
- Licence and/or licence category[8]: icons indicating public/academic/restricted/unknown, and in some cases, such as CC, there may also be a (composite) icon for the specific licence[9]. Also see section 2.5.1.
- A manually or automatically derived quality/significance marker (e.g. 'recommended by CLARIN')

Occurrences of any of the search terms should be highlighted in the resource title as well as its description, as is currently the case on the record page.

2.1.3 Record presentation

The record page currently consists of a single, vertical concatenation of various aspects of a metadata record and the resources it describes/references. This should be replaced with a tabbed layout below a page heading with record title, link to the landing page ("original context") and a navigator (for browsing to the previous or next record in the search results). The actual record information should be distributed over a number of "tabs":

- Basic information (initial selection): description and all other fields currently shown in the first properties table of the record page, and an overview of the described resources (without details). For images, and possibly also for videos, it

[8] See https://www.clarin.eu/content/license-categories

[9] https://commons.wikimedia.org/wiki/Category:Open_Icon_Library_set_of_license_icons provides some icons that could potentially be used

would be possible to show thumbnails in those cases that the resources are accessible and the licence permits.

- Resource details: a paged, tabular listing of referenced resources with available details such as mime type.
- All metadata: a (user friendly) rendering of the entire Components section of the CMD record
- Technical details: all fields currently shown in the "technical details" table at the bottom of the record page

The hierarchical tree (see section 2.6) should be accessible via a separate pane on the left hand side that can be shown or hidden (user's choice should be remembered, initially shown).

Basic citation information for the current record should be provided. Due to the heterogeneity of the metadata, exact citation instructions cannot be derived automatically. However, citation information should be available at the landing page, which can be mentioned explicitly. The PID (if available) should be mentioned but also the potential caveats (i.e. it may not resolve to a human friendly representation) and the information at the original context should be leading.

2.1.4 Help and guidance

The entry page, which has no search results or facet selection boxes stays, but will provide more guidance to the novice user. It will match the new layout of the search page (see 2.1.2) and have a brief, descriptive introduction with basic usage instructions.

It can provide some "recipes" for searches that can serve as an example but also should cover common search scenarios. These recipes should be presented as answers to the question "What are you looking for". Some examples:

- "German annotated corpora"
- "Audio recordings of Turkish"
- "Written resources from Ethiopia"
- "Something else? Type your keywords in the search bar or click here to browse all resources"

On the entry page and throughout the application more guidance can be provided by means of contextual help in the form of tooltips. In particular cases the availability of such guidance can be made explicit by adding icons or question marks to user interface components (as is already the case for the search bar).

The use and intention of the "link" and "report" options should be made clearer. The former can be better framed as a "share" option with an appropriate icon and various sharing options (including e-mail, social media), while the latter can be rephrased (e.g. as "feedback") and also augmented with an icon.

2.2 Facets

2.2.1 Multiple selection/Logical combination

Facets allow the user to narrow down on the set of records shown in the browser, but selecting a single value for a facet can sometimes cause the result to be 'too narrow'. For example, users may be interested in texts in Bulgarian or Macedonian, but this cannot be expressed by means of a single facet based search. The technology underlying the VLO however does support these kinds of queries and it is very common among faceted browsers throughout the web. To match the user's needs and expectations, the VLO

should be extended with an option to select more than one value for any facet, which then is to be interpreted as an OR selection within that facet – selections between facets should remain AND.

The option to also allow AND selections within facets can be considered, but such an option should not reduce ease of use or cause confusion in any way. Performing an AND search within fields is already possible by using the features of the Lucene query syntax, and could also be made available via an advanced search form (see 2.3).

2.2.2 Facet display and conditional facets

The number and precise selection of facets displayed in the VLO's search interface has been the subject of debate for a while, and we have seen proposals for both the addition of new facets and the removal of existing facets[10]. In fact, there is a legitimate need for more (more ore less domain specific) facets while at the same time displaying irrelevant or simply too many facets reduces usability. The proposed solution is to make the display of specific facets conditional, i.e. only show a facet selector for a facet if some condition is met. Several types of such conditions can exist:

- Selection of any value from a specific facet (e.g. only show 'subject' after selection within 'collection')
- Selection of a specific value from a specific facet (e.g. only show 'annotation type' after selection of resource type 'corpus')
- Number of remaining value options (e.g. only show 'genre' if the number of choices is down to 50)

There can still be the option to show all facets (at least in some advanced search mode), but the 'hidden' facets should initially not be in plain sight. Some facets (primarily technical ones) should only appear in the view of all facets, i.e. should be considered 'hidden facets' rather than 'conditional'.

2.2.3 Hierarchical facets

In addition to not displaying all available facets at once, there is also the possibility of not displaying all facet *values* for any one facet at once. For some facets, a more or less natural organisation into (multiple levels of) categories can be made. A good example is the current 'format' facet, which can be divided up into the groups 'video', 'audio', 'image', etc. according to the mime type prefix. Presenting facet values this way will a) make it easier for users to find relevant values and b) open up the possibility to perform a broader search on the union of all subordinate values, for example a search for all records with a resource of any 'video' format.

It has to be noted that useful facet value hierarchies can only realistically be achieved as a result of manual creation and more or less continuous curation, and even then can only be applied to facets that have a fairly restricted vocabulary, or a strict pattern to them such as in the case of mime type. The output of the CLARIN metadata curation task force however promises to make such curation work easier, and vocabularies may be derived from taxonomies within CLAVAS – the list of known licences by type (pub/aca/res) might make a good first use case – also see section 2.5.

2.2.4 New facets

A number of new facets have been proposed and can be implemented by adding fields to the Solr configuration and adding new mapping rules based on concept links:

[10] Notably by the CLARIN-D VLO working group, Jan Odijk, and the ACDH team

- Temporal coverage, originally proposed by the CLARIN-D VLO WG[11], will focus on date/time values and ranges, with support for different levels of specificity. It will need a dedicated value selection mechanism that allows for the selection of ranges Not all types of temporal information can easily be mapped (e.g. 'middle ages', '20th century', 'pre-Columbian') so probably only well specified source information will be taken into account.
- Linguistic annotation type, as proposed by Jan Odijk[12], indicates what kinds of annotations are present in the described data. Odijk proposes a taxonomy based on the linguistic level, which would make this a suitable case for a hierarchical facet (see above). Examples of potential values are 'phonetic transcription' within 'phonology/phonetics', 'pos-tags' within 'morphology and 'synonyms' within 'semantics'. This facet should not be presented to the user unless a relevant resource type has been selected.
- Actor, proposed by the ACDH team as part of their VLO recommendations, has relatively broad semantics and can take authors, researchers, creators, subjects and (other) contributors as values. It should probably not be one of the 'primary' facets, i.e. only show up by default when relevant, for example when only relatively few values remain.

2.3 Full text search

Up and until version 3.2 of the VLO, the 'full text search' queries would take the entered value as a literal phrase. This differs from the usual 'search box' behaviour where each token is interpreted as a distinct search term unless explicitly grouped. The VLO will be reconfigured to adopt this more common strategy, which should lead to higher recall, while improved ranking (see below) should mitigate the loss of precision. To be precise, the full Lucene query syntax[13] will be supported, so that users can compose more advanced queries if they like. A number of fields have unfriendly internal names, so aliases will be configured (e.g. language -> languageName).

To make advanced querying more accessible, an advanced search form can be implemented that allows users to compose a query by adding terms, selecting fields to search in and combining these elements by means of logical operators.

The 'suggestion' or 'auto complete' functionality of the search box can be improved by also providing common multi-word phrases extracted from the index. Furthermore, it could help the user select field values by providing 'content first' suggestions where the user can select both a value and a field. For example, when typing "German" a user could be presented with the option to search for "language:German" to exclude results that just contain the word German if this matches the user's search needs – there should of course still be the option to perform a full text search.

Instead of sorting the results alphabetically by title, the default sorting should be by relevance. Solr can perform a standard *tf-idf* ranking and on top of that weights can be assigned to certain fields (in relation to the search term) or to document properties. The following 'boosts' will be applied:

- Match in record name (exact phrase match preferred)
- Match in description (exact phrase match preferred)
- Hierarchy level of the record (collections first)

[11] In http://www.clarin.eu/sites/default/files/cac2014_submission_30_0.pdf
[12] In http://www.clarin.nl/sites/default/files/Searching%20with%20the%20VLO.pdf
[13] See https://lucene.apache.org/core/2_9_4/queryparsersyntax.html

- Number of collection members (limited)
- Presence of a record name or description

The users can be given the option to sort the results by title instead.

When the returned result set is empty, the VLO could propose a similar, common search term if available (e.g. when searching for 'Yurucare': "No results found. Did you mean to search for 'Yuracare' instead?").

2.4 Metadata quality

The CLARIN metadata curation task force is building a workflow for on-going metadata curation on basis of post-hoc value mapping (complemented by feedback to the metadata provider). This mapping is a step beyond the value post-processing that is currently implemented in the importer and will make use of new data structures that can be edited by curators in a workflow that is detached from the development process. The importer will be adapted to process the resulting definitions. Enhanced facet values are a prerequisite for a number of the proposed VLO interaction improvements, in particular those depending on a clean and structured set of resource type and licence values (see sections 2.1 and 2.5).

2.5 Licence information for search and display

2.5.1 Presentation

Many metadata records contain licence information, either pertaining to the entire record or to individual resources referenced from the record. As of VLO 3.2 a record's licence is mapped to a value in the 'availability' facet and the original licence value(s) are shown in the record page.

It has been noted that it is too easy for users to access a resource while (unwillingly) ignoring its licence. Therefore the licence information should be displayed close to the resource links in the record page, and in case the access by the licence is non-public, this should be clearly indicated with a link to the licence information of the original provider if available.

Licence information of records should also be indicated in the search results where applicable, preferably by means of icons, as described in section 2.1.2.

2.5.2 Faceted search

As mentioned directly above and in section 2.2.3, searching for resources with a specific licence (category) can be much improved by, first, better mapping and curation on this facet and, strongly depending on this, improved presentation of the available options by showing them in a hierarchy based on licence categories. A categorised list of known licences has been assembled within CLARIN[14], which can serve as the basis of these functionalities.

2.6 Metadata hierarchies

CMDI records can refer to other CMDI records, thus constituting a metadata hierarchy. These hierarchies are used by a number of metadata providers (in varying ways) to create collections and sub-collections. It can be very useful to a user of the VLO to be aware of such hierarchies and the position of a search result item in such a hierarchy, especially if the option is provided to traverse these hierarchies. Such a feature is planned for inclusion into the record page of the VLO. Hierarchies will be rendered as interactive trees that can be expanded into arbitrary depth, and via links in this tree any record at a higher or lower level than the record at hand can be navigated to.

[14] https://www.clarin.eu/content/license-categories

Based on the observation that, given a search term, collection records are generally more interesting to users of the VLO that 'leafs', collections that are 'higher up' in the hierarchy they belong to will also receive a higher ranking, and therefore will be displayed first in the search results.

2.7 Infrastructure integration

The VLO functions, to a degree, as a portal into the world of CLARIN (and broader DH) resources. Therefore it is natural for the VLO to integrate with other CLARIN services that operate on the resources that can be discovered via the VLO.

The Virtual Collection Registry (VCR) can be tied in with the VLO user interface, which would provide a means to submit a search result set to a new or existing collection in the user's VCR workspace. A pilot for this has already been developed but not yet merged into a released version of the VLO.

Another service that can be tied in with the VLO is the Language Resource Switchboard, which is currently under development. This service, once completed, will allow users to process, view resources found in one application in any other application within the CLARIN infrastructure that is registered in the switchboard and supports the type of resource at hand. Integration with the VLO comes down to the addition of a user interface component to the record page that passes the PID of the record and/or the PIDs of the individual resources to the switchboard endpoint, which can then guide the user into selecting an application

At a different level of the VLO, namely the importer, integration can be facilitated by detaching the mapping facilities that are built into the importer from the Solr ingestion process, so that other tools and services can easily 'hook into' this mapping logic. This requires some refactoring of the VLO importer so that a *metadata mapping library* can be 'distilled'.

Finally, given the central role that the VLO plays in CLARIN's metadata infrastructure, it must be adapted to support CMDI 1.2 from the moment it is released[15]. This means that the importer needs to be adapted to support the new paths and namespaces, and the same goes for the stylesheet used in the record page. In later versions, some of the features that are new in CMDI 1.2 can be harnessed to improve the mapping and display of metadata records, in particular in those cases where external vocabularies (from CLAVAS) are used, which allows for easier and safer unification of equivalent values within a domain, even across languages.

2.8 Statistics

Secondary to the front facing features of the VLO, the application can provide a lot of valuable information about its content and usage that can, among other things, provide insight into the state of the 'metadata world' and help improving the usability of the VLO.

Anonymous usage statistics are already being collected by means of Piwik[16]. The 'reporting' to Piwik is currently limited to non-Ajax requests, which leads to an incomplete and possibly skewed picture of, for example, the overall search behaviour of the VLO's users. This can be improved by utilising the Piwik API in the context of Ajax based partial page updates.

[15] See https://www.clarin.eu/content/cmdi-12-changes-executive-summary and the CMDI 1.2 implementation plan (currently being written)
[16] https://stats.clarin.eu

Furthermore it would be useful to periodically gather statistics on the state of the index, so that the development of metrics such as number of included records, resources, languages, collections etc. can be tracked over time.

2.9 Performance and scalability

Although the performance of the VLO components thus far has proved to suffice, some improvements can be made to ensure the stability and responsiveness of the VLO as the user base and record count grow.

First, there are some latent improvements to the querying behaviour between the VLO web application and the Solr database. In particular, redundant queries could be eliminated, which would reduce load on both the front end and Solr instances. Second, there is the potential of importing multiple collections in parallel, which would greatly improve scalability of the import process, which may become relevant when large library collections get incorporated into the VLO, as is likely to happen in the near future.

3 Roadmap

3.1 VLO releases (first phase)

Three releases are planned during the first development phase (October 2015 – May 2016). The final release during this stage will have a redesigned front end and will be marked as a new major version (4.0) of the VLO.

The following distribution of changes described in this document over these releases is foreseen:

VLO 3.3 (November 2015)

Focus: advanced search, result ranking, metadata hierarchies

- Support for Lucene query syntax
- Result boosting on basis of field matches and document properties
- Browsing of metadata hierarchies on the record page

VLO 3.4 (February 2016)

Focus: improved display and mapping of licencing information, and support of metadata curation data structures

- Better mapping to the 'availability' facet
- Better mapping to the 'licence' field
- Better display of licence information on the record page
- Improved performance by more sophisticated querying behaviour

VLO 4.0 (May 2016)

Focus: user interface and service integration

- Usage of the common CLARIN style (expected to be available March – April 2016)
- Revision of search page
- Tabbed record page
- Integration with the Language Resource Switchboard

- Integration with the Virtual Collection Registry
- Extraction of a mapping library
- Support for CMDI 1.2

3.2 Further VLO development (second phase)

During the second development phase (November 2016 – June 2017), another three releases are foreseen. The distribution of features and other changes over these releases will be determined after the completion of the first development stage.

We expect to implement the following changes described in this document during this phase:

- Multiple facet value selection
- Resource type selector on search box
- Conditional facets and hidden facets
- Hierarchical facets
- Support for external vocabularies (CMDI 1.2)
- Search examples on entry page
- Custom search result sorting
- Icons in search results (licence type, resource type, …)
- Thumbnails for publicly accessible images and videos
- Addition of a 'Temporal coverage' facet
- Improved gathering of statistics
- Parallelised imports

Depending on feasibility and priority (to be assessed), the following points will also be addressed:

- Addition of a 'Linguistic annotation type' facet
- Addition of an 'Actor' facet
- Boosting on basis of metadata quality score
- "Did you mean…?" in case of empty result set
- Support for 'visualisation cues' (CMDI 1.2)

3.3 Related work

- The CLARIN metadata task force is responsible for the development of a set of instruments for benchmarking and improving the metadata quality in the CLARIN infrastructure. Much of this work takes place in the periphery of the VLO, and partly ties in with the development thereof as is described in this document. See also CLARIN-PLUS task 2.2.1. Initially there will be a focus on post-hoc curation of certain facets, in particular resource type and licence/availability. Ultimately the goal is to have a continuous, largely automated workflow in which metadata records and their mappings can be assessed and corrected, and metadata curators, providers and experts can get insight into these processes, leading to a more usable and sustainable metadata ecosystem. The development of guidelines for the creation and curation of CMDI metadata is also part of this task.

- In a related task (see CLARIN-PLUS 2.2.2), the OAI harvester, which provides input for the VLO importer, will be improved in terms of flexibility, scalability and robustness. This should result in the possibility to run more frequent imports, leading to an index that reflects the current state of the metadata infrastructure more accurately. Feedback towards the metadata providers should also be improved, which is expected to contribute to higher metadata quality as well as more stable state of the VLO.
- Work on the next version of the component metadata infrastructure, CMDI 1.2, will be completed in the first half of 2016. This version has a number of new features that are expected to contribute to improved metadata quality. Next to the Component Registry, the VLO will be one of the first publicly accessible infrastructure components to support this new version of CMDI.