



SSHOC

social sciences & humanities open cloud

Summary

Use Cases for the Virtual Collection
Registry in SSHOC

September 2021

SSHOC, "Social Sciences and Humanities Open Cloud", has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-04-2018, Grant Agreement #823782.

About this document

This document is meant to provide concise information on the Virtual Collection Registry (VCR), particularly with regard to the use cases and its extensions and improvements planned and created during the Social Sciences and Humanities Open Cloud project (SSHOC).

Note that this document is largely a summary from a larger document¹ that holds the common wisdom of the SSHOC Task 3.6 partners with regard to the VCR goals. It serves as documentation and proof-of-work for Task 3.6 from 2020 to 2022 and may inform future discussion on service integrations revolving around the VCR.

In SSHOC task 3.6 the main working goal is to generalise the VCR for use by the SSH community at large and ideally implement a number of SSH use cases with potential to grow e.g. to be taken up by other data-catalogues and services. This acknowledges the language-/ linguistics-related focus of the VCR so far (as a CLARIN service). SSHOC serves as a suitable framework for extending the VCRs use case from language/linguistics to other research areas.

Author List*

Organisation	Name	Contact Information
CLARIN ERIC	Daan Broeder Willem Elbers	d.g.broeder@uu.nl willem@clarin.eu
UGOE	Stefan Buddenbohm	sbudden@gwdg.de

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability.

*The document also reflects the contribution of many SSHOC colleagues that have been involved in discussing the VCR use-cases.

¹ Available as project internal working document "SSHOC Virtual Collection Registry".

Contents

[Introduction](#)

[VCR development and use in other projects, past and present](#)

[VCR developments and use in the context of SSHOC](#)

[Operating a SSHOC VCR](#)

[State of SSHOC VCR activities and planning](#)

[Further VCR outreach activities](#)

[EOSC Portal/Marketplace](#)

[Factsheet](#)

[Discussion with other SSH infrastructures](#)

[Tutorials](#)

[Discussions conducted with \(SSHOC\) communities](#)

[VCR General Use Cases](#)

[Detailed SSHOC VCR use cases from CESSDA](#)

[Detailed SSHOC VCR use cases from SSHOC 5.2 \(Dataverse\)](#)

[Detailed use cases from E-RIHS v0.8 \(Gen 2021\)](#)

[Detailed use case from SSHOC Marketplace](#)

[Detailed use cases from DARIAH-DE/CLARIAH-DE](#)

[The DARIAH-DE Data Federation Architecture and its Collection Registry](#)

[CLARIAH-DE/ DARIAH-DE DFA and the CLARIN VCR](#)

[Analysis & comments of the use cases](#)

[VC creation for bookmarking, sharing and publication; the classical scenario](#)

[Making VCs interoperable with other services](#)

[Saving query results in VCs](#)

[Conclusion](#)

[References](#)

Introduction

The Virtual Collection Registry (VCR) is a long-standing CLARIN service and is being actively maintained and enhanced by CLARIN ERIC, who maintains ownership over the service. This continued ownership of the VCR stands in no conflict to its enhancement, e.g. by additional use cases from other partners or new functionalities implemented within SSHOC. Wherever appropriate it will be made clear in which relation new developments such as for use cases that were suggested by the SSHOC partner infrastructures stand to other entities, e.g. by a label "Powered by SSHOC". This is also important against the background of the VCR as accepted EOSChub thematic service. If anything, the continued ownership of the service by CLARIN is beneficial as it underpinned the discussion with experience, commitment and some predictability.

Documentation on the VCR is available from various sources, addressing various audiences. A general overview document², a [CLARIN short guide](#)³, an updated factsheet⁴ by SSHOC and [research-oriented small tutorials, which may be easily updated via GitHub](#)⁵. Some of these information sources have been updated during the SSHOC project to adapt further audiences. The source code for the VCR is available on GitHub⁶ along with the technical documentation⁷. Most of the linked resources follow a transparent and inclusive approach to invite researchers and institutions to make use of the VCR or even take part in its further development. An emphasis for enhancing the uptake of the VCR could be laid upon targeting specific scientific disciplines.

The CLARIN VCR, that was originally developed by the Institute for the German Language⁸ (concept and operations) and CLARIN ERIC (further development and maintenance), allow users to create collections of resources possibly each hosted by a different repository and describe the collection with descriptive metadata. It allows to indicate both 'intensional' and 'extensional' collections and also different modes/policies wrt. the collection's reproducibility. In a nutshell the VCR allows the researcher to create a persistent set of links to digital objects, which may have varying formats, e.g. research data sets, tutorials, even book chapters. A specific virtual collection may be used as a persistent documentation underlying a research publication or it may describe a tutorial highlighting the use of a certain service along with exemplary research data and documentation. An obvious use is for citing all the different data objects that underlie a research publication and that do not already have a single citable description. A PID (both an EPIC Handle as a DOI) is automatically issued for every created virtual collection for just that purpose. Various use cases are possible and especially those discussed in the context of the SSHOC project are described in this document.

² Available as project internal documentation.

³ https://www.clarin.eu/sites/default/files/virtual_collections-CLARIN-ShortGuide.pdf

⁴ Elisa Gorgaini. (2021). Virtual Collection Registry - SSHOC Service Catalogue's Factsheets Series (v1.0). Zenodo. <https://doi.org/10.5281/zenodo.5482899>

⁵ https://github.com/subugoe/factsheets_sshoc_services

⁶ <https://github.com/clarin-eric/VirtualCollectionRegistry>

⁷ <https://github.com/clarin-eric/VirtualCollectionRegistry/blob/master/doc>

⁸ Leibniz Institut für Deutsche Sprache Mannheim (IDS), <https://www.ids-mannheim.de/>

VCR development and use in other projects, past and present

The CLARIN VCR was further developed in the CLARINplus project⁹ and the EOSC-hub project (2018-2020). The CLARIN VCR is registered in the EOSC Marketplace¹⁰ and EOSC catalogue¹¹ and is an EOSC-hub thematic service¹². One can even go beyond and note that the VCR technically doesn't have to be limited to the SSH research domains but may be useful for other research domains as well.

Currently the CLARIN VCR is, or is planned to be integrated^{13,14} with the following applications:

- The Virtual Collection (VC) metadata is available via OAI-PMH and harvested by the CLARIN VLO and the current EOSC interdisciplinary metadata catalogue "B2FIND"
- The CLARIN VCR accepts authentication of users via the CLARIN Service Provider Federation. *Access via one of the EOSChub authentication options is planned*
- Integration with general data catalogues/ repository systems e.g. B2SHARE (planned)
- Integration with (cloud) storage solutions facilitating copying collection content data into a user's personal workspaces e.g. B2DROP. This is under discussion
- Integrated with the CLARIN Switchboard¹⁵
- Integrated with the CLARIN VLO¹⁶

The status of the EOSC registration for the VCR implies a production level service (Technology Readiness Level of TRL 8 minimally). Still some code restructuring and improvements for improved maintainability and efficiency and some functional enhancements are on the VCR development wishlist. A current summary of these is:

- a simplified better internal data-model
- a simplification of the workflow to create new collections
- the use of ORCID's identifying authors/researchers
- the possibility to enhance VCs after their initial publishing; e.g. dynamic or versioned VCs
- add support for the RDA collections API / fair digital objects (FDOs).

Note that the items on the wish list are often inspired by previous discussions within SSHOC and not considered currently part of the CLARIN ERIC own VCR work plan. Also the wish list is a snapshot being added to whenever new use cases are identified.

VCR developments and use in the context of SSHOC

In the SSHOC DoA (description of task 3.6) a requirement to create an implementation of a SSHOC VCR is tasked for the work package: "A SSHOC virtual collection-registry inspired

⁹ <https://cordis.europa.eu/project/id/676529/de> (2015-2017)

¹⁰ <https://marketplace.eosc-portal.eu/services/virtual-collection-registry>

¹¹ https://catalogue.eosc-portal.eu/service/clarin-eric.virtual_collection_registry

¹² <https://office.clarin.eu/v/CE-2018-1175-EOSC-hub-task71-roadmap.pdf>

¹³ Integration is also being used in the sense of being aware of other applications, thus allowing easy copying of resource references into VCs or using VCs

¹⁴ The VCR integration specification is available as project internal documentation.

¹⁵ <https://switchboard.clarin.eu>

¹⁶ <https://vlo.clarin.eu>

by similar components such as the CLARIN VCR will be integrated with the Switchboard and remember previous data selections allowing (a) populating a personal workspace of the researcher and (b) to share data selections with fellow researchers or broader/public audience. Such virtual collections can be made persistent and published using PIDs and the SSHOC citation recommendations developed in T3.4.” The time span for this task stretches from 2020-2022.

But from the broader SSHOC and task 3.6 descriptions and discussions we see, next to the very specific use case above, a requirement for developing use cases for the VCR and integration opportunities for instance the possibility to invoke the VCR from other applications in the other SSHOC partner infrastructures: CESSDA, DARIAH and E-RIHS. As hinted above it proved useful for the discussion to be held among established ERICs other than conventional project partners. Taken CLARIN and DARIAH alone, both research infrastructures come with a rich experience of developing and providing services and share a common history of cooperation. Furthermore these ERICs represent certain disciplines and are able to discuss use cases adapting to discipline-specific requirements better than just individual institutions or temporary projects.

For identifying such use cases we have scheduled specific discussions with the SSHOC partners. The results of which are available in the chapters below.

Originally also on the basis of the available resources and interests in the 3.6 task, there two partners were available to contribute development work: CLARIN ERIC and UGOE. UGOE showed an interest in implementing the RDA collection registry API¹⁷ for the VCR, unfortunately in Q2 2021 we had to abandon that plan due to non-availability of developer resources at UGOE and discuss other extensions on the basis of available developer resources from our EKUT partner.

Next to task 3.6 in the SSHOC project itself, there are further opportunities for generalising the CLARIN VCR, and making it more useful for the SSHOC stakeholders, for instance through efforts of WP2 and WP6 aiming at an enhanced visibility of the VCR in the EOSC portal or by adding new information and training material.

Some topics that should be pursued with 3.6 resources or on the basis of collaboration with other SSHOC tasks and external projects are:

- Integration of the VCR data in the CESSDA, DARIAH and E-RIHS data discovery mechanisms for instance by harvesting the VCR metadata as is currently the case with the CLARIN VLO.
- Collaborating with task 5.2 integrating the VCR in the new SSHOC Dataverse software for instance as a ‘shopping basket’ for search results.
- Collaborating with SSHOC task 3.4 making the VC citable according to SSHOC recommendations (to be developed by 3.4).

Based on the use cases provided by the different SSHOC communities - CESSDA, E-RIHS, DARIAH - so far, we can identify a clear need for more dynamic collections implemented through a versioning approach. Another shared requirement is functionality for collaborative

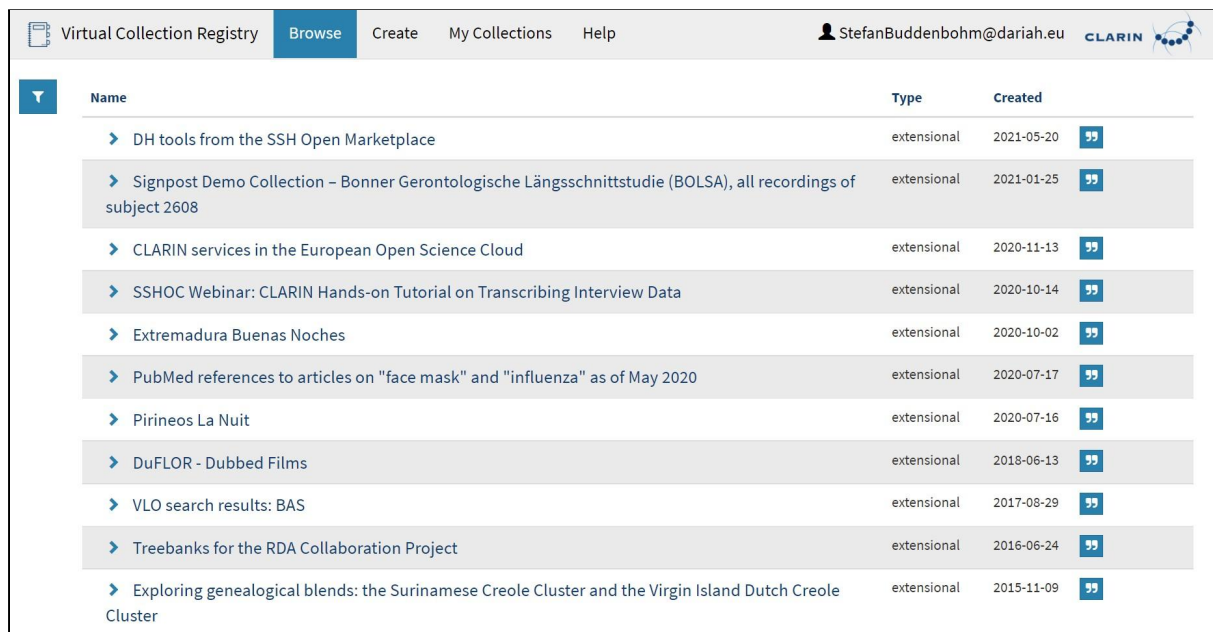
¹⁷ <https://www.rd-alliance.org/api-requirements-document>

editing. This can be implemented for private collections that are still being improved. For published collections, referenced by a persistent identifier and thus citable, a solution for versioning must be available. This implies that editing a published collection will result in a new, independently citable, version of that collection. The major shared use case however is the use of VCs to store repository search results cross-domain and cross- repository. Integrating the VCR in a number of SSH repositories should demonstrate the VCR use for researchers that work with different repositories or even cross SSH domain.

All SSHOC related requirements are grouped under a GitHub milestone:
<https://github.com/clarin-eric/VirtualCollectionRegistry/milestone/9>

Operating a SSHOC VCR

Currently the CLARIN VCR is operated by CLARIN ERIC and hosted by the IDS (German CLARIN centre). CLARIN ERIC plans to keep operating a stable version that may also be useful for other domains. Other versions supporting new (experimental) functionality can be operated by others. In the discussions with CESSDA it was pointed out that there should be clear solutions and guarantees wrt operations of the SSHOC VCR before CESSDA can commit using it¹⁸.



Name	Type	Created
▶ DH tools from the SSH Open Marketplace	extensional	2021-05-20
▶ Signpost Demo Collection – Bonner Gerontologische Längsschnittstudie (BOLSA), all recordings of subject 2608	extensional	2021-01-25
▶ CLARIN services in the European Open Science Cloud	extensional	2020-11-13
▶ SSHOC Webinar: CLARIN Hands-on Tutorial on Transcribing Interview Data	extensional	2020-10-14
▶ Extremadura Buenas Noches	extensional	2020-10-02
▶ PubMed references to articles on "face mask" and "influenza" as of May 2020	extensional	2020-07-17
▶ Pirineos La Nuit	extensional	2020-07-16
▶ DuFLOR - Dubbed Films	extensional	2018-06-13
▶ VLO search results: BAS	extensional	2017-08-29
▶ Treebanks for the RDA Collaboration Project	extensional	2016-06-24
▶ Exploring genealogical blends: the Surinamese Creole Cluster and the Virgin Island Dutch Creole Cluster	extensional	2015-11-09

Figure 1: Virtual Collections for a registered user (<https://collections.clarin.eu/>)

This also touches governance aspects that are discussed in SSHOC WP8.

The same applies for other services which are developed within SSHOC as well. The SSH Open Marketplace, for instance, as a new service will have to find an answer for governance and sustainability after the end of project funding. With the VCR matters are easier as ownership has in the past been successfully maintained by IDS and later CLARIN. This

¹⁸ This has now been cleared with CLARIN ERIC and a commitment was provided that within limits of reason CLARIN ERIC will operate a VCR meeting the requirements of current SSHOC use-cases.

“easier situation” could also be used as a leverage to interest other institutions, such as CESSDA, for the VCR.

State of SSHOC VCR activities and planning

Further VCR outreach activities

The VCR is already a stable and mature service, but lacks the uptake it could have in view of the high potential the VCR has for allowing users to work with resources crossing the boundaries of single repositories and communities ie. being used for cross-repository & cross-community data management. Insufficient exposure is one important reason and a similar situation can be found, for instance, when looking at services of other ERICs, for instance the DARIAH-DE Repository. The limited disciplinary scope and ownership of such services are on the one hand advantageous for the sustainability and product maintenance because they usually are integral parts of broader infrastructure portfolios.

But it should be in the interest of any research infrastructure to explore further usage potentials of its offerings. Of course the disciplinary scope is a limiting factor but interoperability or cross-domain collaboration could be pursued in projects like SSHOC. SSHOC as the social sciences and humanities contribution to the European Open Science Cloud proved useful in this regard. The project, and particularly work package 3, was a fruitful discussion arena for the SSH ERICs to explore linkages, service integrations and exploiting other research and data domains.

However such activities rely on dedicated resources for communicating, networking and disseminating. Often these resources are limited for sustained research infrastructures, which lay - naturally - a focus on sustaining and maintaining the portfolio from an infrastructural and technical point of view.

Therefore SSHOC addresses this desideratum with measures to leverage the visibility and usage of the VCR in the SSH research communities but also beyond.

EOSC Portal/Marketplace

This platform serves as a central dissemination hub for resources and services affiliated or funded within the EOSC framework. Together with other services, such as the Switchboard or the SSH Open Marketplace, the VCR has been onboarded¹⁹.

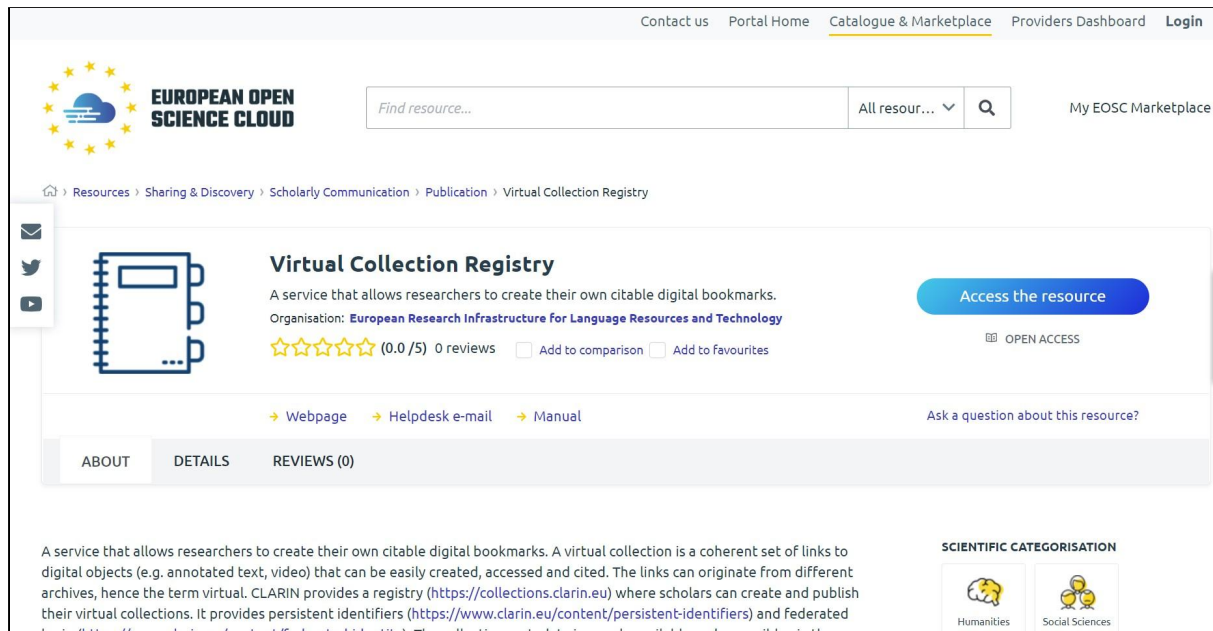


Figure 2: VCR as onboarded service in the EOSC Portal
(<https://marketplace.eosc-portal.eu/services/virtual-collection-registry>)

Factsheet

Together with WP2 and LIBER a new information primer²⁰ has been created. This so-called factsheet is useful as an add-on for the EOSC portal but may be used for other purposes as well, even as a print version handout.

Discussion with other SSH infrastructures

To elaborate further use cases for the VCR intensive discussion with other SSH infrastructures have taken place, respectively with CESSDA, DARIAH, and E-RIHS. Part of the results is documented in this paper as use cases.

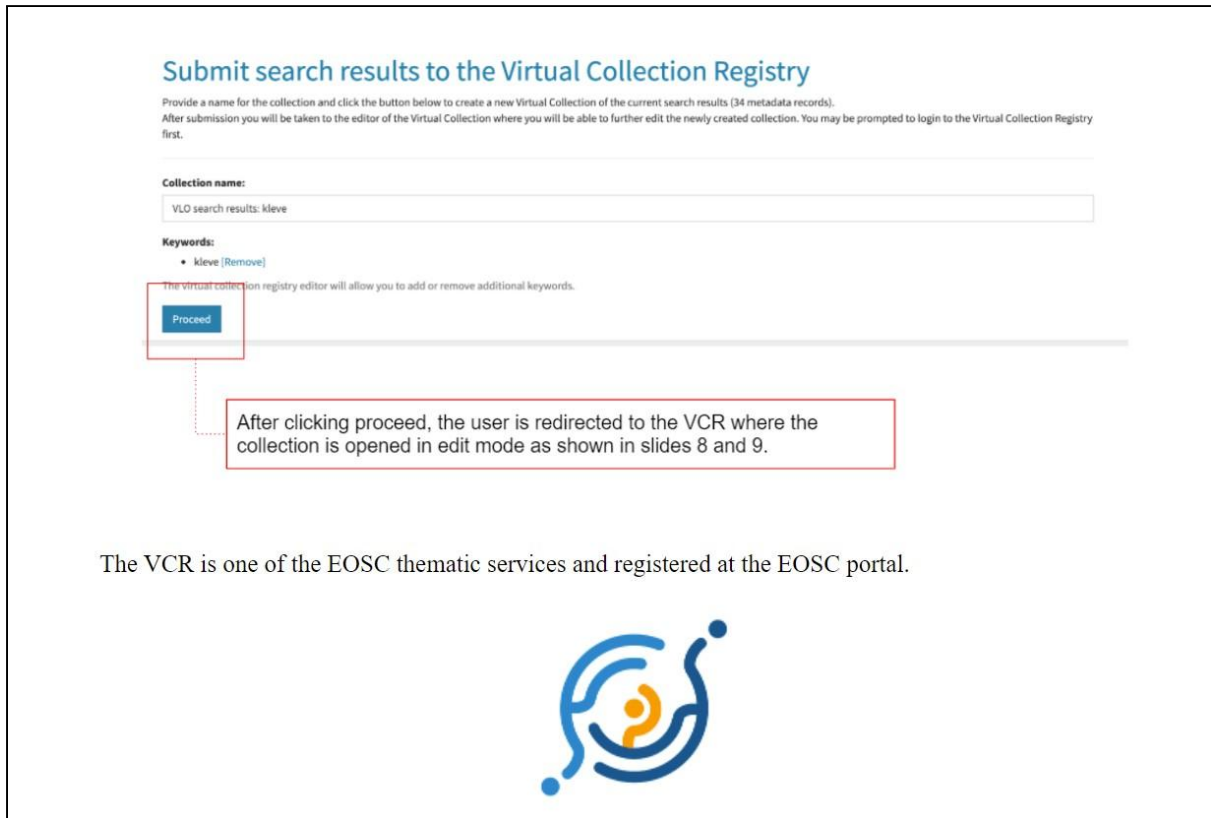
Tutorials

Varying levels of knowledge about services and resources have to be reflected by the information and support material. Whereas the EOSC Portal and the already mentioned factsheet address a very basic level of information and may be best described as

¹⁹ <https://marketplace.eosc-portal.eu/services/virtual-collection-registry>

²⁰ Elisa Gorgaini. (2021). Virtual Collection Registry - SSHOC Service Catalogue's Factsheets Series (v1.0). Zenodo. <https://doi.org/10.5281/zenodo.5482899>

dissemination or outreach information, a tutorial or FAQ service at the service's website addresses a more research case-oriented visitor. Therefore an easy-to-use and easy-to-edit and sustainable tool set has been created. It is based on easy-to-edit markdown files hosted in a GitLab repository (currently at the University of Göttingen). The markdown files have HTML representations, which offer a convenient and illustrative presentation of the individual tutorial. For the VCR such a HTML representation may look like this: https://subugoe.github.io/factsheets_sshoc_services/virtual_collection_registry.html



Submit search results to the Virtual Collection Registry

Provide a name for the collection and click the button below to create a new Virtual Collection of the current search results (34 metadata records). After submission you will be taken to the editor of the Virtual Collection where you will be able to further edit the newly created collection. You may be prompted to login to the Virtual Collection Registry first.

Collection name:
VLO search results: kieve

Keywords:
• kieve [Remove]

The virtual collection registry editor will allow you to add or remove additional keywords.

Proceed

After clicking proceed, the user is redirected to the VCR where the collection is opened in edit mode as shown in slides 8 and 9.

The VCR is one of the EOSC thematic services and registered at the EOSC portal.




Figure 3: Screenshot of a Github-/Markdown-based tutorial for the VCR (https://subugoe.github.io/factsheets_sshoc_services/virtual_collection_registry.html)

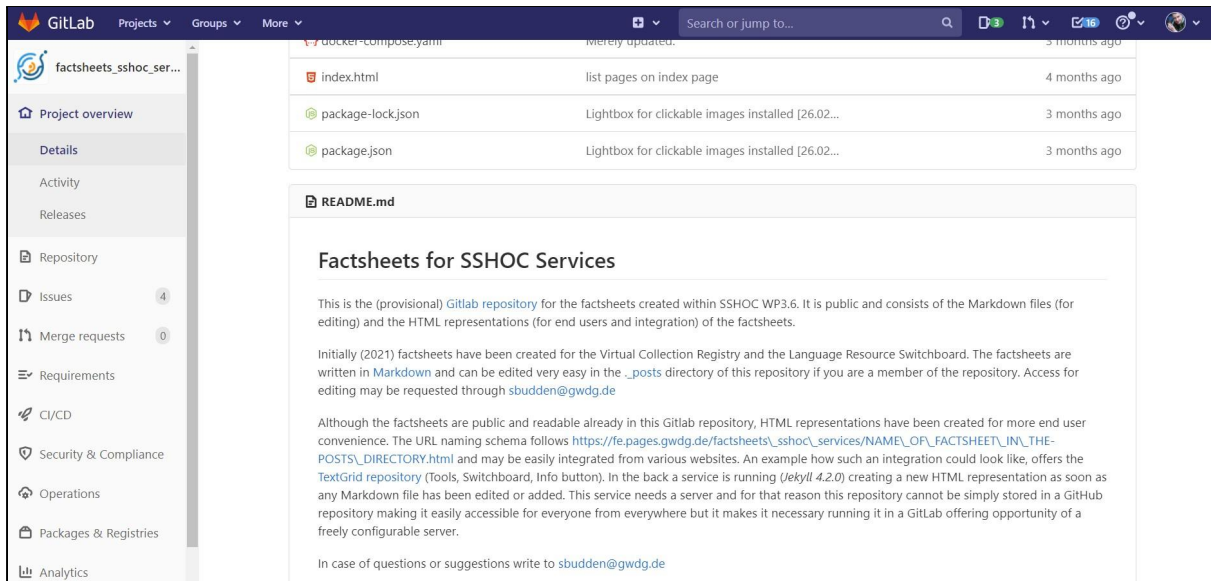


Figure 4: Screenshot of the Github repository used for the VCR tutorial
https://gitlab.gwdg.de/fe/factsheets_sshoc_services

Discussions conducted with (SSHOC) communities

The SSHOC consortium with its 20 partner organisations and 27 associates offers a valuable opportunity to engage various communities on behalf of the VCR. These discussions are not documented in this document but included research infrastructures like CESSDA, DARIAH, E-RIHs and the research library community LIBER. Discussions that resulted in substantial use cases are documented below in a separate chapter.

VCR General Use Cases

As a reminder: in general a virtual collection is a coherent set of links to digital objects. These digital objects can reside in various repositories and can take shape in various formats. The common denominator of a virtual collection may be tutorials for a certain research question, a documentation of used research data and tools for a journal article, or a research data collection for a research group.

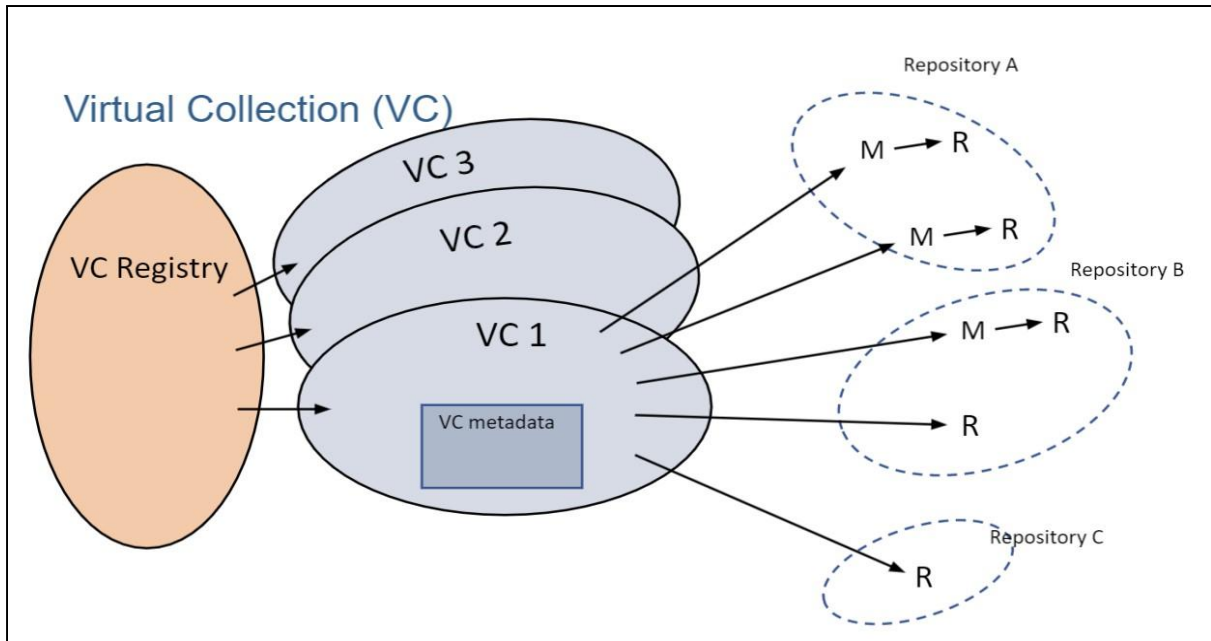


Figure 5: Schema of the VCR

To enhance the functionalities and uptake of the VCR during the SSHOC project, several additional use cases for the VCR have been elaborated. Not all will be implemented but nevertheless may serve as a discussion base in the future.

The use cases below, some of them already known at the start of the project, seem general enough to be of use for the broad SSH and beyond.

1. Create a personal data collection on a specific topic consisting out of possibly (links to) many different digital objects from different collections spanning different catalogues.
 - a. [Already supported]
2. Save search results for future use
 - a. Requires integration in a catalogue supporting search
 - i. [should be supported in the external data catalogue]
 - b. Possibly extend by allowing search results to be added incrementally to an existing search result virtual collection
 - i. [Requires new feature [#106](#)] done
 - ii. [Requires new feature [#83](#)] in progress
 - iii. [Requires new feature [#104](#)] done
3. Downloading the content of a VC into personal researcher's workspace
 - a. This implies two AAI challenges: obtaining access to the user's workspace and secondly to the digital objects to be downloaded. The first challenge should be solvable if only for a limited number of such workspace systems. What comes to mind are: B2DROP (and other ownCloud²¹ technology based solutions) and (especially in the E-RIHS context) PARTHENOS platform. If the workspace can be mounted on a personal computer such as with ownCloud type of technology the problem is reduced to downloading the resources to

²¹ <https://owncloud.com>

- the local machine. The second challenge would be only solvable with a universal SSO system in place, and for now can not be solved and leaves the user iterating over each repository log-in procedure
- i. [Requires new feature [#105](#)] (to be) realised via CMDI-Explorer
 4. Research Objects²²; allow a user/community to create research objects
 - a. Need a RO template, possibly community specific supporting RO metadata and typed relations
 - i. [Requires new feature [#107](#)] open
 5. Write access via the VCR API based on a token (e.g. to create a VC from a script).
 - a. [Requires new feature [#82](#)] done
 6. Minting a DOI for a VC, as to enhance citability. Now possible in VCR v1.5.1

Detailed SSHOC VCR use cases from CESSDA

From a discussion with CESSDA experts & users we have determined the following use cases.

- A. Saving search results as a VC
 - A user browses the CESSDA data-catalogue (or from a national CESSDA node?) and executes a search query. The query results in a result set.
 - [should be supported in the external data catalogue]
 - If the user chooses he can save the result set as a VC.
 - [should be supported in the external data catalogue]
 - If he initiates the creation of a VC, he is forwarded to the VCR application to authenticate.
 - [Supported via the submission API]
 - [Requires new feature [#108](#)] done
 - A2. Using existing search result VCs
 - A user browses the CESSDA data-catalogue (or from a national CESSDA node?)
 - [should be supported in the external data catalogue]
 - He wants to inspect the content of a previous set of search results that were saved as a VC, for instance to compare with other queries
 - To get a list of previous saved search results he is *forwarded* to the VCR after which he obtains a list of VCs with saved search results. Note that this requires typing of search-results e.g. {search-result, CESSDA search result, CESSDA data-catalogue search-result}
 - [Requires new feature [#103](#)] (done on collection level)
 - A3. Users may add new search results to existing search-result VC
 - A user browses the CESSDA data-catalogue (or from a national CESSDA node?) and executes a search query. The query results in a result set.
 - [should be supported in the external data catalogue]
 - If the user chooses he can add the result to an existing unpublished search-result VC
 - [Requires new feature [#106](#)] done
- B. Collaborative editing of (yet unpublished) VC with colleagues

²² <http://www.researchobject.org/>

- A user can send a colleague a link to an unpublished VC
 - [Requires new feature [#81](#)] open
- The possession of the link allows the colleague to also edit and extend the VC as described in A3 (limited security, but simple implementation)
 - [Requires new feature [#81](#)] open

Detailed SSHOC VCR use cases from SSHOC 5.2 (Dataverse)

SSHOC task 5.2 is creating a Cloud version of the Dataverse repository system. There have been two meetings between 3.6 and 5.2 to discuss use cases for integrating the Switchboard and VCR in Dataverse. At 10-6-2020 we discussed a number of use cases that are already partly overlapping with those from other communities:

- A. Consolidation of the resources of a VC into Dataverse. Similar to General Use Case 3
- B. Saving Query results similar to CESSDA use case A

Detailed use cases from E-RIHS v0.8 (Gen 2021)

1. Creation of the resource
 - a. Selecting one or more data collections and perform a search query i.e. databases regarding materials or pigments in infrared spectroscopy, X-ray fluorescence or RAMAN spectrometry selecting used elements²³.
 - i. [should be supported in the external data catalogue]
 - b. Creating a personal collection with the data results of the search query.
 - i. [supported via the submission API]
 - c. Save your search in your personal space
 - i. [Similar to General Use Case 3]
2. Edit saved searches
 - a. copy the data of a research already done by another user being able to add/delete/modify data and resources
 - i. [Requires new feature [#104](#)] done
 - b. ability to add metadata or resources to the search (such as a description of the searched data collection) (either use modified metadata or annotations concepts) *this can be used to compare alternative result sets, or different versions of a collection*
3. Post processing (Interpretation of data).
 - a. possibility to compare different research data, make a gap analysis (useful for identification of studied elements as pigments or materials)
 - i. [should be supported in the external data catalogue, out of scope for the VCR]
 - b. search for previous searches that contain comparable catalogs. The different results obtained should be comparable among them (useful for report production) → extend to store query (if it exists) with extensional collections.

²³ Available as project internal document "VCR E-RIHS Spec Data Catalogue".

- i. [Requires new feature [#103](#)] done
4. Publication
- a. create a safe (persistent and stable) link for publishing or share.
 - i. [supported by assigned PIDs]
 - ii. [Requires new feature [#69](#) to support DOIs] done
 - b. create a workgroup in which users can share, edit, add search results (falls under collaborative editing)
 - i. [Requires new feature [#81](#)] open

Detailed use case from SSHOC Marketplace

CLARIN as a partner in the SSHOC project proved useful to elaborate various use cases in more detail than usually possible. A valuable community engagement was conducted as a SSHOC webinar, discussing the use virtual collections to offer training packages where the collection bundles slides (hosted on e.g. Zenodo), videos (e.g. hosted on youtube), web page and potentially other relevant resources (hosted by various repositories). The webinar is documented and could serve for future activities: [slides](#) + [video\(s\)](#) + [webpage](#).

The discussion between the SSH Open Marketplace work package and Task 3.6 (responsible for the VCR) has led to the identification of additional use cases briefly listed below. A detailed discussion and evaluation had been published as working paper on Zenodo²⁴.

- A. Creation of a Virtual Collection with Records from the MP
 - a. Unique identifiers from MP are necessary, PIDs strongly recommended. Function currently not yet implemented.
- B. Inclusion of the VCR as a feature in the MP
 - a. Would require some work to implement as a feature on the detail page of a record with likely a low cost-benefit ratio.
 - b. To include the VCR as a useful service in the FAQ or manual of the SSH Open Marketplace could be a first and easy start to link the two services. The user would then just read how to create a VC with records from the SSH Open Marketplace.

Detailed use cases from DARIAH-DE/CLARIAH-DE

Against the background of various integration activities between CLARIAH-DE and SSHOC/CLARIN, the following chapter describes the context and use cases in a more detailed way.

DARIAH-DE is since 2011 an established research infrastructure for the arts and humanities and since 2019 merged with CLARIN-D in the CLARIAH-DE initiative. CLARIAH-DE on the other hand provides its offerings sustainably via the "Verein für geistes- und

²⁴ Buddenbohm, Stefan, Broeder, Daan, Eisner, Marthe Irene, Illmayer, Klaus, & Durco, Matej. (2020). Collaborative Use Cases between SSH Open Marketplace and the Language Resource Switchboard and Virtual Collection Registry. Zenodo. <https://doi.org/10.5281/zenodo.4442320> and Buddenbohm, Stefan, & Broeder, Daan. (2021). Addendum to Collaborative Use Cases between SSH Open Marketplace, Switchboard and Virtual Collection Registry. Zenodo. <https://doi.org/10.5281/zenodo.5215987>

kulturwissenschaftliche Forschungsinfrastrukturen e.V.", an association under German law. The offerings of CLARIAH-DE are through these means also available to the NFDI, the National Research Data Initiative. Currently CLARIAH-DE is contributing either directly or indirectly through partners to Text+, NFDI4Culture, NFDI4Memory, and NFDI4Objects, the humanities-related infrastructure consortia of the NFDI.

The DARIAH-DE Data Federation Architecture and its Collection Registry

DARIAH-DE's research infrastructure is aiming to a large part at text-based resources²⁵. To provide the researcher access to text-based resources, DARIAH-DE offers the Data Federation Architecture (DFA)²⁶ which sets the services and tools into a joint framework. The title DARIAH-DE is maintained for the DFA although it is now a CLARIAH-DE offering²⁷.

An important element of the DFA is the DARIAH-DE Collection Registry (ColReg) which provides information on currently 220 collections²⁸ (some also from the CLARIN context). The DARIAH Generic Search with its central search index²⁹ serves as front end for the user, who wants to search for collections.

It is understood, that the DARIAH ColReg offers a similar use case as the CLARIN VCR. It allows the registered user to create own collections in a very convenient way, receive a PID for the collection and even to base such a collection on content uploaded by him to the DARIAH Repository (dhrep). An example for such a user-created collection may be found here³⁰, which offers metadata on the collection *"Revistas culturales históricas en lengua española desde el modernismo hasta las vanguardias: procesos de modernización y formación de redes transnacionales spanish-language Cultural Magazines from Modernismo to Avant-Garde: Processes of Modernization and Transnational Network Formation"*. The metadata is structured into descriptive metadata - what is the collection about or describing? -, administrative metadata - who is the creator of the collection, what kind of licences do apply? - and technical metadata - such as identifiers. The data this specific collection is based on, is archived in the dhrep. In sum these components (and some more) of the DFA come together and offer the researcher a very convenient way to create individual collections of research data.

During the CLARIAH-DE project phase the data and searches behind DARIAH-DE and CLARIN-D have been linked with one another on various levels of integration³¹.

²⁵ To orient on the broad range of possible use cases in DARIAH (and concurrently taking CLARIN uses cases into account as well), the Text+ user stories may be consulted: Rißler-Pipka, Nanette, Barthauer, Raisa, Buddenbohm, Stefan, Calvo Tello, José, Friedrichs, Sonja, & Weimer, Lukas. (2021). Community Involvement in Research Infrastructures: The User Story Call for Text+ (1.0.0). Zenodo. <https://doi.org/10.5281/zenodo.5384085>

²⁶ English documentation on the DFA is available here: <https://de.dariah.eu/en/data-federation-architecture>

²⁷ This relates to a similar discussion about the branding of the VCR and Switchboard as SSHOC results.

²⁸ <https://colreg.de.dariah.eu/colreg-ui/collections/>

²⁹ Other than the federated approach of CLARIN, where - for instance - a search request in the FCS is sent to several search indexes.

³⁰ <https://hdl.handle.net/21.11113/0000-000D-1CFF-6>

³¹ See Thomas Eckart, Tobias Gradl, Robin Jegan, Eliza Margaretha, Antonina Werthmann, Felix Helfer, Stefan Buddenbohm. "CLARIAH-DE Cross-Service Search: Prospects and Benefits of Merging Subject-specific Services". DARIAH-DE Working Papers Nr. 41. Göttingen: DARIAH-DE, 2021. URN: <urn:nbn:de:gbv:7-dariah-2021-1-9>

DARIAH-DE-related repositories have become CLARIN Data Centres and vice versa CLARIN collections have been made searchable through the DFA.

CLARIAH-DE/ DARIAH-DE DFA and the CLARIN VCR

Against this background - an already existing similar service in CLARIAH-DE - the VCR use case has to revolve more around interoperability and enhanced user experience (more research data and services available) than on integrating DARIAH-DE/CLARIAH-DE in the existing VCR. Basic interoperability between the two communities is already given and conveniently possible by using an existing DARIAH account for the VCR.

Apart from this basic interoperability some specific integration options can be described:

1. Elaboration of integration options with look at the DFA and the CLARIN CMDI (<https://www.clarin.eu/content/component-metadata>) to enhance interoperability between data domains of CLARIN and DARIAH.
2. Enhancement of the CLARIN VCR in a way that it may access the DARIAH Generic Search and by this acquires humanities-related data. In this use case, the DFA could be understood as the humanities interface of the VCR.
3. Extension of 2: Enhancement of the CLARIN VCR in a way that it is able to save search queries to the DARIAH Generic Search and is able to aggregate them instantly.

In terms of the specific functionalities as mentioned by the other ERICs above, CLARIAH-DE would request the following ones:

- A. Expansion of the available repositories to the VCR (e.g. through a simple visual integration of the VCR in the individual repository interface or FAQs).
 - a. Testing and possibly indexing of larger data domain for the VCR. Discipline-specific?
 - b. This could include a showcase demonstrating the user how he may add resources from humanities collections. Such a showcase could rely on a DARIAH-DE data domain.
- B. Personal space
 - a. Sharing of VCs: a user should be able to access VCs from other creators and should be able to copy them to its his own space.
- C. Citing of VCs
 - a. VCs should be easily citable via PIDs.
 - b. Export functionality: download via the CMDIExplorer³²

Apart from the abovementioned use cases some other scenarios emerged in the discussion along the SSHOC project:

1. Review of the CLARIN VCR API and the RDA Collection API and evaluation of possible acquisitions enhancing the VCR. In view of the problems getting developer resources for this we may have to postpone it beyond the SSHOC scope. Integration with other repository APIs such as FDO API may need to be prioritised.
2. Work on a possible VCR extension of the marketplace, e.g. that VCs may contain items from the marketplace.

³² <https://weblicht.sfs.uni-tuebingen.de/CMDIExplorer/>

Analysis & comments of the use cases

From an analysis of the provided use cases by the different SSHOC stakeholder infrastructures and the comments from their peers we consider the following use cases as useful and applicable for the broad SSH:

1. VC creation for bookmarking, sharing and publication; the classical scenario

A researcher creates a virtual collection in the VCR with as purpose (1) a private set of bookmarks to resources interesting for his research (2) potentially being able to share the collection with his colleagues and (3) potentially publishing the collection by the VCR enabling harvesting of the VC metadata by other infrastructure components.

Sub-case (2) deserves attention wrt the need for shared editing of a VC by multiple researchers and secondly (2) also needs an extension wrt. versioning e.g. which version gets shared. No special integration with other infrastructure components is necessary with the exception of the AAI solutions chosen by the different SSHOC stakeholder infrastructures.

2. Making VCs interoperable with other services

Additional functionality that can be offered by additional VC metadata would be to provide interoperable services ques wrt to the creator's intention for this VC. In this way interoperable services such as the Switchboard could rank & offer services based not only on the VC constituent's data-type but also on the VC creator's intention for the VC constituents. Evidently special metadata is required in addition to the current descriptive metadata. Other types of services that can be beneficially made interoperable with VCs are workflow engines such as WebLicht. Workflow engines can be made to iterate over the VC constituents and (under the right circumstances) deposit also the results into the VC together with appropriate provenance data.

3. Saving query results in VCs

Currently in many data-catalogues there is a possibility to save the results (result-set) of a query for future inspection. However these result-sets are only interoperable with one specific data-catalogue only. A strong extension to VC applicability is to use the VCs to support the incremental aggregation of result-sets potentially spanning different data-catalogues.

In a VCR interoperable data-catalogue a user is able to add a query result-set to a new VC or to an already existing VC.

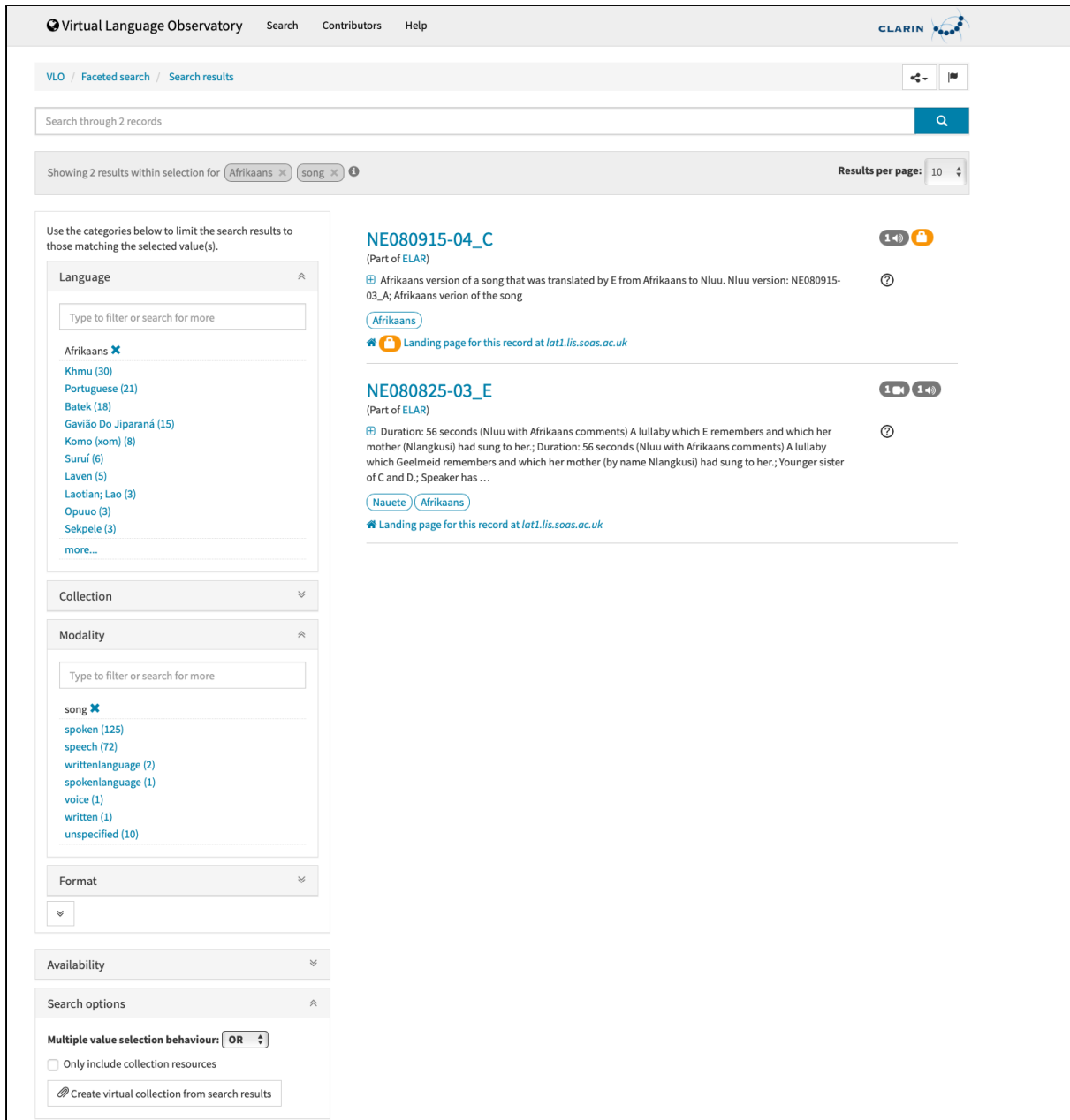
- The VCR requires specific metadata that supports result-sets (e.g. saving the query although the result-set VC is extensional)

Current scenario / basic integration (currently available³³):

- A data-catalogue builds a post request with various parameters describing the virtual collection and it's references.

³³ VCR Integration Documentation: <https://github.com/clarin-eric/VirtualCollectionRegistry/blob/master/doc/Integration.md>

- The user-agent executes to POST request at the VCR submission endpoint (thus the user has left the data-catalogue domain and is now on the VCR domain)
 - If no authentication session is available, the VCR caches the submitted data and initiates a SAML login workflow
 - If a login session is available after completing the SAML workflow or because of an already existing login session, the cached values are used to prefill the virtual collection creation form and the user can update the values if needed before creating the collection.
-
- Ideal scenario / deep integration (requires implementation): This requires a data-catalogue to use the VCR API and to appropriately authenticate on behalf of the user or use an application token.
 - An OAUTH2 approach to support user delegation from the data catalogue to the VCR is technically feasible but currently not organisational SSH wide. On a small scale and possible short-term, a 'golden-token' confidence relation between the VCR and selected data-catalogues is feasible. This implies the creation of application tokens for each data-catalogue. This token can then be used by the data-catalogue to authenticate to the VCR API. There still needs to be a mechanism to link user's authenticated in the data-catalogue to user's in the VCR since these are two separate user domains in this scenario.



Virtual Language Observatory Search Contributors Help CLARIN

VLO / Faceted search / Search results

Search through 2 records

Showing 2 results within selection for Afrikaans song Results per page: 10

Use the categories below to limit the search results to those matching the selected value(s).

Language

Type to filter or search for more

- Afrikaans **X**
- Khmu (30)
- Portuguese (21)
- Batek (18)
- Gavião Do Jiparaná (15)
- Komo (kom) (8)
- Suruí (6)
- Laven (5)
- Laotian; Lao (3)
- Opuuo (3)
- Sekpele (3)
- more...

Collection

Modality

Type to filter or search for more

- song **X**
- spoken (125)
- speech (72)
- writtenlanguage (2)
- spokenlanguage (1)
- voice (1)
- written (1)
- unspecified (10)

Format

Availability

Search options

Multiple value selection behaviour: OR

Only include collection resources

Create virtual collection from search results

NE080915-04_C (Part of ELAR) 1 1

Afrikaans version of a song that was translated by E from Afrikaans to Nluu. Nluu version: NE080915-03_A; Afrikaans verion of the song

Afrikaans

Landing page for this record at lat1.iis.soas.ac.uk

NE080825-03_E (Part of ELAR) 1 1


Duration: 56 seconds (Nluu with Afrikaans comments) A lullaby which E remembers and which her mother (Nlangkusi) had sung to her; Duration: 56 seconds (Nluu with Afrikaans comments) A lullaby which Geelmeid remembers and which her mother (by name Nlangkusi) had sung to her; Younger sister of C and D; Speaker has ...

Nauete Afrikaans

Landing page for this record at lat1.iis.soas.ac.uk

Figure 7: The CLARIN VLO query for songs in the Afrikaans language. The option to create Virtual Collection from the recalled resources is present in the lower left corner.

Virtual Collection Registry
Browse
Create
My Collections
Help

daan.broeder_di.huc.knaw.nl@clarin.eu


Create / Edit Collection: Advanced Mode

Name:

Description:

New collection description

Note: Markdown supported ([cheat sheet](#))

Keywords:

Authors:

↑ ↓ ↻ ✎
🗑️

Family Name:

Given Name:

Email:

Affiliation:

Press <enter> to add this author

References:

Add new reference by URL or PID

Set a title for this new reference

Press <enter> to add this reference

NE080915-04_C (hdl:2196/00-0000-0000-000D-AB23-C@format=cmdi) ⊗

Afrikaans version of a song that was translated by E from Afrikaans to Nluu. Nluu version: NE080915-03_A

↑ ↓ ↻ ✎
🗑️

NE080825-03_E (hdl:2196/00-0000-0000-000D-AA22-A@format=cmdi) ⊗

Duration: 56 seconds (Nluu with Afrikaans comments)

A lullaby which E remembers and which her mother (Nlangkusi) had sung to her.

↑ ↓ ↻ ✎
🗑️

Save Collection
Cancel

Figure 8: The CLARIN VLO query for songs in the Afrikaans language and creation of a virtual collection with these resources.

Conclusion

This document summarised the use case discussions aiming at the Virtual Collection Registry (VCR) as part of SSHOC's task 3.6 "Lifting Services in the SSH Cloud". During the course of the project 2019-2022 these discussions led not only to the implementation of new functionalities or widening of the VCR's scope but also to a set of ideas, for which SSHOC cannot be a suitable framework for implementation. These ideas are nevertheless worthwhile to follow up on in the future and have therefore been documented here.

Looking at the potential to scale up the uptake of the VCR it became clear that the service is - despite its origin in the CLARIN research infrastructure - ready to be adapted for other disciplines. The VCR as such is agnostic of the provenance of the resources in its virtual collections. On the other hand, the basis of the service in the CLARIN research infrastructure proves useful for the sustainability and continued product maintenance so far. Getting rid of this environment seems not to be advisable, especially since CLARIN has vouched for the VCR availability for the other SSHOC infrastructures. This point is particularly proven by the description of the DARIAH-DE Collection Registry in this document. Similarities between the two services are obvious and particularly visible with regard to the benefits of ownership by a research infrastructure.

The implementation status of the above discussed use cases and additional functionalities is documented in a separate "Implementation Report" (Deliverable 3.8) to be published at the end of the SSHOC project in 2021. This implementation report also includes integration scenarios for another CLARIN service, the Switchboard.

References

Buddenbohm, Stefan, Broeder, Daan, Eisner, Marthe Irene, Illmayer, Klaus, & Durco, Matej. (2020). Collaborative Use Cases between SSH Open Marketplace and the Language Resource Switchboard and Virtual Collection Registry. Zenodo. <https://doi.org/10.5281/zenodo.4442320>

Buddenbohm, Stefan, & Broeder, Daan. (2021). Addendum to Collaborative Use Cases between SSH Open Marketplace, Switchboard and Virtual Collection Registry. Zenodo. <https://doi.org/10.5281/zenodo.5215987>

CLARIN (2021): Github Repository for the Virtual Collection Registry. <https://github.com/clarin-eric/VirtualCollectionRegistry>

CLARIN Short Guide on Virtual Collections (2010):
https://www.clarin.eu/sites/default/files/virtual_collections-CLARIN-ShortGuide.pdf

DARIAH UGOE (2021): Factsheets for SSHOC Services. https://github.com/subugoe/factsheets_sshoc_services

Elbers, Willem/ Broeder, Daan (2021): VCR Integration Documentation:
<https://github.com/clarin-eric/VirtualCollectionRegistry/blob/master/doc/Integration.md>

Eckart, Thomas/ Gradl, Tobias/Jegan, Robin/ Margaretha, Eliza/Werthmann, Antonina/Helfer, Felix/ Buddenbohm, Stefan. "CLARIAH-DE Cross-Service Search: Prospects and Benefits of Merging Subject-specific Services". DARIAH-DE Working Papers Nr. 41. Göttingen: DARIAH-DE, 2021. URN: [urn:nbn:de:gbv:7-dariah-2021-1-9](https://nbn-resolving.org/urn:nbn:de:gbv:7-dariah-2021-1-9)

EOSC Portal (2021): Virtual Collection Registry. <https://marketplace.eosc-portal.eu/services/virtual-collection-registry>

Gorgaini, Elisa (2021). Virtual Collection Registry - SSHOC Service Catalogue's Factsheets Series (v1.0). Zenodo. <https://doi.org/10.5281/zenodo.5482899>

RDA - Research Data Collections Working Group (2016): API Requirements Document.
<https://www.rd-alliance.org/api-requirements-document>

Rißler-Pipka, Nanette, Barthauer, Raisa, Buddenbohm, Stefan, Calvo Tello, José, Friedrichs, Sonja, & Weimer, Lukas. (2021). Community Involvement in Research Infrastructures: The User Story Call for Text+ (1.0.0). Zenodo. <https://doi.org/10.5281/zenodo.5384085>

Uytvanck, Dieter van/Goosen, Twan/Elbers, Willem (2018): EOSC-Hub Task 7.1 Roadmap