https://doi.org/10.20378/irb-51559

CLARIAH-DE Cross-Service Search
-
Prospects and Benefits of Merging Subject-specific Services

Thomas Eckart, Tobias Gradl, Robin Jegan, Eliza Margaretha, Antonina Werthmann, Felix Helfer, Stefan Buddenbohm

DARIAH-DE Working Papers Nr. 41

Eckart, T.; Gradl, T.; Jegan, R.; Margaretha, E.; Werthmann, A.; Helfer, F.; Buddenbohm, S.: CLARIAH-DE Cross-Service Search : Prospects and Benefits of Merging Subject-specific Services
Göttingen : GOEDOC, Dokumenten- und Publikationsserver der Georg-August-Universität, 2021
(DARIAH-DE working papers 41)

Verfügbar:
PURL: http://resolver.sub.uni-goettingen.de/purl/?dariah-2021-1
URN: http://nbn-resolving.de/urn:nbn:de:gbv:7-dariah-2021-1-9

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über <http://dnb.dnb.de> abrufbar.

Erschienen in der Reihe
DARIAH-DE working papers

ISSN: 2198-4670

Herausgeber der Reihe
DARIAH-DE, Niedersächsische Staats- und Universitätsbibliothek

Mirjam Blümm, Thomas Kollatz, Stefan Schmunk und Christof Schöch

Abstract: CLARIAH-DE combines services and offerings of CLARIN-D and DARIAH-DE. This includes various search applications which are made directly available to researchers. These search applications are presented in this working paper based on their main characteristics and compared with a focus on possible harmonizations. Opportunities and risks of different forms of technical integration are highlighted. Identified challenges can be explained in particular considering the background of different organizational and technical frameworks as well as highly specific and discipline-dependent requirements. The integration work that has already been carried out and the experiences gained with regard to future work and possible integration of further applications are also discussed. The experiences made in CLARIAH-DE can especially be of interest for other projects in the field of digital research infrastructures.

Keywords: Digitale Forschungsinfrastruktur, Such- und Recherchesysteme, Serviceintegration, Metadaten, CLARIAH-DE, CLARIN, DARIAH

Digital research infrastructure, search systems, service integration, metadata, CLARIAH-DE, CLARIN, DARIAH

CLARIAH-DE Cross-Service Search

Prospects and Benefits of Merging Subject-specific Services

Thomas Eckart[1] Tobias Gradl[2] Robin Jegan[2]

Eliza Margaretha[3] Antonina Werthmann[3] Felix Helfer[1]

Stefan Buddenbohm[4]

[1] Institut für Informatik, Universität Leipzig
[2] Otto-Friedrich-Universität Bamberg, Lehrstuhl für Medieninformatik
[3] Leibniz-Institut für Deutsche Sprache, Mannheim
[4] Niedersächsische Staats- und Universitätsbibliothek Göttingen

DARIAH-DE
Working Papers

Thomas Eckart, Tobias Gradl, Robin Jegan, Eliza Margaretha, Antonina Werthmann, Felix Helfer, Stefan Buddenbohm: "CLARIAH-DE Cross-Service Search". *DARIAH-DE Working Papers* No. 41. Göttingen: DARIAH-DE, 2021. URN: urn:nbn:de:gbv:7-dariah-2021-1-9.

The *DARIAH-DE Working Papers* are published by Mirjam Blümm, Thomas Kollatz, Stefan Schmunk and Christof Schöch.

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Abstract

CLARIAH-DE combines services and offerings of CLARIN-D and DARIAH-DE. This includes various search applications which are made directly available to researchers. These search applications are presented in this working paper based on their main characteristics and compared with a focus on possible harmonizations. Opportunities and risks of different forms of technical integration are highlighted. Identified challenges can be explained in particular considering the background of different organizational and technical frameworks as well as highly specific and discipline-dependent requirements. The integration work that has already been carried out and the experiences gained with regard to future work and possible integration of further applications are also discussed. The experiences made in CLARIAH-DE can especially be of interest for other projects in the field of digital research infrastructures.

Schlagwörter

Digitale Forschungsinfrastruktur, Such- und Recherchesysteme, Serviceintegration, Metadaten, CLARIAH-DE, CLARIN, DARIAH

Keywords

Digital research infrastructure, search systems, service integration, metadata, CLARIAH-DE, CLARIN, DARIAH

Contents

1 Introduction 4

2 Generic Search (GS) 5
- 2.1 Organisational Context . . . 6
- 2.2 Available Resources and Resource Inventory . . . 7
- 2.3 Search Index, Search Logic, and Query Language . . . 7
- 2.4 User Interface and Presentation of Results . . . 8
- 2.5 Ranking . . . 8
- 2.6 Customization . . . 9

3 Virtual Language Observatory (VLO) 9
- 3.1 Organisational Context . . . 10
- 3.2 Available Resources and Resource Inventory . . . 11
- 3.3 Search Index, Search Logic, and Query Language . . . 12
- 3.4 User Interface and Presentation of Results . . . 12
- 3.5 Ranking . . . 13
- 3.6 Customization . . . 13

4 Federated Content Search (FCS) 13
- 4.1 Organisational Context . . . 14
- 4.2 Available Resources and Resource Inventory . . . 15
- 4.3 Search Index, Search Logic, and Query Language . . . 15
- 4.4 User Interface and Presentation of Results . . . 16
- 4.5 Ranking . . . 16
- 4.6 Customization . . . 16

5 Overlaps and Major Issues 16
- 5.1 Organisational and Technical Context . . . 16
- 5.2 Search Syntax and Search / Filter Functionality . . . 17
- 5.3 Presentation of Results . . . 18
- 5.4 Corporate Design . . . 18

6 Realized Harmonizations 18
- 6.1 Cross-Service Search: Integrated Access to the Search Services . . . 18
- 6.2 Generic Search as an Endpoint of the Federated Content Search . . . 19
- 6.3 Educational and Training Materials with Generic Search . . . 20
- 6.4 Adaptations of VLO and FCS . . . 20

7 Lessons Learned and Perspectives 21

8 Annex 24

Bibliography 27

1 Introduction

This concept paper presents opportunities and difficulties related to the use of search and research systems within the context of the CLARIAH-DE project.

Figure 1: Logo of CLARIAH-DE.

CLARIAH-DE[1] is a contribution to the digital research infrastructure for the Humanities and related disciplines. It is funded by the Federal Ministry of Education and Research (BMBF), grant no. 01UG1610 A to I. Due to the merger of the infrastructures CLARIN-D[2] and DARIAH-DE[3], CLARIAH-DE creates added value for researchers: existing offers and services are further developed and jointly made available in accordance with the FAIR principles[4] Findable, Accessible, Interoperable, and Reusable. The project's portfolio includes the sustainable provision of research data, technical infrastructures, digital tools, and virtual research environments for the Humanities, information and teaching material, as well as instructions on standards and proceedings.

As part of CLARIAH-DE, various search and research systems are available, of which the following ones have been identified as particularly relevant:

- Generic Search[5] based on the DARIAH-DE Data Modelling Environment
- CLARIN(-EU) Virtual Language Observatory (VLO[6]) based on the Component Metadata Infrastructure (CMDI)
- CLARIN(-EU) Federated Content Search (FCS[7])

Below, the characteristics of these services are described and compared regarding possible future harmonizations. First and foremost, the findability of resources takes centre stage, i.e. it is most important to offer (new) users the possibility of comprehensive access to the complete CLARIAH-DE resources inventory. Since a standardisation of such services, e.g. with regard to metadata interoperability or discipline-specific characteristics, is always accompanied by compromises (cf. "Overlaps and Major Issues") and specific searches from the user's perspective remain desirable in many cases (e.g. an

[1] https://clariah.de/en/
[2] https://www.clarin-d.net/en/
[3] https://de.dariah.eu/
[4] https://www.go-fair.org/fair-principles/
[5] https://search.de.dariah.eu/search/
[6] https://vlo.clarin.eu
[7] https://contentsearch.clarin.eu/

exclusive search for teaching materials), the three search services will be provided and maintained on an individual basis in the future.

The relevant characteristics also include:

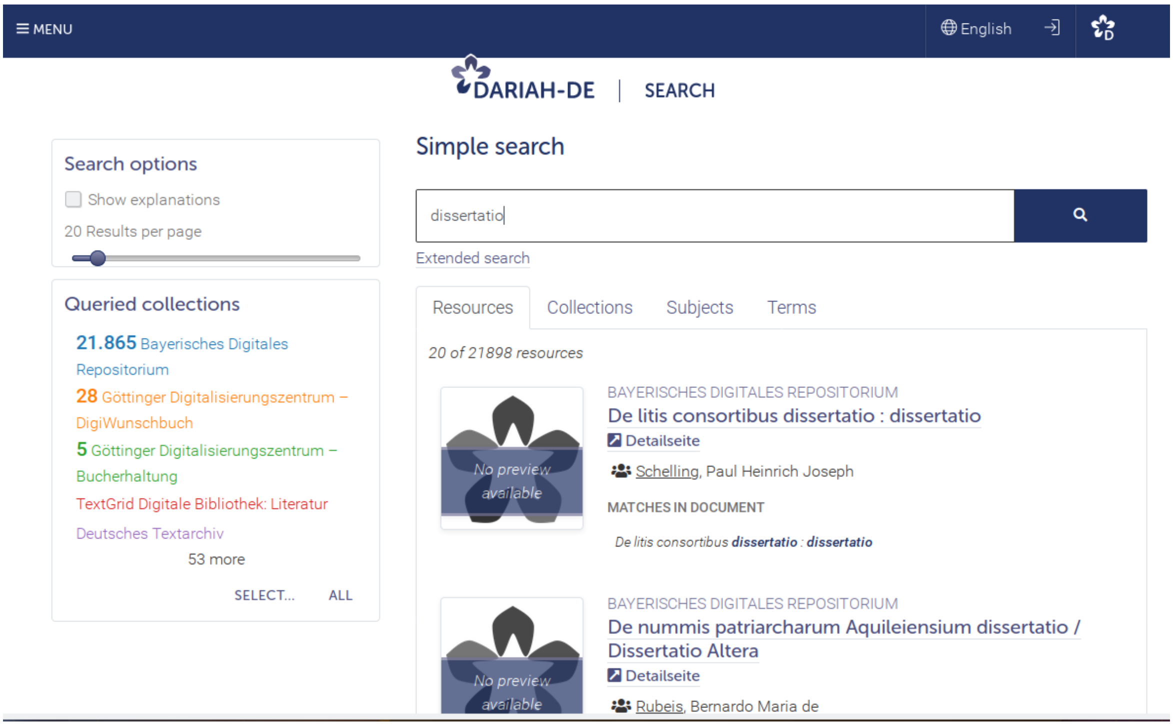
- Addressed / Available resource types
  - Data import (push / pull)
- Search logic and retrieval model
  - Query language, full text search, faceted search, AND / OR combination of search terms, etc.
  - Index used, use of local vs. distributed index, etc.
- Presentation / User Interface (UI)
  - UI search structure
  - Internationalisation / Localisation
  - Recognition value, "corporate design", embedding, or cross-referencing external projects
  - Adaptability of user interfaces, icons, used / presented search metaphors
- Possibilities / Effort of (project related) customisation
  - References to CLARIN-EU[8]
  - References to Text+[9]
- Usability

2 Generic Search (GS)

The Generic Search offers a comprehensive search option for the scientific collections and data sources stored in the Collection Registry.

[8] https://clarin.eu

[9] https://www.text-plus.org/en



Simple search

dissertatio

□ Show explanations

20 Results per page

Queried collections

21.865 Bayerisches Digitales Repositorium

28 Göttinger Digitalisierungszentrum – DigiWunschbuch

5 Göttinger Digitalisierungszentrum – Bucherhaltung

TextGrid Digitale Bibliothek: Literatur

Deutsches Textarchiv

53 more

20 of 21898 resources

BAYERISCHES DIGITALES REPOSITORIUM
De litis consortibus dissertatio : dissertatio
Schelling, Paul Heinrich Joseph
MATCHES IN DOCUMENT
De litis consortibus **dissertatio** : **dissertatio**

BAYERISCHES DIGITALES REPOSITORIUM
De nummis patriarcharum Aquileiensium dissertatio / Dissertatio Altera
Rubeis, Bernardo Maria de

Figure 2: Generic Search (GS).

2.1 Organisational Context

The GS is embedded in the **federation concept** of DARIAH-DE, which aims at providing integrative services. Based on cross-collection and interdisciplinary views, researchers are supported in identifying and integrating relevant collections as well as relevant resources. Therefore, the searchable data usually do not originate from DARIAH-DE itself. The following components of the DARIAH-DE architecture, in combination with the Generic Search, are responsible for enabling the search function and the use of other tools in DARIAH-DE:

- The federation layer comprises the Collection Registry (CR)[10], in which collections are registered and described, as well as the Data Modelling Environment (DME)[11], in which data models and mappings between data models are applied.
- The data layer describes the accessible collections, which can be reached via interfaces such as OAI-PMH.
- In addition to the GS, the service layer contains other applications such as the Cosmotool[12] which allows researchers to gain access to humanities data or offers additional services to process (e.g. annotate, enrich, analyse) the search results, i.e. the data from the GS.

The GS can also be used for both depth and breadth first search.

[10] https://colreg.de.dariah.eu/colreg-ui/

[11] https://dme.de.dariah.eu/dme/registry/

[12] https://cosmotool.de.dariah.eu/cosmotool/personsearch/

2.2 Available Resources and Resource Inventory

The target group of the GS are scholars of the humanities and cultural studies in research and teaching who work with digital resources and methods. The indexed data sources can mostly be characterized as **scientific collections** that bundle both **research data and metadata**. Accordingly, the GS enables access not only to pure textual data, but also creates a basis on which multimedia, its metadata, graphic evaluations or visualizations of data volumes can be searched.

The GS makes use of other building blocks of the DARIAH-DE federation architecture, in particular the CR and the DME. The content that is indexed in the GS is previously registered as a collection in the CR and described there. Data models and mappings amongst them can be created in the DME in order to create connections between different data sources and their types.

In addition to the different metadata schemes that are supported by the DME in this architecture, the **DARIAH Collection Description Data Model (DCDDM)**[13] was created, which was adapted specifically to the requirements of collections in the humanities.

The execution of queries in the GS is carried out on the contents of the collections that are registered in the CR and whose data models are described by the DME. A necessary requirement is that data can be accessed via interfaces such as OAI-PMH.

Queries can be carried out both as a full text search on all available collections (**Simple Search**), and by selecting one or more metadata fields. This **advanced search** is made possible by the use of DME mappings. A mapping describes rules for mapping a source data model to a target data model[14]. A search query can be faceted as part of the advanced search using a selectable data model (in the case of the primary instance of the GS mostly Dublin Core[15]). By using mappings, the search query is translated into locally used formats and can therefore be executed on heterogeneous data.

The selection of one or more integration models (e.g. for the faceting of queries) takes place dynamically, but depends on the existence of mappings on these integration models. This flexibility means that the GS can be used in different contexts. In addition to the primary CLARIAH-DE instance with currently 47 collections with over 1.2 million documents (as of 2021-01-20), there are other installations, for example as MWW joint search (27 collections, 250,000 documents) or CLARIAH-DE Tutorial Finder (8 collections, 558 documents). Both search solutions are currently under construction.

2.3 Search Index, Search Logic, and Query Language

The search and analysis software **Elasticsearch**[16], which is based on Apache Lucene, is used in the GS. By entering a collection in the CR, the metadata and access mechanisms to this collection are described and stored. The data model of this collection is entered into the DME, if it does not already exist, and mappings and the transformation rules contained therein are created. The data models and mappings

[13] https://github.com/DARIAH-DE/DCDDM

[14] cf. e.g. https://doi.org/10.1515/bfp-2016-0027

[15] https://dublincore.org

[16] https://www.elastic.co/de/elasticsearch/

are analysed during the data indexing in order to obtain a ”maximum possible level of integrated semantics”[17], which can be used for a more efficient execution during query processing.

The query language is based on the **Elasticsearch Query Language**[18], which supports Boolean operators such as AND / OR as well as the search for phrases and the use of wildcards in search queries. The user interface, which is available in both German and English, offers full-text searches in all collections as well as more specific queries using the advanced search through the selection of metadata fields. In the advanced search, several search terms can be used in different metadata fields in order to narrow down the results more precisely. For this purpose, the elements of the **Dublin Core** metadata standard[19] are offered as facets in the standard setting, but elements of the other data models entered into the DME can also be selected.

2.4 User Interface and Presentation of Results

For the simple search, the GS presents a text field in which search terms can be entered. The advanced search supports several text fields and the associated metadata fields, which are displayed as a drop-down selection. The number of search terms can differ in the advanced search. A single search expression or many different search queries using different metadata fields are possible.

In the standard setting, the **results are presented as a list with 20 results**, however, the number of results per page can be chosen between 10 and 150 hits. In addition to the resources themselves, other types of results can also be selected, for example in which collections the search term was found. Furthermore, a list of the searched collections can be used to create a filter for the considered sources. The selection of a specific result opens a detailed view of the resource with further information, the original data set and links to the resource. These links can be used to view the resource in its original state on the websites of the collection, which was entered into the DARIAH-DE federation architecture using the CR.

2.5 Ranking

The ranking of the result list is based on the Elasticsearch index, which is used in two variants. For a simple search, a full text search is carried out on all available data, both primary data and metadata. In the extended search, only those fields are used that were selected by the user in the search query, based on the schema that was selected as the data model in the DME and on which the data was indexed. If appropriate mappings are available, faceted requests are also translated into semantically associated fields in the target data models. Furthermore, facets for selective searches in individual collections, a selection of collections, or all collections can also be made here.

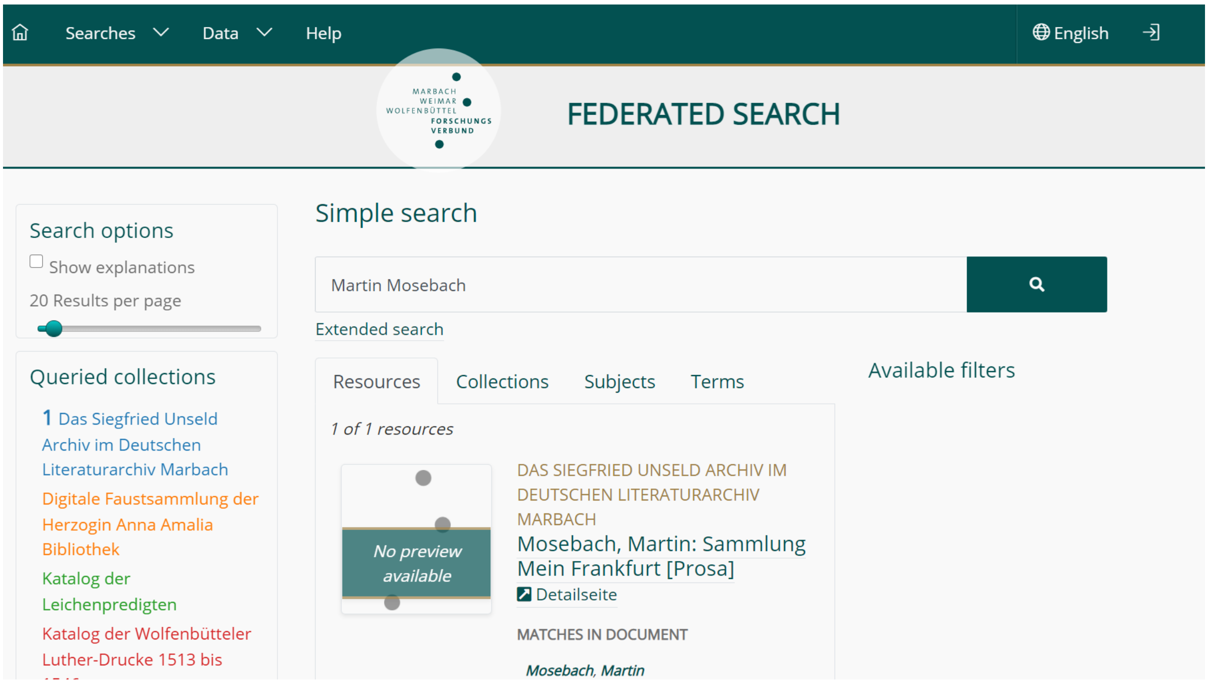
[17] For more information about indexing and query processing: https://www.degruyter.com/view/journals/bfup/40/2/article-p222.xml

[18] https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-simple-query-string-query.html

[19] https://www.dublincore.org/specifications/dublin-core/dces/

2.6 Customization

Alongside the CR and the DME, the GS is a central part of the DARIAH-DE federation concept. In addition to the central installation of services within the DARIAH-DE / CLARIAH-DE infrastructure, there are also feasibility studies and projects whose implementation is based on dedicated installations of these components. Examples are the federated search of the Marbach Weimar Wolfenbüttel Research Association (MWW)[20], a first joint study with the Germanisches National Museum (GNM)[21], and the CLARIAH-DE Tutorial Finder[22].



□ Show explanations

20 Results per page

1 Das Siegfried Unseld Archiv im Deutschen Literaturarchiv Marbach

Digitale Faustsammlung der Herzogin Anna Amalia Bibliothek

Katalog der Leichenpredigten

Katalog der Wolfenbütteler Luther-Drucke 1513 bis

Martin Mosebach

1 of 1 resources

DAS SIEGFRIED UNSELD ARCHIV IM DEUTSCHEN LITERATURARCHIV MARBACH

Mosebach, Martin: Sammlung Mein Frankfurt [Prosa]

Mosebach, Martin

Figure 3: Federated search at the MWW.

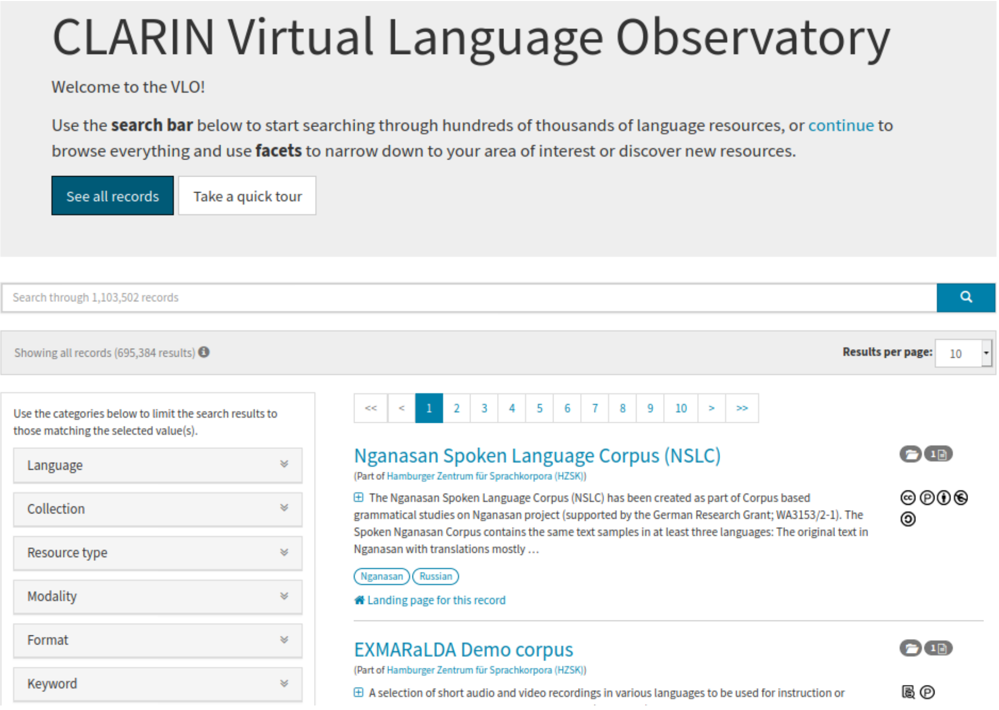
3 Virtual Language Observatory (VLO)

The Virtual Language Observatory (VLO) is a search engine for metadata with a specific focus on language resources including data, services, and tools.

[20] https://search.mww-forschung.de

[21] http://genericsearch.gnm.de

[22] https://teaching.clariah.de



CLARIN Virtual Language Observatory

Welcome to the VLO!

Use the **search bar** below to start searching through hundreds of thousands of language resources, or continue to browse everything and use **facets** to narrow down to your area of interest or discover new resources.

Showing all records (695,384 results)

Results per page: 10

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

Nganasan Spoken Language Corpus (NSLC)

(Part of Hamburger Zentrum für Sprachkorpora (HZSK))

⊞ The Nganasan Spoken Language Corpus (NSLC) has been created as part of Corpus based grammatical studies on Nganasan project (supported by the German Research Grant; WA3153/2-1). The Spoken Nganasan Corpus contains the same text samples in at least three languages: The original text in Nganasan with translations mostly …

EXMARaLDA Demo corpus

(Part of Hamburger Zentrum für Sprachkorpora (HZSK))

⊞ A selection of short audio and video recordings in various languages to be used for instruction or

Figure 4: Virtual Language Observatory (VLO).

3.1 Organisational Context

The **VLO is strongly embedded in the European CLARIN infrastructure** and is based on their general framework, structures, and goals. This includes among others:

- The development concept which relies on developers of CLARIN ERIC and several national consortia (currently primarily CLARIN-D and CLARIN-AT).
- The accepted data model that solely includes metadata following the CMDI standard[23] and which is often defined using the CLARIN Concept Registry[24]. External metadata stocks are included in the VLO using transformations into equivalent CMDI schemata.
- The hosting and maintenance infrastructure provided by CLARIN ERIC.
- The integration concept, which focuses on a deep integration with other (European) CLARIN applications and services. This currently includes the CLARIN Centre Registry[25], CLARIN Federated Content Search (FCS), the Language Resource Switchboard[26], the Virtual Collection Registry[27], and the CLARIN Metadata Curation Module[28] including its link checker.

[23] https://www.clarin.eu/content/component-metadata

[24] https://www.clarin.eu/ccr

[25] https://centres.clarin.eu

[26] https://switchboard.clarin.eu

[27] https://collections.clarin.eu

[28] https://curate.acdh.oeaw.ac.at

- The corporate design which is based on general style guides, color palettes, and logos of the European CLARIN project.

CLARIN-D is actively involved in many of these areas for years, but represents only one of several participating organisations[29].

3.2 Available Resources and Resource Inventory

The VLO focuses on scientific users and enables **searching in metadata of various resources for all language- and text-oriented disciplines**. This includes metadata to corresponding data stocks (like text / language corpora, treebanks, dictionaries), tools (such as taggers or classifiers), and (Web) services (such as annotation services).

All available metadata are **based on XML schemata following the CMDI standard** (in parts using transformations from other source schemata like Dublin Core, MODS[30], or the Europeana Data Model EDM[31]). The import process and the records' representation are thus based on possibilities and features of the CMDI standard. This includes a very flexible design of the actual metadata payload[32] that largely allows all XML structures (tree structure, attributes, etc.) and are interpreted **on the basis of external data or concept categories** (such as the CLARIN Concept Registry CCR or Dublin Core). Information can also be extracted on the basis of schema-specific XPATH expressions. The CMDI header common to all CMDI-based metadata records only contains general organisational and descriptive metadata (such as PID or collection). The *Resources* header, which is also mandatory, allows references to other resources, landing pages or the **construction of (arbitrarily complex) hierarchy trees of resources**. The latter can be used to structure collections as well as to separate complex resources into their constituents (like a corpus consisting of several subcorpora) and is supported by the VLO via a searchable hierarchy component.

All metadata inventory available in the VLO is queried via the **standard interface OAI-PMH** from participating metadata providers and is usually updated twice a week **by a central harvester** and the VLO importer. In the event that a metadata record is not accessible anymore via OAI-PMH, the respective entry is kept in the index for at least one week. As a consequence, the available metadata is not static but changes (to a small degree) from day to day. **Requirement** for participation in the central metadata harvesting and VLO import is the **inclusion of the respective endpoint in the CLARIN Centre Registry**[33].

At present, the metadata of approximately 1.1 million resources can be searched in the VLO.

[29] https://vlo.clarin.eu/contributors

[30] https://www.loc.gov/standards/mods/

[31] https://pro.europeana.eu/page/edm-documentation

[32] There are currently around 180 different CMDI schemata publicly available. In addition, there are various "private" schemata.

[33] Which was completed, for example, for the DARIAH-DE repository and the TextGrid repository.

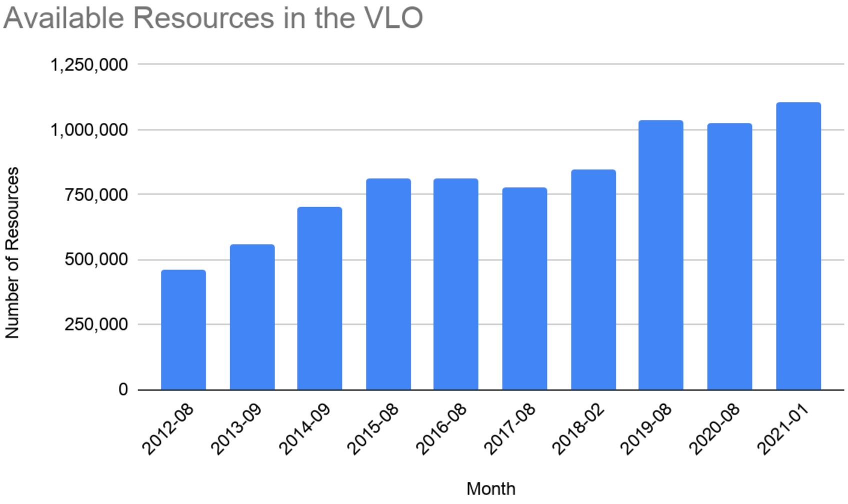


Figure 5: Growth of the amount of resources in the VLO.

3.3 Search Index, Search Logic, and Query Language

The VLO uses the search engine solution Apache Solr in a single node configuration as its data backend, which is based on **Apache Lucene**. The **full text** of every metadata record and selected information types for several dozen data fields are indexed, of which **14** can currently be queried as **search facets**. They are based in their selection and definition on the typical information needs of scientists in language-oriented disciplines (such as search facets for language, multilingualism, resource type, etc.). **The values of these search facets** are **based on manually maintained mappings**[34].

The **user interface is entirely in English** and allows a **full text search** via a central **input field with autocomplete / suggestion functionality**. **Single terms** as well as their **combination via Boolean operators** are supported as search queries. **Search facets** are available on the left side of the user interface and present the ten most frequent values after user interaction. By default, only a subset of search facets is displayed (sorted according to assumed relevance); displaying all search facets requires user interaction ("expansion").

3.4 User Interface and Presentation of Results

Results are presented as lists with (by default) ten results per page, whereby only a partial view of all records with elementary fields (title, description, licence, language) or prominent links (landing page) are shown. **Metadata records with a high similarity** (based on extracted data like title and language) **are recognized as duplicates** and are only displayed in full after user interaction. Lack of accessibility of linked resources is marked on the basis of the ACDH Link Checker.

[34] https://github.com/clarin-eric/VLO-mapping

A click on an entry opens its **detailed description in a tab-based representation**. Tabs are available for the extracted search facets, linked resources, licence information, technical details, representation of the resource hierarchy (if available), and an HTML representation of the entire metadata file.

This **detailed metadata description directly links to relevant data or services**, such as the provider's landing page, data files, querying the resource via CLARIN FCS, processing of the resource via the Language Switchboard etc. **Links to similar resources** in the index are based on similar term vectors using Solr / Lucene functionality[35] and are shown below the resource description.

3.5 Ranking

Ranking of search results is based on the underlying Lucene index. Via specific Solr configurations, **resources with the following properties are given priority**:

- Existence of the fields *Name* and *Description*
- Existence of a landing page at the resource provider
- Availability of (directly referenced) data files
- Open licence or licence allowing academic use
- Position on top of a "resource hierarchy" (e.g. preferring a corpus record over the records of its constituent parts)
- Number of referenced resources (more is better)
- Number of days since the metadata record was imported or updated the last time (less is better)

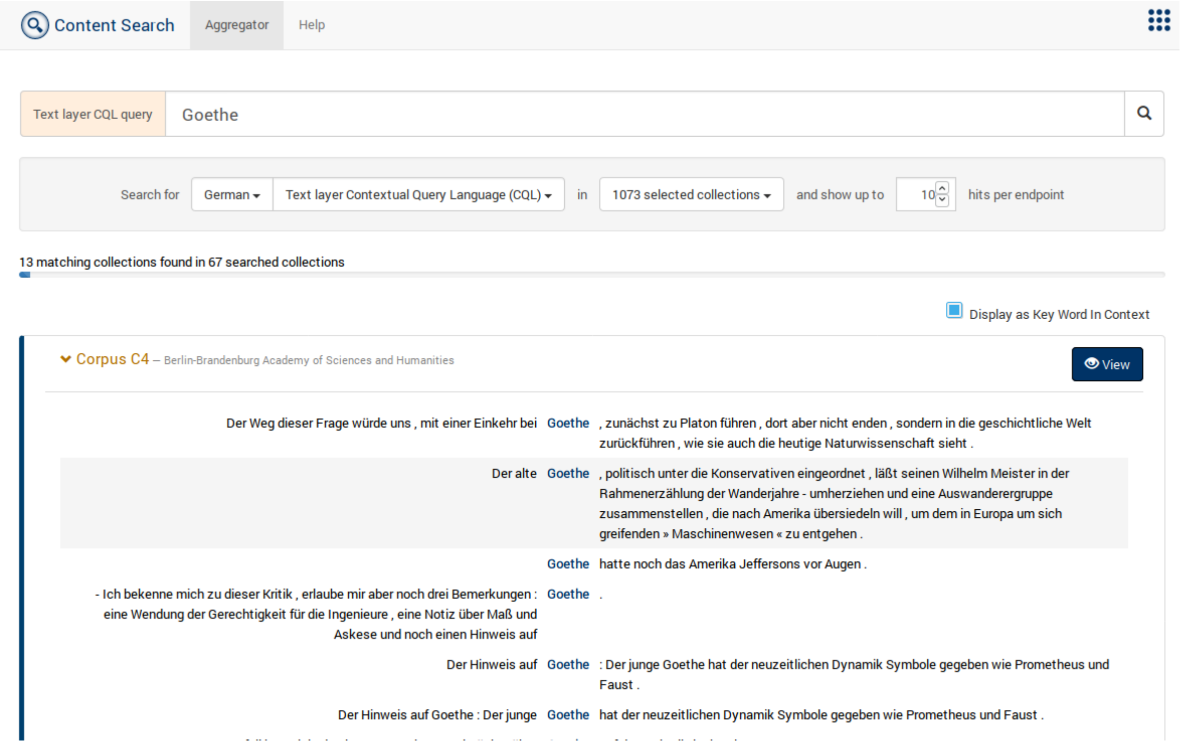
3.6 Customization

The software project is **divided into several subprojects** (including database, data importer, user interface, etc.). The VLO as the central search application of the CLARIN project was not primarily developed to be adapted and installed in other contexts or projects. However, improvements in this direction were made during the EOSC-Hub project[36]. This includes the simplified customization of CSS as well as support of non-English UI languages. There are experiences with the creation and deployment of customized VLO instances, but this usually requires a certain development effort.

4 Federated Content Search (FCS)

The Federated Content Search (FCS) is a federated search engine for examples and references using distributed text corpora from multiple data providers.

[35] https://lucene.apache.org/solr/guide/7_2/morelikethis.html
[36] https://marketplace.eosc-portal.eu/services/virtual-language-observatory



Text layer CQL query: Goethe

Search for German, Text layer Contextual Query Language (CQL), in 1073 selected collections, and show up to 10 hits per endpoint

13 matching collections found in 67 searched collections

☑ Display as Key Word In Context

Corpus C4 – Berlin-Brandenburg Academy of Sciences and Humanities

Der Weg dieser Frage würde uns , mit einer Einkehr bei	Goethe	, zunächst zu Platon führen , dort aber nicht enden , sondern in die geschichtliche Welt zurückführen , wie sie auch die heutige Naturwissenschaft sieht .
Der alte	Goethe	, politisch unter die Konservativen eingeordnet , läßt seinen Wilhelm Meister in der Rahmenerzählung der Wanderjahre - umherziehen und eine Auswanderergruppe zusammenstellen , die nach Amerika übersiedeln will , um dem in Europa um sich greifenden » Maschinenwesen « zu entgehen .
	Goethe	hatte noch das Amerika Jeffersons vor Augen .
- Ich bekenne mich zu dieser Kritik , erlaube mir aber noch drei Bemerkungen : eine Wendung der Gerechtigkeit für die Ingenieure , eine Notiz über Maß und Askese und noch einen Hinweis auf	Goethe	.
Der Hinweis auf	Goethe	: Der junge Goethe hat der neuzeitlichen Dynamik Symbole gegeben wie Prometheus und Faust .
Der Hinweis auf Goethe : Der junge	Goethe	hat der neuzeitlichen Dynamik Symbole gegeben wie Prometheus und Faust .

Figure 6: Federated Content Search (FCS)

4.1 Organisational Context

Like the VLO, the FCS is heavily **integrated into the European CLARIN infrastructure** and is based on its general framework, structures, and goals. This concerns the following:

- The development concept, that primarily relies on developers from different national consortia. At the moment, this includes primarily SWE-CLARIN (with contributions from CLARIN-D) for the central search page of the FCS and distributed work at participating endpoints from a variety of national consortia (including SWE-CLARIN, CLARIN-D, CLARIN Latvia, etc.).
- Specifications that were developed in the context of CLARIN[37] and which are based on and extend the standards SRU[38] (as transport protocol) and CQL[39] (as query language).
- The integration concept which includes the deep integration with other (European) CLARIN applications, like the CLARIN Centre Registry, the Language Resource Switchboard, and the annotation platform WebLicht[40].
- The corporate design in use, which is based on general style guides, colour palettes, and logos of the European CLARIN project.

[37] https://www.clarin.eu/content/federated-content-search-clarin-fcs
[38] http://www.loc.gov/standards/sru
[39] https://www.loc.gov/standards/sru/cql
[40] https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page

- Taken together the aforementioned issues outline the legacy and ownership issues that accompany the FCS and pose a challenge for CLARIAH-DE.

4.2 Available Resources and Resource Inventory

The FCS is a **federated search application for textual data** (or resources that can be serialized as text) in which a **central aggregator**[41] queries participating, **distributed endpoints** and presents their results in an extended list representation. The used data model and the query language FCS-CQL enables querying and representation of different kinds of resources ranging from simple plain text to annotated texts. **Supported annotations include various types of linguistic information**, such as lemma, part of speech, and various transcription or normalization information. The protocol **supports the creation of (arbitrarily complex) hierarchy trees** of resources. This can be used to divide complex resources into their constituents (like a corpus consisting of several sub-corpora).

The FCS aggregator **automatically configures itself on the basis of the CLARIN Centre Registry** and relies on all FCS endpoints listed there[42]. The **resources provided at the endpoints are queried daily** by the aggregator to update its central index. As with the GS, the actual data (text and its annotations) is stored decentralised at the end points.

At the moment, 38 endpoints in 19 locations participate in the search, whereby one endpoint can provide multiple resources. A total of around 180 collections are currently available, from which many are divided into further substructures. Including all subcollections, around 3,500 resources can be accessed. In addition to written text, these also contain other formats such as voice recordings. The new query language FCS-QL is currently only supported by a subset of the endpoints.

4.3 Search Index, Search Logic, and Query Language

Apart from an overview of the supported endpoints and their resources as well as some general metadata (such as the language of the material), the FCS has **no central search index**. The query language and the data model are standardized by the FCS specification and must be supported by all endpoints. Among other things, this often includes the implementation of a wrapper service that translates between the CLARIN-FCS / SRU protocol and the endpoint's corpus engine.

The **query language is similar to standard corpus query languages** (CQL, CQP[43], etc.) in terms of syntax and functionality. This enables, among other things, AND / OR connections of search terms (also within sequences), comparisons based on substrings and searching using regular expressions. In addition, the query language allows the selection of sequences based on the annotation layers supported by the endpoint[44]. **Endpoints provide different sets of annotation layers** (at least token-separated full text) and provide this information according to the specification.

[41] https://contentsearch.clarin.eu

[42] https://centres.clarin.eu/fcs

[43] http://cwb.sourceforge.net/files/CQP_Tutorial/

[44] The query "*[word='car'] [pos='VERB']*", for example, can be used to search for the word "car" followed by any verb.

4.4 User Interface and Presentation of Results

A central search field supports, on the one hand, a simple keyword search and, on the other hand, the formulation of multi-layered FCS-QL queries based on the various annotation layers. Queries can be compiled from modular elements using a graphic interface[45] which also presents the resulting FCS-QL expression. Additional UI elements allow the selection of relevant languages, the collections to be searched and the number of text extracts to be displayed per collection.

For the results, the data model supports a simple keyword-in-context representation (KWIC) as well as the aligned representation of results with all queried annotation layers. Results can be exported in various file formats or sent to the WebLicht platform for further processing. In addition, there is a link to the website of the respective provider of the data collection.

4.5 Ranking

The FCS aggregator **does not rank the results on its own**. Results of the federated query are presented in the order in which they were received and grouped according to the individual queried resources. The **ranking is retrieved fully from the respective endpoint** and is therefore not comparable across resources or endpoints. The FCS aggregator is intended to serve primarily as an exploratory first step in a search, for identifying relevant collections, which can then be examined in more detail in the corpus engine of the concrete data centre.

4.6 Customization

As the central search application of the CLARIN project for text and corpus data, the FCS was not primarily developed to be adapted and installed in other contexts or projects. Preparatory work for customization therefore mainly exists in the context of wrapper services to make it easier for data providers connecting their local infrastructure to the aggregator, but not for porting the aggregator itself.

5 Overlaps and Major Issues

5.1 Organisational and Technical Context

The three CLARIAH-DE search services include 2.2 million resources in total. The content is aimed primarily at researchers from disciplines in the humanities, and cultural and linguistic sciences. The three services presented were created in different research projects and will in part continue to be developed in these. As a consequence, they are designed with a focus on different goals and – often – research-specific requirements. Because of their affiliation to the European CLARIN project and their intended functional complementarity, the VLO and the FCS are more similar to each other than to the

[45] c.f. https://contentsearch.clarin.eu/?&query=%5B%20word%20%3D%20%22goethe%22%20%5D&queryType=fcs

GS, which was created within the framework of DARIAH-DE. A large part of the differences between these services presented below is a result of this situation and the varying development models and goals. An obvious difference is, for example, the focus on different resource types or data models: CMDI-based metadata (VLO), any data model (GS) or linguistically annotated text data (FCS) and, as a result, sometimes different addressed user groups. As a conclusion one has to bear in mind that the different appearances of the services are to some degree connected to differing research-related requirements.

This also affects less flexibility when adapting the CLARIN applications, as these are heavily embedded in (or developed at) the European context. The alignment with a common CLARIAH-DE standard is constrained by the role of these CLARIN applications as central index or aggregator for CLARIN-EU and their extensive integration with other CLARIN applications. Systematically, this would currently only be possible in the form of an – undesirable – separation from the European development and integration model. There is no comparable connection to CLARIAH-DE-external projects for the GS. In addition, its focus on generic data models and the possibility of creating independent instances facilitate various adjustments.

Another difference is the origin of incorporated resources. In the case of the GS and the VLO, content that has been registered in a registry is stored in a central index and updated at regular intervals. In the case of the FCS, a central aggregator is used that queries distributed endpoints. The data is indexed regularly, but is stored decentralised at the respective endpoints.

5.2 Search Syntax and Search / Filter Functionality

Search queries can be entered in all three services via a text field, which has been implemented using different search syntax depending on the service. The FCS uses the query language FCS-QL, which is based on the corpus query language CQP, whereas the GS and VLO projects use Apache Lucene: in the case of the GS the Elastic Search Query Language, which is based on Lucene, and in the case of the VLO, the Lucene Query Parser. As a result, many similarities can be found between the GS and the VLO.

Features present in all three services include some common search options. For example, all services support Boolean operators, albeit with slightly different syntax. In addition, phrases can be used in search queries for all three systems.

More search options can be used before and after the search query. On the start page, the FCS only offers filters for language, query language or collection to be searched. The GS in the "Advanced Search" allows the selection of facets based on the data model of the respective collection. The VLO in turn offers a number of predefined search facets, for example for selecting suitable languages, formats or resource types. All systems also support filtering on the result pages. This is explained in more detail in the following section.

5.3 Presentation of Results

A search activity returns a list of results in all three systems, which can be examined in different ways, but can also be further refined.

In all services, in addition to the specific resource, the collection in which the data record is embedded is highlighted and made accessible via direct links. The FCS specifically returns a list with the collections in which the search results are located, the VLO shows the metadata of the respective collections directly in the user interface via a dedicated "hierarchy" tab, and the GS offers an overview of the returned collections also via a specific tab.

Further options for modifying the presentation include the number of results per page, available for all three services, or the use of additional filters, which vary in scope depending on the service. This includes for example the collection to be searched, the keywords, the language or even the query language. A detailed list of the services' options is given in the overview table in the appendix.

5.4 Corporate Design

The search services presented here (GS, VLO and FCS) have primarily been developed and expanded within the DARIAH (for GS) or CLARIN infrastructure (for VLO and FCS) with the participation of various partners and are therefore operated in different organizational and technical frameworks. Various European projects were actively involved in the development of the CLARIN services, who, among other things, use a common CLARIN style guide[46]. As a consequence, a redesign of the three independent applications using a joint CLARIAH-DE corporate design or the adaptation of their functionality to CLARIAH-DE requirements is a non-trivial task.

6 Realized Harmonizations

From the showcased search services and, particularly, their differences, it is evident that a complete harmonization entails significant effort. This section will present the concrete adaptations of services that have been realized up to this point.

6.1 Cross-Service Search: Integrated Access to the Search Services

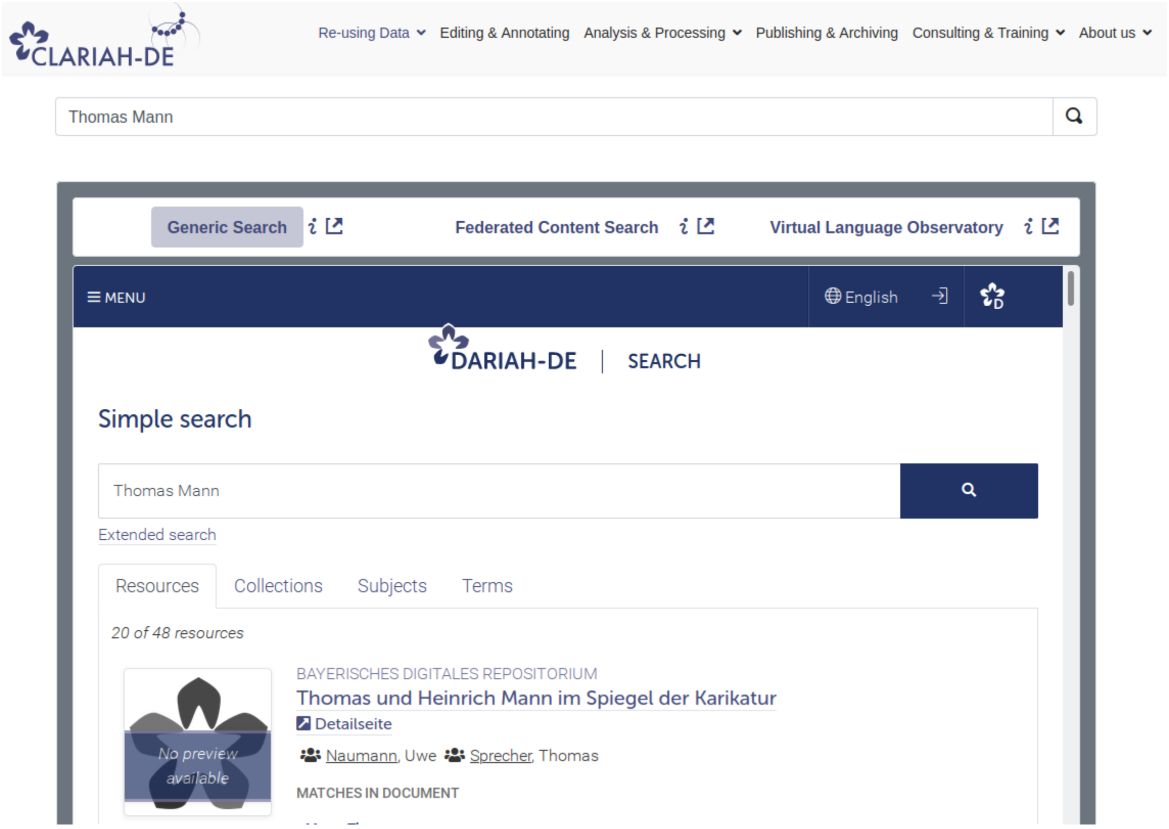
Following the principle of a uniform branding, a web portal was developed in the context of CLARIAH-DE which enables search functionalities spanning the three search services (GS, VLO and FCS) and was integrated into the CLARIAH-DE website under the menu item *Cross-Service Search*[47]. The portal is designed in line with the CLARIAH-DE corporate design, allowing for a high recognition value. Like

[46] https://www.clarin.eu/content/style-guide

[47] https://clariah.de/en/re-use-data/find-data/clariah-de-cross-service-search

other content of the CLARIAH-DE website, the portal for search and research services is available in German and English.

The portal offers important information about the three services, which are each integrated as an iframe and with a text field for search queries. After executing the search query, the results will be displayed inside the iframes and it is possible to switch between the three search services at any time. The search results of a particular service can also be opened in a new tab. An information button can provide background information for the currently active search service and also offers assistance for formulating search queries.



Thomas Mann

Simple search

Thomas Mann

20 of 48 resources

BAYERISCHES DIGITALES REPOSITORIUM

Thomas und Heinrich Mann im Spiegel der Karikatur

Naumann, Uwe Sprecher, Thomas

Figure 7: Results presentation of the search services GS, FCS and VLO as integrated iframes, by example of the term "Thomas Mann".

The search results of the embedded individual services are held in the corporate design of their respective developing group – DARIAH-DE and CLARIN respectively. The iframe integration emphasizes the nature of CLARIAH-DE: the different services' affiliations are clearly distinguishable, however, their embedding into the CLARIAH-DE corporate design promotes the merger of DARIAH-DE and CLARIN-D.

6.2 Generic Search as an Endpoint of the Federated Content Search

Besides search via metadata, the GS also encapsulates access to full text resources, provided they are available in affiliated collections. Thus, the GS enables search over a multitude of full texts. However, due to its previous focus, it cannot offer dedicated query options beyond the scope of the Elasticsearch query language. To close this gap and allow for language oriented queries for full text, the GS implements a CQL-based interface which enables the FCS to access full text content from the GS, thus making them available for such queries. A preliminary feasibility study concerning the connection of the GS as an FCS endpoint was completed with the implementation of the *Basic Search*[48] functionality based on CQL. A further implementation of the interface to support additional layers will be pursued.

6.3 Educational and Training Materials with Generic Search

The CLARIAH-DE Tutorial Finder was devised and implemented in the summer of 2020 and provides a solution for making searchable educational and training materials available in text and video. The search solution builds on components of the GS, i.e. the Collection Registry CR for the registration and description of the collections and the DME for the modelling of the collection descriptions. By integrating YouTube videos as a multimedia resource type a use case can be presented here that bridges text-based resources with other types of media.

The harmonization effort can be illustrated using two characteristics. On the one hand, the joint search across educational and training materials represents a search engine that is positioned from the start to include data both from CLARIN and DARIAH, like the collection of the teaching material collection TeLeMaCo[49] from CLARIN and the DARIAH-Campus[50]. This data will be aggregated and made searchable after registration in the respective components. The search therefore constitutes a new service for CLARIAH-DE with a comprehensive import of data from both preceding projects. On the other hand, the common corporate design of CLARIAH-DE was realized in a search service, with its corresponding user interface elements, logo, and fundamental user guidance. The theme was created similarly to the previously existing DARIAH theme and is available via GitHub[51].

6.4 Adaptations of VLO and FCS

During the problem analysis, the difficulty of low adaptability was identified on the parts of the CLARIN applications (VLO and FCS), due to the context of their developments. Furthermore, their roles as central European entry points into resource collections means that the creation and operation of parallel instances and entailing necessary adaptations were not part of the original development focus. To minimize these problems, preliminary groundwork was made for the VLO.

This work was done both in the context of CLARIAH-DE as well as part of the general development work on the VLO. The latter also includes work on the VLO integration into the European Open Science Cloud (EOSC), like the incorporation into the EOSC-Hub Marketplace[52]. More precisely, the following work was carried out:

[48] https://office.clarin.eu/v/CE-2017-1046-FCS-Specification.pdf, Section 2.2.1

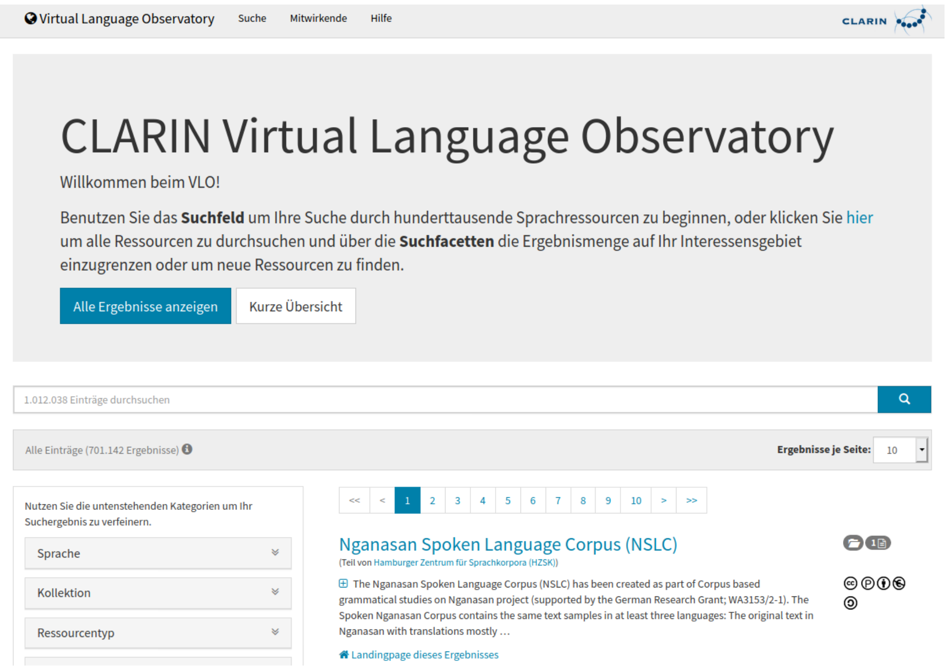
[49] https://fedora.clarin-d.uni-saarland.de/hub/browse/

[50] https://campus.dariah.eu/

[51] https://gitlab.rz.uni-bamberg.de/dariah/themes/clariah-theme

[52] https://www.eosc-hub.eu/clarin-vlo

- Modification of the VLO for simplified internationalization / localization, in part to support non-English user interfaces[53] and localization of size and time measurements
- Easier deployment of the application by utilizing Docker-based container technology
- Easier access / edit of style sheets and improved documentation.



Virtual Language Observatory
Suche
Mitwirkende
Hilfe
CLARIN

CLARIN Virtual Language Observatory

Willkommen beim VLO!

Benutzen Sie das **Suchfeld** um Ihre Suche durch hunderttausende Sprachressourcen zu beginnen, oder klicken Sie hier um alle Ressourcen zu durchsuchen und über die **Suchfacetten** die Ergebnismenge auf Ihr Interessensgebiet einzugrenzen oder um neue Ressourcen zu finden.

Alle Ergebnisse anzeigen
Kurze Übersicht
1.012.038 Einträge durchsuchen
Alle Einträge (701.142 Ergebnisse)
Ergebnisse je Seite: 10
Nutzen Sie die untenstehenden Kategorien um Ihr Suchergebnis zu verfeinern.
Sprache
Kollektion
Ressourcentyp
<< < 1 2 3 4 5 6 7 8 9 10 > >>

Nganasan Spoken Language Corpus (NSLC)

(Teil von Hamburger Zentrum für Sprachkorpora (HZSK))

The Nganasan Spoken Language Corpus (NSLC) has been created as part of Corpus based grammatical studies on Nganasan project (supported by the German Research Grant; WA3153/2-1). The Spoken Nganasan Corpus contains the same text samples in at least three languages: The original text in Nganasan with translations mostly …

Figure 8: German localization of the VLO.

For the FCS, a preliminary status analysis was conducted concerning analogous work.

7 Lessons Learned and Perspectives

In this concept paper, the essential differences and similarities of the existing CLARIAH-DE search services as well as the reasons for their variety were presented.

CLARIN-D and DARIAH-DE have developed their services over the years in similar environments and research disciplines, but nevertheless for different user groups. This pertains the research-specific requirements (or research interests) as well as the scope of data. For rather generic components, such as AAI, integration and joint solutions are easier to achieve than for specialised scientific services such as the described search functionality. The focus on different target groups and information needs resulted in a functional encapsulation of the search services, each having a different thematic focus. The complex organizational and technical initial situation leads to a rather high integration and harmonization effort. Various forms of measures were identified, which are implemented on different levels. They range from

[53] https://github.com/clarin-eric/VLO/issues/263

the simple adaptation of style sheets to complex and time-consuming adaptations of the data models and query languages used. Various preliminary work has already been implemented as part of the work package or is still in active development.

The analysis and completed work demonstrate the problems when integrating established applications into a larger, shared infrastructure. Lessons learned can also be helpful for future integration work in the context of national and international research infrastructures. This specifically concerns the participation of CLARIAH-DE in the Text+ consortium as part of the National Research Data Infrastructure (NFDI)[54] and the integration in the context of European projects, such as the aforementioned European Open Science Cloud or the Social Sciences and Humanities Open Cloud[55] (SSHOC). They include the following:

- The highly differentiated target groups require diverse search functionalities in order to satisfy different research interests and tasks. These differ among others with respect to the selected preparation / presentation, available query options, and the selection of provided resources.
- The use of standardized interfaces and protocols (such as Apache Lucene, OAI-PMH, FCS etc.) is a necessary prerequisite for successful and systematic integration measures[56].
- The potential use of a tool outside of the originally intended context (like a specific department, community, etc.) must already be taken into account in the design phase and during initial development[57]. The close connection between CLARIAH-DE and the various scientific communities represents a major advantage that has to be used in all future developments.
- The integration of existing user interfaces in portals or in completely new contexts is made significantly easier by the support of "rebranding" (e.g. by exchanging respective style information). Ideally, this can be controlled and modified for a running system via parameters.
- The use of standard software facilitates the integration and development of harmonized applications. As an example in this context, the standard indexing software Apache Lucene, which is used as a data backend in both the VLO and the GS, simplifies complex integration measures, due to the same underlying query language and a compatible data model.
- The option of directly accessing the respective data backend (e.g. Apache Lucene) allows integration tasks down to the data model level. In this sense, individual user interfaces can only be facades in front of a highly integrated data inventory.
- The use of iframes in the implementation of the Cross-Service Search portal allows independent content to be easily and effectively integrated into the common corporate design of a new portal and presented in parallel, despite varying technical requirements. However, there is no control over the content, design and responsiveness of the websites that are linked via the iframes.

[54] https://www.dfg.de/en/research_funding/programmes/nfdi/index.html

[55] https://www.sshopencloud.eu

[56] In this context, it allowed, among other things, the direct connection of the DME to the Federated Content Search based on the FCS specification.

[57] As an example, the focus of the CLARIN applications on a purely English localization turned out to be problematic.

- The user perspective should be at the centre of all technical harmonization considerations so that in case of doubt a solution in terms of "user interfaces as facades in front of a highly integrated data inventory" can be the method of choice. The different views on an inventory are usually due to the requirements of the specific scientific community and must not be lost during integration.
- A – however deeply – integrated solution to accessing and representing the complete resource inventory at a single place, as shown prototypically in the CLARIAH-DE portal, is important to show connectivity and expandability to other actors, especially in the context of EOSC and NFDI. Such an access point can function as a technical and "organizational" invitation to others.

On the basis of the analysis carried out, concrete future work areas include:

- Expansion of the specialized search solutions with a focus on the data domains addressed by the Text+ consortium. Adaptation and further development of the FCS for lexical resources is already planned, as well as increased support for research question or subject-specific search solutions, as it has already been implemented in the form of the CLARIAH-DE teaching and training material collection.
- Further expansion of theming and localization capabilities of all discussed applications; creation and improvement of appropriate documentation to simplify these tasks.
- Increased consideration of the different research focus of the individual applications in the entire CLARIAH-DE infrastructure making use of the contrast between "global" search applications and detailed searches, especially for various research questions or searches in specific fields of study; increased user-oriented integration of search solutions into the infrastructure.
- Further development of the existing infrastructure to improve connectivity to relevant external research infrastructures (including other NFDI consortia such as NFDI4Culture[58] or the European Open Science Cloud EOSC).

[58] https://nfdi4culture.de

8 Annex

Table 1: Comparative overview of the three search systems.

	Federated Content Search (FCS)	Generic Search (GS)	Virtual Language Observatory (VLO)
Aggregated Search in all Systems	https://www.clariah.de/dienste.html	https://www.clariah.de/dienste.html	https://www.clariah.de/dienste.html
URL	https://clariah.de/en/re-use-data/find-data/clariah-de-cross-service-search	https://clariah.de/en/re-use-data/find-data/clariah-de-cross-service-search	https://clariah.de/en/re-use-data/find-data/clariah-de-cross-service-search
Origin	CLARIN	DARIAH-DE	CLARIN
Facets	On the entry page: Query language, language, collection to be searched, number of results After search query: Same number of facets as before the search	On the entry page: None After search query: Number of results, searched collections via “Advanced Search”: Dublin Core Elements	On the entry page after scrolling down or clicking “See all records”: 14 facets After search query: Same number of facets as before the search
Search syntax	Query language “FCS-QL”, which is based on the Corpus Query Language CQP: https://office.clarin.eu/v/CE-2017-1046-FCS-Specification.pdf	Based on Elasticsearch Query Language: https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-simple-query-string-query.html	Based on Lucene Query Parser: https://lucene.apache.org/core/2_9_4/queryparsersyntax.html
Search syntax - Boolean operators	Version 1.0 (CQL): “and” | “or” | “not” | “prox” (Lower case according to CQL specification) Version 2.0 (FCS-QL): Restrictions on “token stream”, “&” und “|”	+’ and “|”	Boolean operators “AND”, “OR”, “NOT”, “+” and “-”. (according to Lucene specification)

	Federated Content Search (FCS)	**Generic Search (GS)**	**Virtual Language Observatory (VLO)**
Search syntax - Phrases	In CQL, a term can be a single token or a phrase, i.e. tokens separated by spaces	Phrases with "", see Elastic Search	Phrases with "" and ranges with "TO"
Syntax search – More	"Searching in annotated data that is represented in annotation layers." Ex.: Lemma, POS, phonetic, etc.	See Elastic Search	Search in certain fields: Language, country, continent, modality, genre, subject, format, organisation, resource type, keyword, resources
Result limitation	10 – unlimited, but a maximum of 250 per endpoint	10-150	5, 10, 25, 50, 100
Result display (list)	Collections, with collapsible details pane on the hits in the documents. Link to the collection	Tabs: Resources, collections, subjects, terms List of hits, information about: Collection, document, type and link to the hit	List of documents in which the search query was found with expandable metadata for the record, number of available documents for the record, link to the record, licence information.
Result display (Single document)	Link to the collection -> Pop-up with metadata, highlighted hits in the documents, download option and further links	Redirect to the document found -> Redirect to new page Link(s) to resource, data record in the original, link to collection and preview / thumbnail if available	Redirect to the document found -> Redirect to new page Metadata for the hit, further links, "more like this" hits, "All metadata"

	Federated Content Search (FCS)	**Generic Search (GS)**	**Virtual Language Observatory (VLO)**
Results navigation	After closing a document of the results (pop-up), return to the list of results	After clicking on "Back" to return to the list of results (if you are on the page of a found document): Search has not been saved and search bar & search results are empty	After clicking on "Back" to return to the list of results (if you are on the page of a found document): Search has been saved and displays the result list again
Language selection	With facet "Language"	Via "Advanced Search"	After successful search in facet "Language"
Options on homepage	Filter for language, request language, collection, number of results	Link to advanced search, tag cloud incl. links of keywords	Button to open more options ("See all records")
Information on homepage	Number of collections	Number of collections and documents, tag cloud of keywords	Number of documents (records)
Automatic suggestions for search terms	No	No	Yes
Loading bar / progress bar	Loading bar with collections & documents found so far	Loading symbol (animated wheel)	Loading symbol (animated wheel)
Localization	English	German & English	English

Bibliography

Buddenbohm, Stefan. "CLARIAH-DE - Aligning two Research Infrastructures: Experiences and Challenges". Talk at the Scholarly Primitives - DARIAH Annual Event 2020, Zagreb, Croatia.

Eckart, Thomas, and Tobias Gradl. "Working towards a Metadata Federation of CLARIN and DARIAH-DE". CLARIN Annual Conference 2017, Budapest, Hungary, 2017.

Goosen, Twan, and Thomas Eckart. "Virtual Language Observatory 3.0: What's New?". CLARIN Annual Conference 2014, Soesterberg, The Netherlands, 2014.

Henrich, Andreas, and Tobias Gradl. "The Integration of Research Data. How Can Research Infrastructures Help?". In Eva-Marie Seng, Frank Göttmann (Eds.): DOKUMENT - OBJEKT - GENESE: DE GRUYTER, S. 767–802.

Henrich, Andreas, Robin Jegan, and Tobias Gradl. "Data Retrieval". Praxishandbuch Forschungsdatenmanagement. De Gruyter Saur, 2021, 427-449. https://doi.org/10.1515/9783110657807

Klammt, Anne, and Roberta Toscano. "CLARIAH-DE. Ein Beitrag zur Entwicklung einer wissenschaftsgeleiteten Forschungsinfrastruktur für die text- und sprachbasierten Geisteswissenschaften". Archivar - Zeitschrift für Archivwesen 73, 2020, 25-30. https://www.archive.nrw.de/archivar/hefte/2020/Ausgabe-1/Archivar-1_2020.pdf

Olsson, Leif-Jöran. "Federated Content Search Engine v2". CLARIN-PLUS Deliverables 2.9, 2017. https://office.clarin.eu/v/CE-2017-1035-CLARINPLUS-D2_9.pdf

Wilkinson, Mark, Michel Dumontier, IJsbrand Jan Aalbersberg, et al. "The FAIR Guiding Principles for scientific data management and stewardship". Sci Data 3, 160018, 2016. https://doi.org/10.1038/sdata.2016.18