Merging Subject-Specific Searches of CLARIN and DARIAH in CLARIAH-DE: Challenges of Technical Integration

Stefan Buddenbohm, RDD, Göttingen State and University Library, sbudden@gwdg.de

Thomas Eckart, NLP Group, Leipzig University, teckart@informatik.uni-leipzig.de

DANS-CLARIAH.NL-Seminar, 2021-03-23

Slides: 10.5281/zenodo.4628889

clariah.de

@CLARIAHde

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
FÖRDERKENNZEICHEN 01UG1910 A bis I

Stefan Buddenbohm

Project officer, Göttingen State and University Library

sbudden@gwdg.de

https://orcid.org/0000-0002-3469-6101

#DARIAH-DE, CLARIAH-DE Office, DARIAH-EU, DINI e.V., SSHOC, research infrastructures, quality assurance for publication repositories

Dr. Thomas Eckart

NLP Group Leipzig University

teckart@informatik.uni-leipzig.de

http://asv.informatik.uni-leipzig.de/staff/Thomas_Eckart

#CLARIN-D, CLARIAH-DE Lead Infrastructural Work Package, CLARIN, research infrastructures, metadata infrastructure

Structure (30 minutes)

1. What is CLARIAH-DE? (very briefly!)
2. Levels of challenges in merge scenarios
3. CLARIAH-DE levels of integration (basic infrastructure, application, research data)
4. Examples from the three categories (in depth: merging subject-specific searches)
5. Experiences & lessons learned

Sources:

➔ Slides: 10.5281/zenodo.4628889

➔ *Eckart, Thomas et al.* (to be published): CLARIAH-DE Cross-Service Search: Prospects and Benefits of Merging Subject-specific Services". *DARIAH-DE Working Papers. Göttingen.*

➔ Result poster AP4 Technical Integration: https://zenodo.org/record/4572610

➔ Project website: https://www.clariah.de/

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Introduction & Background of CLARIAH-DE

clariah.de

@CLARIAHde

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
FÖRDERKENNZEICHEN
01UG1910 A bis I

Funded by the BMBF 2019-2021 - 3.3 mio. € - 13 funded partners - Merger project CLARIAH-DE

Building blocks for the NFDI SSH consortia

Language- and text-based research data

Sustainable provision of research data and technical infrastructure, digital tools, teaching material

Levels of Challenges in Merge Scenarios

clariah.de

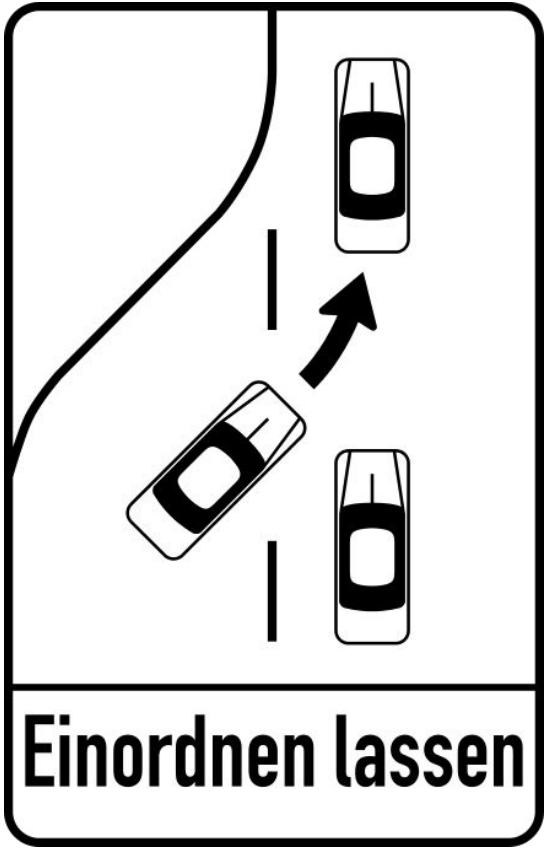
@CLARIAHde

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
FÖRDERKENNZEICHEN
01UG1910 A bis I

Level of Challenges

- further reading: https://en.wikipedia.org/wiki/Organizational_theory

- Cultural: Specific organisational structures are reflected in decision-making and working structures. Tendencies of perseverance of the status quo are common.
- Structural: Although CLARIAH-DE is about merging, both CLARIN-D and DARIAH-DE have to maintain roles on a European level: ERIC.
- Technological: Different research communities = different solutions (handling of metadata, organising data: fulltexts, collections, differing curation approaches, ownership of data).
    - For generic infrastructure components synergies may be quick off the mark, for research-nearer components not. Provider-related benefits of merging (lower maintenance) easier to achieve than scientific-related ones.



Source: https://commons.wikimedia.org/wiki/File:Hinweiszeichen_23c.svg (Public domain)

CLARIAH Levels of Integration

clariah.de

@CLARIAHde

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
FÖRDERKENNZEICHEN
01UG1910 A bis I

CLARIAH-DE

Levels of Integration

	Complexity / Effort	Standardization	Visibility / User expectation	External dependencies
Basic infrastructure components	-	+	o	o
Applications	++	o	++	+
(Research) data	++	-	++	++

Integration Examples

Basic Infrastructure Integration - Application Integration - (Research) Data Integration

clariah.de

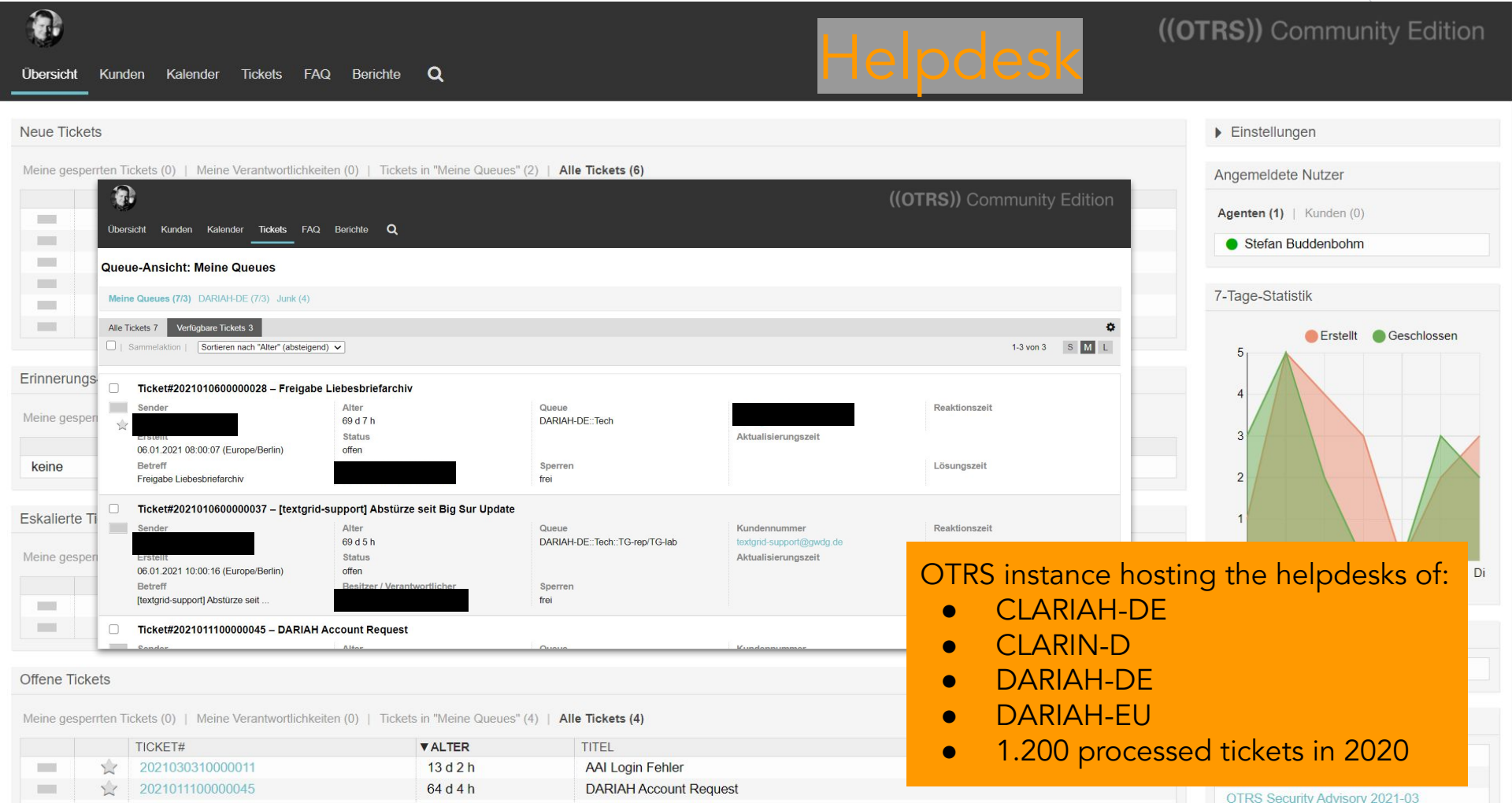
@CLARIAHde

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
FÖRDERKENNZEICHEN
01UG1910 A bis I

Basic infrastructure components

- evident category of infrastructure components for merging because of its low complexity (in comparison) and standardisation
- visibility and appreciation by users usually lower (in comparison)
- both CLARIN and DARIAH developed similar solutions for requirements like user support, access to resources, or maintaining services
    - OTRS ticket system
    - AAI federation with LDAP for group management
    - Icinga-based monitoring

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung


Übersicht
Kunden
Kalender
Tickets
FAQ
Berichte
Helpdesk
((OTRS)) Community Edition
Neue Tickets
Meine gesperrten Tickets (0) | Meine Verantwortlichkeiten (0) | Tickets in "Meine Queues" (2) | Alle Tickets (6)
Einstellungen
Angemeldete Nutzer
Agenten (1) | Kunden (0)
Stefan Buddenbohm
7-Tage-Statistik
Erstellt
Geschlossen
Queue-Ansicht: Meine Queues
Meine Queues (7/3) DARIAH-DE (7/3) Junk (4)
Alle Tickets 7
Verfügbare Tickets 3
Sortieren nach "Alter" (absteigend)
1-3 von 3
Ticket#2021010600000028 – Freigabe Liebesbriefarchiv
Sender
Alter
69 d 7 h
Queue
DARIAH-DE::Tech
Reaktionszeit
Erstellt
06.01.2021 08:00:07 (Europe/Berlin)
Status
offen
Aktualisierungszeit
Betreff
Freigabe Liebesbriefarchiv
Sperren
frei
Lösungszeit
Ticket#2021010600000037 – [textgrid-support] Abstürze seit Big Sur Update
Sender
Alter
69 d 5 h
Queue
DARIAH-DE::Tech::TG-rep/TG-lab
Kundennummer
textgrid-support@gwdg.de
Reaktionszeit
Erstellt
06.01.2021 10:00:16 (Europe/Berlin)
Status
offen
Aktualisierungszeit
Betreff
[textgrid-support] Abstürze seit ...
Besitzer / Verantwortlicher
Sperren
frei
Ticket#2021011100000045 – DARIAH Account Request
Offene Tickets
Meine gesperrten Tickets (0) | Meine Verantwortlichkeiten (0) | Tickets in "Meine Queues" (4) | Alle Tickets (4)
TICKET#
▼ALTER
TITEL
2021030310000011
13 d 2 h
AAI Login Fehler
2021011100000045
64 d 4 h
DARIAH Account Request
OTRS instance hosting the helpdesks of:
CLARIAH-DE
CLARIN-D
DARIAH-DE
DARIAH-EU
1.200 processed tickets in 2020

gn in via the CLARIN Service Provider Federation

AAI federation

where you work or study. Signing in here will allow you to
ble to users who have logged in. If you cannot find your
unt and use your CLARIN website credentials. If you don't
please contact spf@clarin.eu.

Previously chosen home organisation

DARIAH
Germany

Home organisation list

Search for your home organisation...

All countries

clarin.eu website account
European Union

AAI@EduHr Single Sign-On Service
Croatia

Aalborg University
Denmark

DARIAH AAI AUTHENTICATION AUTHORISATION INFRASTRUCTURE

DE: please confirm access permissions

Geo-Browser needs to access data from **DARIAH-DE Storage**

You need to grant the following permissions to use the application:

- read
- write

GRANT PERMISSIONS CANCEL

DARIAH-DE in 2020:

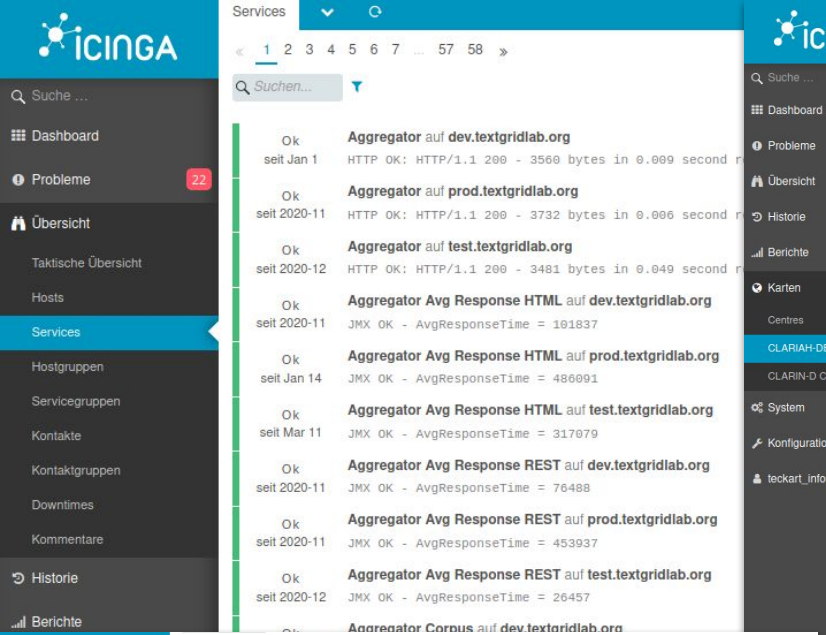
- 43.000 user interactions
- 3.600 accounts
- 104 organisations in the AAI

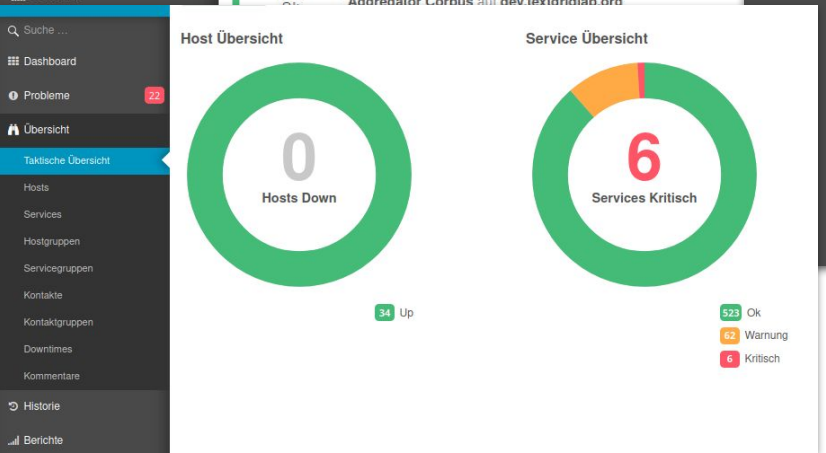
DARIAH-DE

Digitale Forschungsinfrastruktur fÃ¼r die Geistes- und Kulturwissenschaften

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

FÃ¶rderkennzeichen 01UG1610A bis J





3 Screenshots from the Icinga system.

Fun fact: Geolocalisation of services is easier for CLARIN than for DARIAH.

Monitoring

Application Integration

- evident category for integration IF use cases are compatible and promising because the user gains access to a broader methodical portfolio (i.e. tools)
- with higher visibility comes (potentially) higher appreciation by users
- but are higher complexity and effort result in trade-offs between user and provider benefits
- importance of research-specific requirements or customs

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

1. https://textgridrep.org/browse/tbz8.0

Metadata

File Type
text/xml

PID
hdl:11858/00-1734-0000-0004-916D-3
Citation Suggestion

Download

Object (TEI)
Metadata (XML)
Tech. Metadata (XML)
Plain Text (txt)
E-Book (epub)
HTML
ZIP

Tools

2.

Voyant
Annotate
Switchboard

[523] [577]Thomas Mann's »Buddenbrooks«

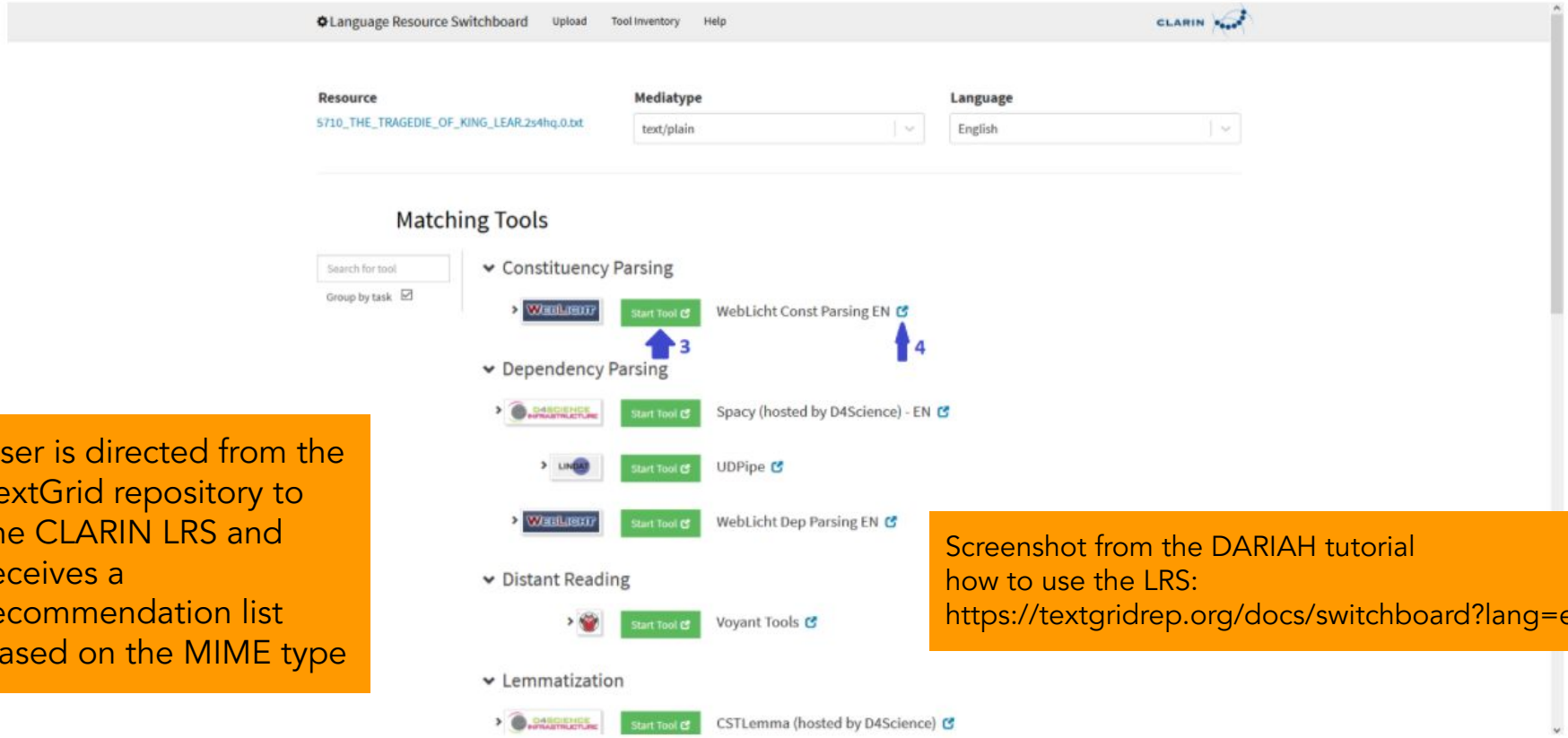
Man wird sich diesen Namen unbedingt notieren müssen. Mit einem Roman von elfhundert Seiten hat Thomas Mann einen Beweis von Arbeitskraft und können gegeben, den man nicht übersehen kann. Es handelte sich ihm darum, die Geschichte einer Familie zu schreiben, welche zugrundegeht, den »Verfall einer Familie« Noch vor einigen Jahren hätte ein moderner Schriftsteller sich damit be eigen, den Letzten, der an sich und seinen Vätern stirbt. Thomas Ma n Schlußkapitel die Katastrophe zusammenzudrängen, an welch hat, gewissenhaft, dort begonnen, wo [577] der höchste Glücksst r diesem Höhepunkt notwendig der Abstieg beginnen muß, ers äher und jäher und schließlich senkrecht abfallend in das Nichts.

So war er also vor die Notwendigkeit gestellt, das Leben von vier Generationen zu erzählen, und die Art wie Thomas Mann diese ungewöhnliche Aufgabe gelöst hat, ist so überraschend und interessant, daß man, obwohl es Tage lostet, die beiden gewichtigen Bande Seite für Seite mit Aufmerksamkeit und Spannung liest ohne zu ermüden, ohne etwas zu überschlagen, ohne das geringste Zeichen von Ungeduld oder Eile. Man hat Zeit, man muß Zeit haben für die ruhige und natürliche Folge dieser Begebenheiten; gerade weil nichts in dem Buche für den Leser da zu sein scheint, weil nirgends, über die Ereignisse hinweg, ein überlegener Schriftsteller sich zu dem überlegenen Leser neigt, um ihn zu überreden und mitzureißen, – gerade deshalb ist man so ganz bei der Sache und fast persönlich beteiligt, ganz als ob man in irgend einem Geheimfach alte Familienpapiere und Briefe gefunden hatte, in denen man sich langsam nach vorn liest, bis an den Rand der eigenen Erinnerungen.

User gets offered tools for further examination of the data. Note: all tools are supported by tutorials (the “i”).

Using tools

1. https://beta-switchboard.clarin.eu/tools

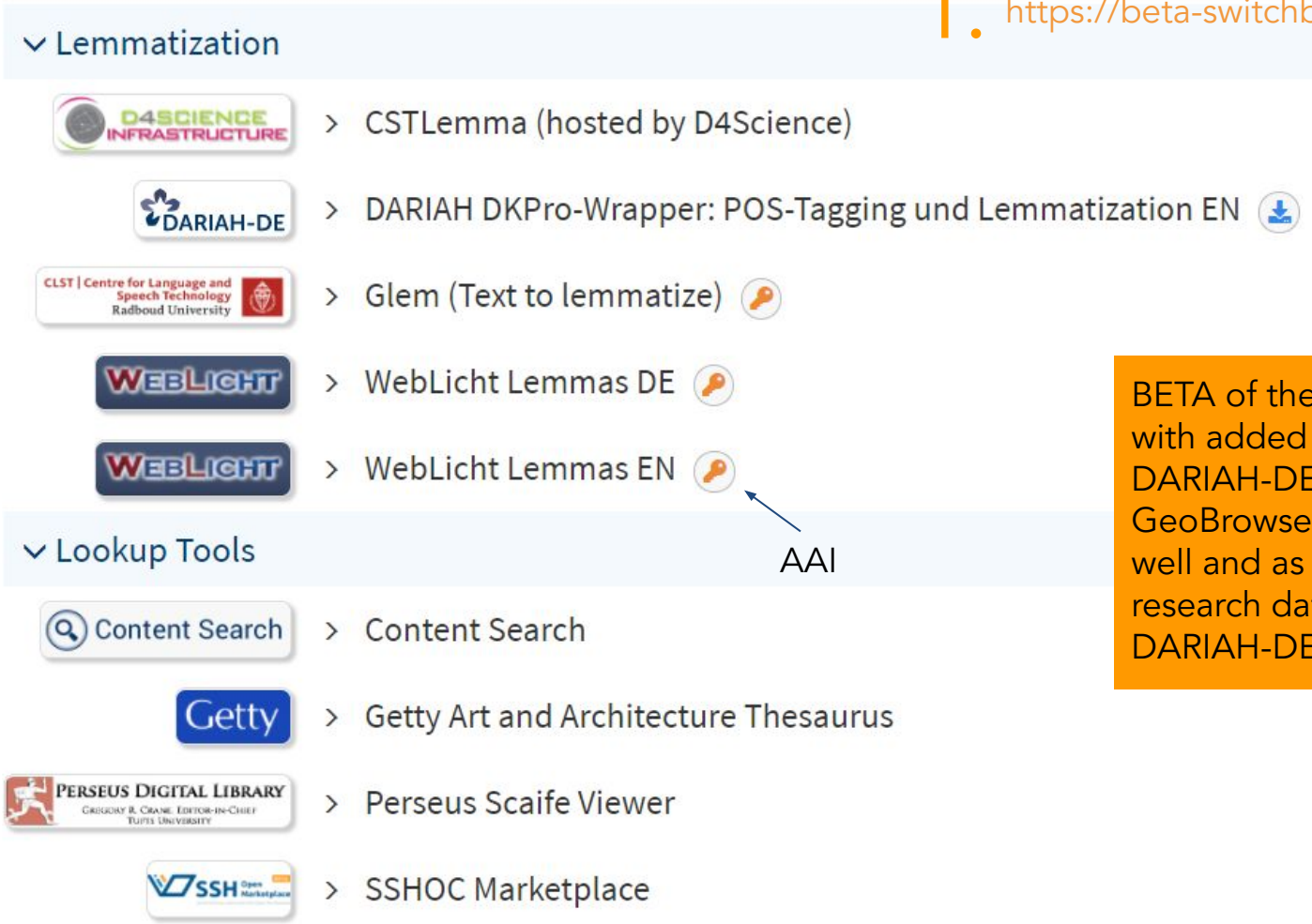


User is directed from the TextGrid repository to the CLARIN LRS and receives a recommendation list based on the MIME type

Screenshot from the DARIAH tutorial how to use the LRS: https://textgridrep.org/docs/switchboard?lang=en

To apply a tool to your selected text, click on the green "Start Tool" button (arrow 3). You will then be redirected to the website of the tool that has already loaded your text and prepared it for processing. Sometimes the processing starts automatically, sometimes you have to start it by yourself. If you wish documentation of the tools, please click on the blue symbol next to the tool name (arrow 4).

1. https://beta-switchboard.clarin.eu/tools



BETA of the CLARIN LRS with added services. From DARIAH-DE side the GeoBrowser is included as well and as source for research data the DARIAH-DE repository.

(Research) Data Integration

- most difficult category due to the research-specific requirements towards the organisation, description, presentation and usage of research data
- legacy issue: most research data resources are linked with other resources
- high complexity and effort resulting in trade-offs between research-specific requirements and a possibly reduced granularity of research data presentation
- often high on the “wish list” of users

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Search services	CLARIN-D	DARIAH-DE
Stock	VLO: 1.200.947 mio. records; FCS: 34 endpoints at 16 institutions querying 4000 collections	DARIAH: 47 collections with well over 1.2 mio. resources; MWW: 27 collections with well over 250k resources; CLARIAH tutorial finder: 6 collections with 554 resources
Organisation	Two searches: Virtual Language Observatory / Federated Content Search 1. CLARIN ERIC as maintaining entity 2. CLARIN Centre Registry (=endpoints harvested via OAI-PMH) 3. LRS & VCR integrated into the search 4. CLARIN Metadata Curation Module	Concept with three layers: 1. Federation layer: CR (=collection registry) and DME (=mapping between data models) 2. Data layer: the accessible collections 3. Service layer: Generic Search (GS) and other services (e.g. Geo-Browser)
Types of resources	FCS focussing on full texts and corpora; VLO: metadata describing text- and language corpora, treebanks, dictionaries; including tools such as taggers, classificators and web-services	Scientific collections; mostly textual data but not limited to it; tailored services such as Geo-Browser (GIS) or Cosmotool (biographical information)
Inter-operability	VLO: CMDI as common ground for all CLARIN data centres (currently 180 public CMDI schemas); deep technical integration of components with one another and within CLARIN; FCS relies on enabled access to the textual content of data sets	DCDDM - DARIAH Collection Description Data Model; simple and extended search relying on DCsimple, no data centres
Search	Solr; 14 search facets based on https://github.com/clarin-eric/VLO-mapping; FCS Query Language	Elasticsearch; Elasticsearch Query Language; DCsimple for faceting search results; Customization, e.g. MWW

Searches | Data | Help | English | DARIAH

DARIAH-DE | SEAI

1. https://search.de.dariah.eu/search/

Extended search

2.

Data structures

oai_dc

Search options

Show explanations

20 Results per page

Queried collections

1 TextGrid Digitale Bibliothek: Literatur

Göttinger Digitalisierungszentrum – Bucherhaltung

Göttinger Digitalisierungszentrum – DigiWunschbuch

Göttinger Digitalisierungszentrum – Göttinger Universitätsgeschichte - Gedruckte Werke

Göttinger Digitalisierungszentrum –

|→ Title | Buddenbrooks

Simple search

+ ADD FACET

Resources | Collections | Subjects | Terms

1 of 1 resources

No preview available

TEXTGRID DIGITALE BIBLIOTHEK: LITERATUR

Thomas Mann's »Buddenbrooks«

Detailseite

Rilke, Rainer Maria

MATCHES IN DOCUMENT

*Thomas Mann's »**Buddenbroo*

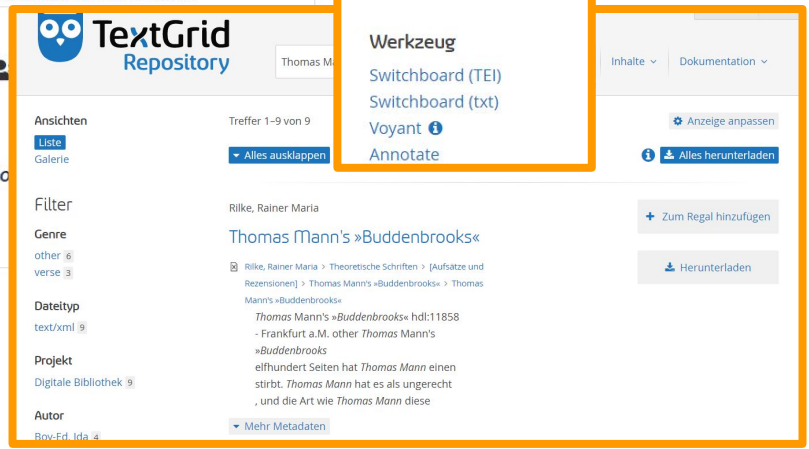
Herunterladen

Objekt (TEI)
Metadaten (XML)
Tech. Metadaten (XML)
Plain Text (txt)
E-Book (epub)
HTML
ZIP

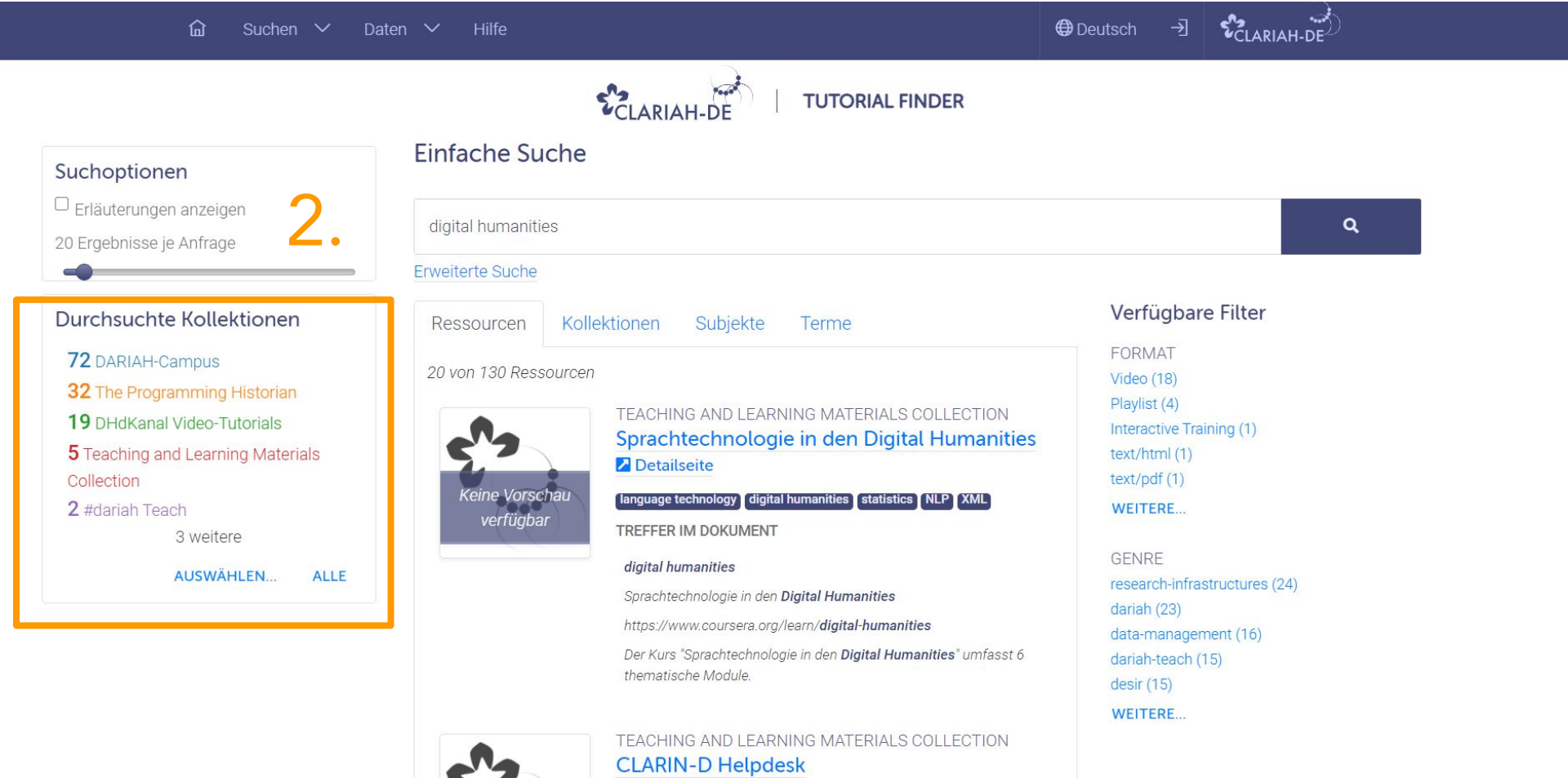
Werkzeug

Switchboard (TEI)
Switchboard (txt)
Voyant
Annotate

3.



1. https://teaching.clariah.de/search/search/simple/



Content Search | Aggregator | Help

1. https://contentsearch.clarin.eu/

2.

Text layer CQL query: buddenbrooks

Search for: Any Language | Text layer Contextual Query Language (CQL) | in: 4421 selected collections | and show up to: 10 hits per endpoint

12 matching collections found in 44 searched collections

3.

The Språkbanken corpora — Unknown Institution, FCS v2.0

@ AmandaLBjorkman + ett extra stilpoäng för **buddenbrooks** i bokhyllan

The Språkbanken modern corpora — Unknown Institution, FCS v2.0

@ AmandaLBjorkman + ett extra stilpoäng för **buddenbrooks** i bokhyllan

The SUC corpus — Unknown Institution, FCS v2.0

DWDS Core Corpus — Berlin-Brandenburg Academy of Sciences and Humanities

Thomas **Manns** **Buddenbrooks** erschienen 1901 .

Noch etwa in **Thomas** **Manns** **Buddenbrooks** besteht das Gerüst des Romans aus der chronologis
Buddenbrook betreffen , und wenn Flaubert , in vieler Hinsicht ein Vorläufer , lange und grundsätzlic
, die die Handlung kaum vorwärtstreiben , so bleibt doch in Madame Bovary (aber wie wäre es mit E
weitersickerndes Sichannähern zuerst an Teilkrisen , schließlich an die abschließende Katastrophe

Tagesspiegel — Berlin-Brandenburg Academy of Sciences and Humanities

Seit 1928 im selben Haus Nun ist diese Erfurt-Dynastie – vier Generationen von Steinhäusern habe
Thomas **Manns** **Buddenbrooks** .

View

Virtual Language Observatory Search Contributors Help CLARIN

VLO / Faceted search / Search results / Record: Thomas Mann's »Buddenbrooks«

1. https://vlo.clarin.eu/

Thomas Mann's »Buddenbrooks« 2.

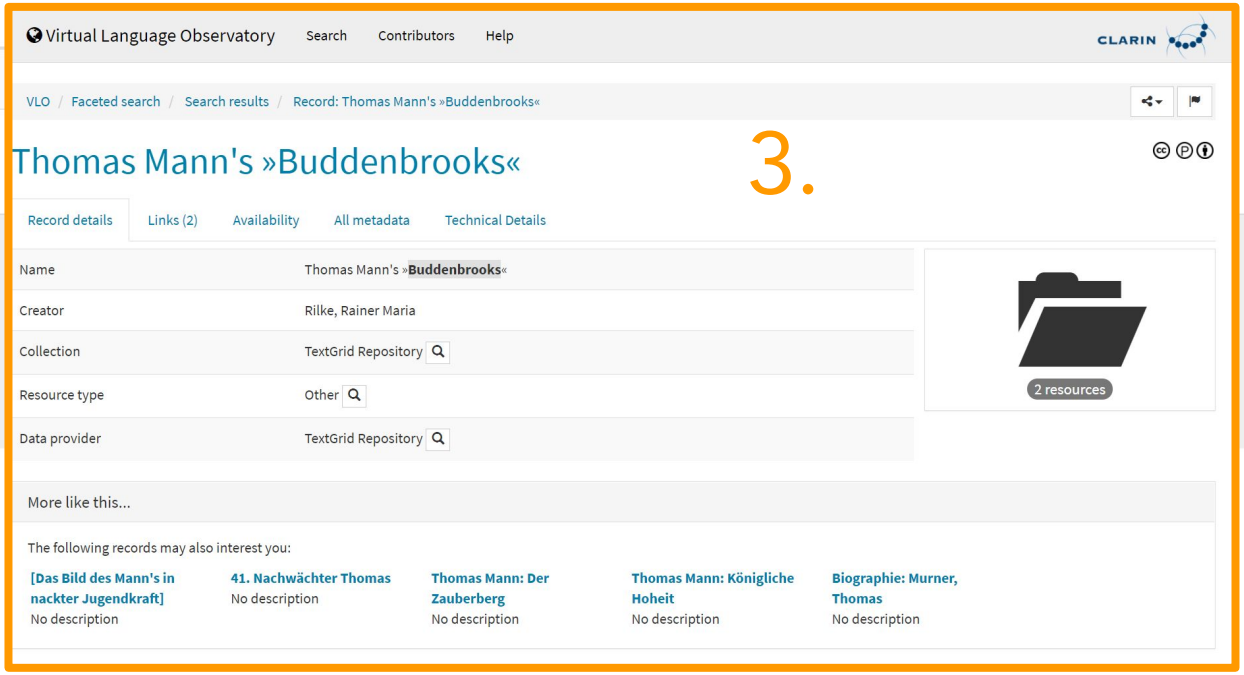
Record details | Links (2) | Availability | All metadata | Technical Details

Name

textgrid:tbz6.0

HDL hdl:11858/00-1734-0000-0004-916C-5

About
v4.9.2



CLARIAH-DE Cross-Service Search

CLARIAH-DE Cross-Service Search provides an integrated access to three different search engine systems developed within the CLARIN and DARIAH projects:

- Scientific collections and resources registered in the DARIAH-DE Collection Registry can be found via the **Generic Search (GS)** by searching in corpus data and metadata.
- Corpus data available through CLARIN-FCS endpoints can be accessed by using search queries in the **Federated Content Search (FCS)**.
- Resources, services and tools available within CLARIN can be found via the **Virtual Language Observatory (VLO)** by searching in metadata.

CLARIAH-DE Cross-Service Search provides a quick and broad overview of the freely available tools and collections that can be used in own research contexts. The search form below enables searching simple terms on the three systems simultaneously. The search results of the systems can be viewed by switching between the corresponding tabs. In FCS, the query is conveyed to the system as it can be seen in the search field, but the search results are not immediately shown. Eventually the search button within the FCS tab must be clicked again. An advanced search with specific criteria such as metadata fields (in GS and VLO, e.g. language, data format) and annotation layers (in FCS, e.g. PoS, lemma) can be conducted on the individual websites of the systems.

This functional search limitation is deliberate because there is a trade off between integrating the available resources more extensively and the representation of the search results. The Cross-Service Search is a functional demonstrator which shows the merging of different search spaces, namely GS, FCS and VLO.

e.g. Annotation

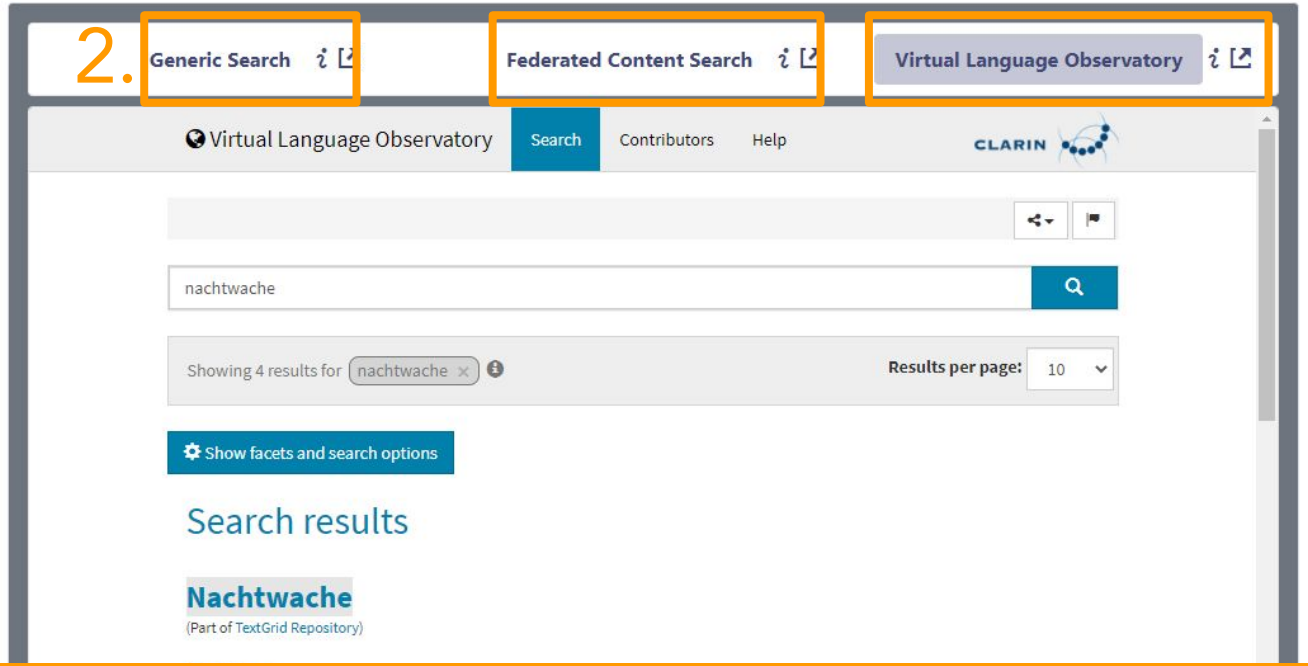
https://www.clariah.de/en/re-use-data/find-data/clariah-de-cross-service-search

CLARIAH-DE

Re-using Data | Editing & Annotating | Analysis & Processing | Publishing & Archiving | Consulting & Training | About us

This functional search limitation is deliberate because there is a trade off between integrating the available resources more extensively and the representation of the search results. The Cross-Service Search is a functional demonstrator which shows the merging of different search spaces, namely GS, FCS and VLO.

1. nachtwache



https://www.clariah.de/en/re-use-data/find-data/clariah-de-cross-service-search

Cross-Services Search, but...

- 2.2 mio resources in all!
- but…(here are the main differences)
  - differing user requirements, e.g. expressed by the resource types/data models:
    - VLO: CMDI-based metadata
    - FCS: linguistic annotated text
    - GS: no preselected data model or resource type
  - legacy: more or less deep integration in European contexts
  - not a technical problem: (FCS-QL leaning on corpus query processor/language CQP and GS & VLO leaning on Apache Lucene-based solutions, i.e. Elasticsearch, Lucene Query Parser)

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Experiences & Lessons Learned

clariah.de

@CLARIAHde

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
FÖRDERKENNZEICHEN
01UG1910 A bis I

Experiences & Lessons learned: Data integration

CLARIAH-DE

1. User perspective vs. (technical) harmonisation
2. Varying expectations towards result presentation:
   a. VLO: “take this data for your research question and use an analysis tool of your choice”
   b. FCS: “this word or construction appears in the following resources”
   c. GS: “this word is appearing in the following collection descriptions”
3. Flat integration/simple rebranding (exchangeable stylesheets)

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Wrap up: Looking at CLARIAH-DE as merger

- various categories of challenges: cultural, technological, structural, resources
- self-concept: a data centre comes with other resources, knowledge, motives as a research institution; a research infrastructure is something in between: advantage?
- trade offs have to be considered
- level of infrastructure components predetermines the outcomes
    - e.g. AAI, Helpdesk have worked out really well, application integration is case-dependent, data integration is difficult
    - differentiation between maintainer- or provider-related outcomes and user- or research-related outcomes

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Thank you for your attention!

Sources:

- Slides: 10.5281/zenodo.4628889
- *Eckart, Thomas et al.* (to be published): CLARIAH-DE Cross-Service Search: Prospects and Benefits of Merging Subject-specific Services". *DARIAH-DE Working Papers. Göttingen.*
- Result poster AP4 Technical Integration: https://zenodo.org/record/4572610
- Some information available at: https://www.clariah.de/

Stefan Buddenbohm, RDD, Göttingen State and University Library, sbudden@gwdg.de

Thomas Eckart, NLP Group, Leipzig University, teckart@informatik.uni-leipzig.de

DANS Seminar, 2021-03-23

clariah.de

@CLARIAHde

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
FÖRDERKENNZEICHEN 01UG1910 A bis I