CLARIAH-DE

Aligning two Research Infrastructures: Experiences and Challenges

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

Scholarly Primitives - DARIAH Annual Event 2020, Zagreb, Croatia, November 10-13, 2020 (Session Infrastructural Challenges)

clariah.de @CLARIAHde

Structure (10 min talk)

1. What is CLARIAH-DE?
2. Level of Challenges - Unified search as example
3. Experiences & Lessons learned

Sources:

- *Eckart, Thomas et al.* (to be published): Zusammenführung fachspezifischer Suchen in CLARIAH-DE: Herausforderungen technischer Integration". *DARIAH-DE Working Papers. Göttingen.*
- Information available at https://www.clariah.de/
- Slides@ https://zenodo.org/communities/dariahannualevent2020

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Funded by the BMBF 2019-2021 - 6 WPs - 13 funded partners - Merger project CLARIAH-DE

Building blocks for the NFDI www.text-plus.org

Language- and text-based research data

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

Level of Challenges

- Cultural: Specific organisational structures are reflected in decision-making and working structures. A tendency of perseverance of the status quo and passed on practices may be a problem.
- Structural: Although CLARIAH-DE is about merging, both CLARIN-D and DARIAH-DE have to maintain roles on a European level: CLARIN ERIC and DARIAH ERIC. In absence of a “new, mutual, unified identity” old structures persist and consume resources.
- Technological: Different research communities lead to different solutions (such as the handling of metadata, way of 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 one.

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

CLARIAH-DE

CLARIAH-DE
Unified Search

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

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*

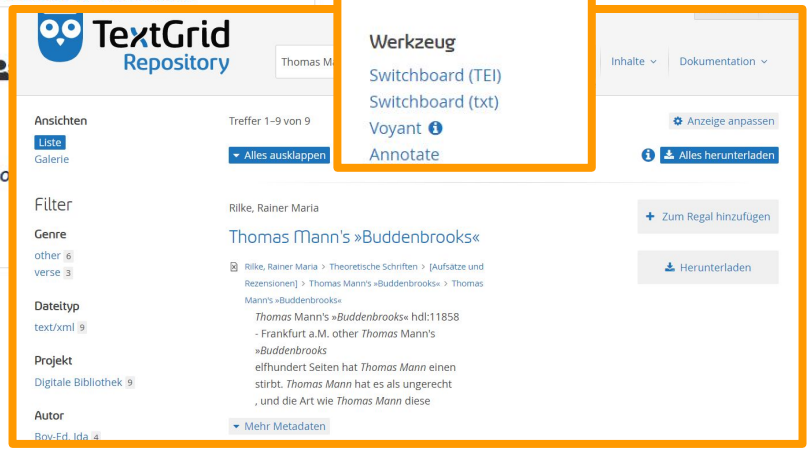
Herunterladen

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

Werkzeug

Switchboard (TEI)
Switchboard (txt)
Voyant
Annotate

3.



Federated Content Search | Aggregator | Help

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

2. Text layer CQL query: Thomas Mann Buddenbrooks

Search for: Any Language | Text layer Contextual Query Language (CQL) | in: 3997 selected collections | and show up to: 10 hits per endpoint

7 matching collections found in 90 searched collections

Display as Key Word In Context

The Språkbanken corpora – Unknown Institution, FCS v2.0

Exception: Read timed out

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

Exception: Read timed out

The SUC corpus – Unknown Institution, FCS v2.0

Exception: Read timed out

3.

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 .

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.

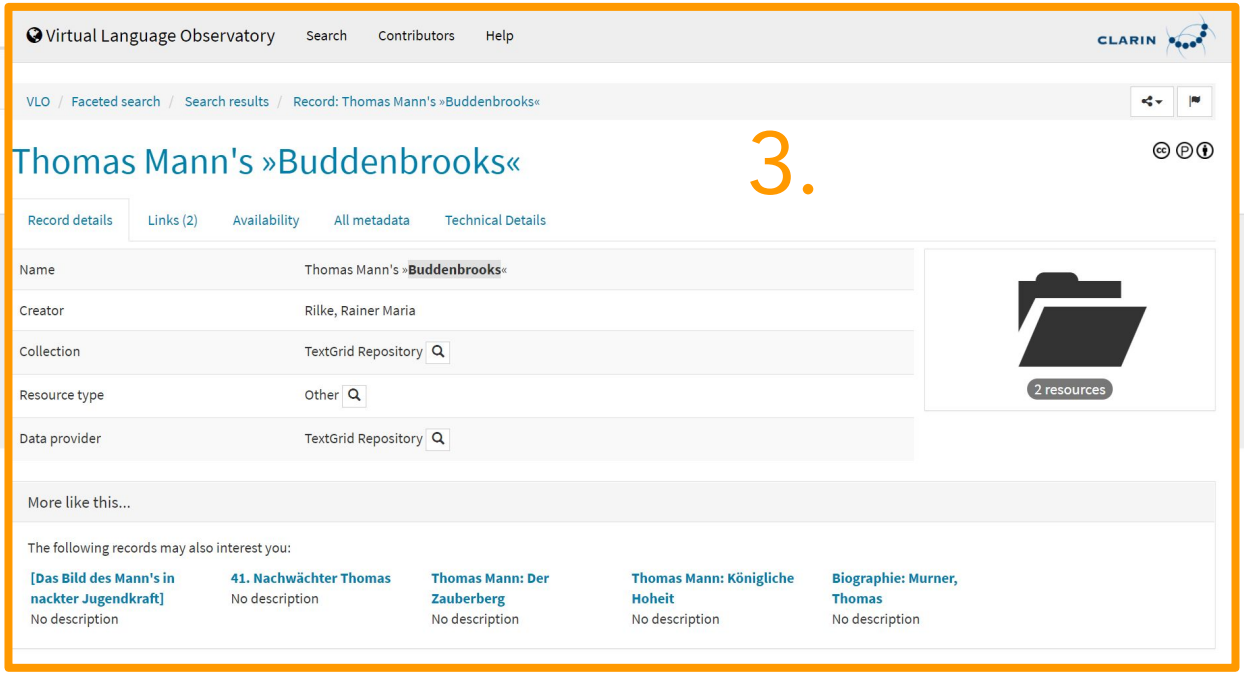
Record details | Links (2) | Availability | All metadata | Technical Details

Name

textgrid:tbz6.0

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

About
v4.9.2



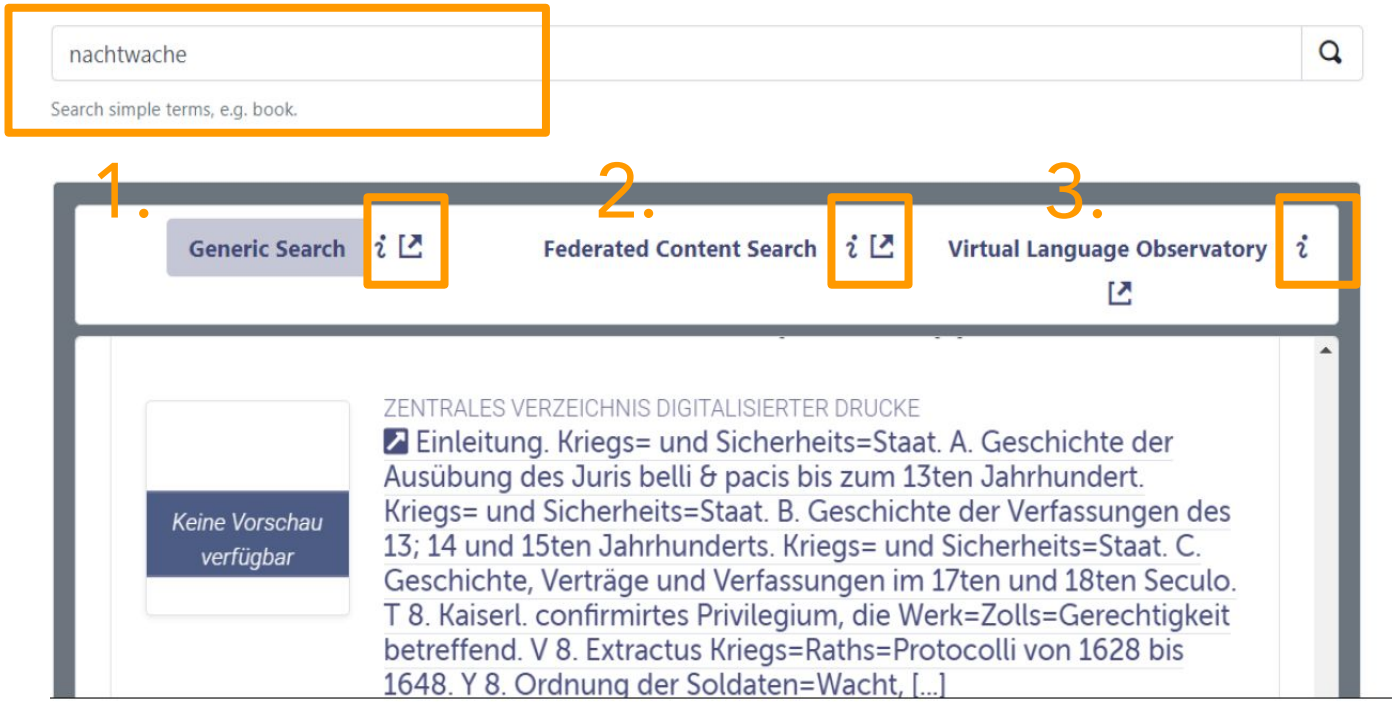
Unified search	CLARIN-D	DARIAH-DE
Stock	VLO: 1.200.947 mio. records; FCS: 34 endpoints at 16 institutions querying 4000 collections	DARIAH: 47 collections with well over 1.2 mio. resources; MWW: 27 collections with well over 250k resources; CLARIAH tutorial finder: 6 collections with 554 resources
Organisation	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 facetting search results; Customization, e.g. MWW

Unified 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 CQP and GS & VLO leaning on Apache Lucene-based solutions, i.e. Elasticsearch, Lucene Query Parser)

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

CLARIAH-DE Search Prototype



Experiences & Lessons learned

1. Varying use cases/search spaces, which will be a problem as soon as leaving the top level search slot (e.g. varying expectations how to sort and present results).
    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”
2. Simple rebranding (= exchangeable stylesheets) may allow for a quite simple solution to integrate services within other portals or environments (flat integration such as the prototypical iFrame implementation of CLARIAH-DE demonstrates).
3. Always consider trade off of user perspective vs. (technical) harmonisation
4. (ugly) iFrame “solution” with two main assets: delusion of (or better: access to) a larger stock of data AND organisational/technical invitation to others (EOSC, NFDI)

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Wrap up: Looking at CLARIAH-DE as merger

- various categories of challenges: cultural, technological, structural, resources
- expectations have to align with a realistic time horizon
- trade offs have to be considered
  - e.g. unified search: user perspective vs. harmonisation
- character of infrastructure components predetermines the outcomes
  - e.g. AAI, Helpdesk have worked out really well
  - differentiation between maintainer- or provider-related outcomes and user- or research-related outcomes

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

CLARIAH-DE

Thank you for your attention!

Sources:

- *Eckart, Thomas et al.* (to be published): Zusammenführung fachspezifischer Suchen in CLARIAH-DE: Herausforderungen technischer Integration". *DARIAH-DE Working Papers.* Göttingen
- Information available at https://www.clariah.de/

Slides:

- https://zenodo.org/communities/dariahannualevent2020

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

Scholarly Primitives - DARIAH Annual Event 2020, Zagreb, Croatia, November 10-13, 2020 (Session Infrastructural Challenges)

clariah.de

@CLARIAHde

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