Aligning two Research Infrastructures: Experiences and Challenges

Stefan Buddenbohm (Göttingen State and University Library)

clariah.de @CLARIAHde

Structure (15 min talk)

1. What is CLARIAH-DE? - A few key facts
2. What lies behind? - CLARIN-D & DARIAH-DE
3. Which challenges? - Unified search as example
4. Lessons learned

Sources:

- ➔ Thomas Eckart et. al. (2020): Zusammenführung fachspezifischer Suchen in CLARIAH-DE: Herausforderungen technischer Integration". DARIAH-DE Working Papers. Göttingen: DARIAH-DE Working Papers.
- ➔ information available at https://www.clariah.de/

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 upcoming NFDI (Text+)

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
Unified Search

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Searches Data Help English

DARIAH-DE

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

Search options

Show explanations

20 Results per page

Queried collections

764 TextGrid Digitale Bibliothek: Literatur

9 Bayerisches Digitales Repositorium

3 Deutsches Textarchiv

Göttinger Digitalisierungszentrum – Bucherhaltung

Göttinger Digitalisierungszentrum – DigiWunschbuch

49 more

Simple search

2. Clemens Brentano Des Knaben Wunderhorn

Extended search

Resources | Collections | Subjects

Results in 3 collections

TextGrid Digitale Bibliothek: Literatur

Terms | Subjects

TextGrid Repository

Suchbegriff

Erweiterte Suche

Metadaten

Dateityp
text/xml

PID
hdl:11858/00-1734-0000-0002-114F-0
Zitationsvorschlag

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

Werkzeug
Switchboard (TEI)
Switchboard (txt)
Voyant
Annotate

Inhaltsverzeichnis
Lied, mit welchem die Kinder

Arnim, Ludwig Achim von › Gedichte › Des Knaben Wunderhorn › Anhang: Kinderlieder › Lied, locken › Lied, mit welchem die Kinder die Schnecken locken

Lied, mit welchem die Kinder die Schnecken locken

3.

Klosterfrau im Schneckenhäußle,
Sie meint, sie sey verborgen?
Kommt der Pater Guardian,
Wünscht ihr guten Morgen!

Der annotierte Datenbestand der Digitalen Bibliothek inklusive Metadaten sowie dav
eine Abwandlung des Datenbestandes von www.editura.de durch TextGrid und werd
Commons Namensnennung 3.0 Deutschland Lizenz (by-Nennung TextGrid, www.edit
bezieht sich nicht auf die der Annotation zu Grunde liegenden allgemeinfreien Texte
Lizenzbestimmungen).

Lizenzvertrag

Eine vereinfachte Zusammenfassung des rechtsverbindlichen Lizenzvertrages in allge

Hinweise zur Lizenz und zur Digitalen Bibliothek

Zitationsvorschlag für dieses Objekt
TextGrid Repository (2011). Arnim, Ludwig Achim von. Gedichte. Des Knaben W
Lied, mit welchem die Kinder die Schnecken locken. Lied, mit welchem die Kind
Bibliothek. TextGrid. https://hdl.handle.net/11858/00-1734-0000-0002-114F-0

Federated Content Search | Aggregator | Help

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

2.

Text layer CQL query | Bertolt Brecht

Search for | Any Language | Text layer Contextual Query Language (CQL) | in | 4509 selected collections | and show up to | 10 | hits per endpoint

7 matching collections found in 94 searched collections

Corpus C4 – Berlin-Brandenburg Academy of Sciences and Humanities

Nun sahen wir die Aufführung der Komödie unter dem Titel " Der Hofmeister " dur
Caspar Neher .

Der Gothaer Parteitag hatte die Naturalismus-Debatte zum Zentrum , die ihrerseit
im Disput über den Expressionismus zwischen **Bertolt** **Brecht** , Georg Lukács un

Es war Helmut Schmidt vorbehalten , in seinem Nachwort immerhin an Gestalten wie **Bertolt** **Brecht** , Kurt Masur und Bernhard Heisig zu erinnern .

Kein zweites Stück von **Bertolt** **Brecht** scheint durch die Implosion des " real existierenden Sozialismus " stärker widerlegt , als das im Januar 1932 erstmals

3.

Corpus C4 – Homepage

Berlin-Brandenburg Academy of Sciences and Humanities
Corpus C4
German

Display as Key Word In Context | Download | Use Weblicht

As CSV file
As ODS file
As Excel file
As TCF file
As Plain Text file

Nun sahen wir die Aufführung der Komödie unter dem Titel " Der Hofmeister " durch das Berliner Ensemble im Deutschen Theater , inszeniert von	**Bertolt**	Brecht und Caspar Neher .
Der Gothaer Parteitag hatte die Naturalismus-Debatte zum Zentrum , die ihrerseits auf den Sickingen-Streit zwischen Marx und Lassalle zurückging und ihre Fortsetzung im Disput über den Expressionismus zwischen	**Bertolt**	Brecht , Georg Lukács und Ernst Blo … n sich ja dürfen .
Es war Helmut Schmidt vorbehalten , in seinem Nachwort immerhin an Gestalten wie	**Bertolt**	Brecht , Kurt Masur und Bernhard Heisig zu erinnern .
Kein zweites Stück von	**Bertolt**	Brecht scheint durch die Implosion des " real existierenden Sozialismus " stärker widerlegt , als das im Januar 1932 erstmals aufgeführte Lehrstück " Die Mutter " . Brecht adaptierte den gleichnamigen Roman von Maxim Gorki aus dem Jahr 1907 , in dem Gorki schildert , wie die Arbeiterwitwe Pelagea Nilowna aus Twer durch das Miterleben der Streikkämpfe von Nishni Nowgorod und Sormowo selbst zur Klassenkämpferin wird .

Virtual Language Observatory | Search | Contributors | Help | CLARIN

VLO / Faceted search / Search results

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

clemens brentano: des knaben wunderhorn

2.

Showing 1 to 10 of 763 results for clemens brentano: des knaben wunderhorn

Results per page: 10

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

Language

Type to filter or search for more

German (4)

Collection

Resource type

Modality

<< < 1 2 3 4

Des Knaben Wunder

(Part of TextGrid Repository)

No description

[Rezension zu:] [...] D
Lieder, herausgegebe
Brentano

(Part of Deutsches Textarchiv (1600–1900))

Historical German text source (16
Archive) collection according to the

German

Landing page for this record

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

OLAC-DcmiTerms

creator	Arnim, Ludwig Achim von
date	2011-12-26T20:14:30Z
format	text/tg.edition+tg.aggregation+xml
identifier	https://textgridrep.org/textgrid:k6c9.0
identifier	hdl:11858/00-1734-0000-0002-0D46-0
relation	TGPR-372fe6dc-57f2-6cd4-01b5-2c4bbefcfd3c
relation	**Des Knaben Wunderhorn**
relation	textgrid:k6cb.0
rights	http://creativecommons.org/licenses/by/3.0/de/legalc
source	Arnim, Ludwig Achim von
source	Achim von Arnim und **Clemens** Brentano: **Des Knaben Wunderhorn**. 1979.
source	Stuttgart

Unified search	CLARIN-D	DARIAH-DE
Stock	VLO: 1.200.947 mio. records; FCS: 34 endpoints at 16 institutions querying 100 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, yes but...

CLARIAH-DE

- 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)
- leading to a prototypical implementation leaning on iFrames & some lessons learned

GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung

Conclusions

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

Summary: 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

Thank you for your attention!

Sources:

Thomas Eckart et. al. (2020): Zusammenführung fachspezifischer Suchen in CLARIAH-DE: Herausforderungen technischer Integration". DARIAH-DE Working Papers. Göttingen: DARIAH-DE Working Papers.

Information available at https://www.clariah.de/

Stefan Buddenbohm, Göttingen State and University Library

sbudden@gwdg.de

clariah.de

@CLARIAHde

GEFÖRDERT VOM

Bundesministerium für Bildung und Forschung

FÖRDERKENNZEICHEN 01UG1910 A bis I