

Aligning two Research Infrastructures: Experiences and Challenges

Stefan Buddenbohm (Göttingen State and University Library)



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/



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





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

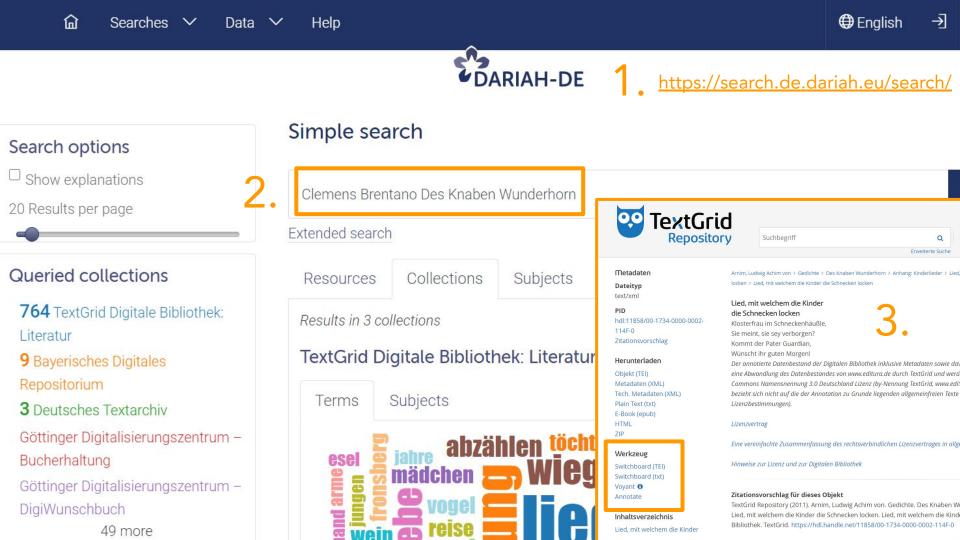
Level of Challenges

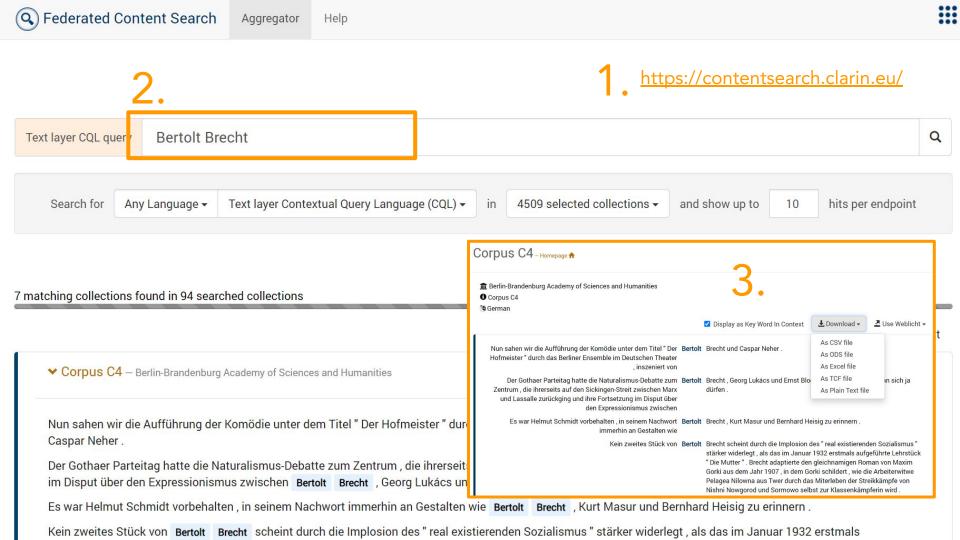


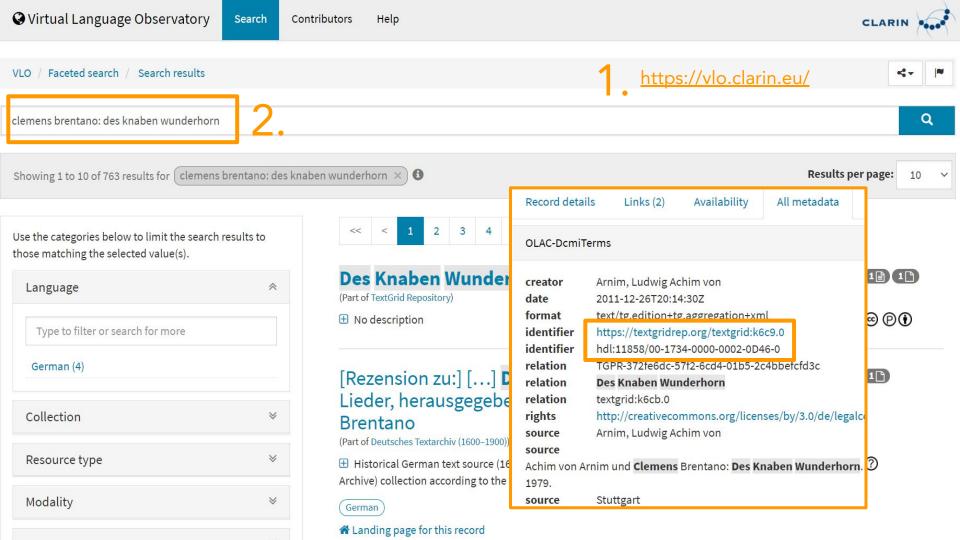
- Cultural: Specific organisational structures are reflected in decision-making and working structures. A tendency of perseverance of the status quo and passed on practices may be a problem.
- Structural: Although CLARIAH-DE is about merging, both CLARIN-D and DARIAH-DE have to maintain roles on a European 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.



CLARIAH-DE Unified Search







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

Elasticsearch; Elasticsearch Query Language; DCsimple for

facetting search results; Customization, e.g. MWW

Search

Solr; 14 search facets based on

https://github.com/clarin-eric/VLO-mapping; FCS Query Language

Unified search, yes 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)
- leading to a prototypical implementation leaning on iFrames & some lessons learned



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"
 - 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)



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





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





