Generische und fachspezifische Integration in der Forschungsinfrastruktur CLARIAH-DE

Buddenbohm, Stefan
sbudden[at]gwdg.de
SUB Göttingen, Deutschland
ORCID-iD: 0000-0002-3469-6101

Eckart, Thomas
teckart[at]informatik.uni-leipzig.de
Universität Leipzig, Deutschland
ORCID-iD: 0000-0001-6692-0713

Gradl, Tobias
tobias.gradl[at]uni-bamberg.de
Universität Bamberg, Deutschland
ORCID-iD: 0000-0002-1392-2464

Helfer, Felix
helfer[at]informatik.uni-leipzig.de
Universität Leipzig, Deutschland
ORCID-iD: 0000-0002-5739-2047

Jegan, Robin
robin.jegan[at]uni-bamberg.de
Universität Bamberg Deutschland
ORCID-iD: 0000-0002-0388-7220

Zusammenfassung. CLARIAH-DE führt die beiden etablierten Forschungsinfrastrukturen CLARIN-D und DARIAH-DE im Rahmen eines BMBF-geförderten Projekts zu einem gemeinsamen Angebot für geistes- und sozialwissenschaftlich Forschende zusammen. Mit seiner nachhaltigen Bereitstellung von Forschungsdaten, technischer Infrastruktur, digitalen Werkzeugen, virtuellen Forschungsumgebungen sowie Informations- und Schulungsmaterialien trägt CLARIAH-DE auch zur Nationalen Forschungsdateninfrastruktur (NFDI) bei. Die Integrationserfahrungen des Projektes liegen insbesondere auf organisatorischer und technischer Ebene, wobei sich dieser Vortrag auf die technische Integration und die Konsequenzen bezieht, die sich aus unterschiedlichen Sichtweisen auf Forschungsdaten und deren Nutzung in verteilten Umgebungen für integrative Maßnahmen ergeben. Problem- und

Erfolgsfälle werden anhand konkreter Beispiele illustriert. Als eine zentrale Schlussfolgerung ergibt sich: je fachspezifischer Anforderungen, und damit Komponenten, sind, desto komplexer eine Zusammenführung. Dabei sind zukünftige Kontexte ebenfalls zu berücksichtigen, bspw. die NFDI oder EOSC, und gefällte Entscheidungen gegenüber der Community plausibel zu begründen.

1 Einleitung

Die Zusammenführung zweier etablierter Forschungsinfrastrukturen zum gemeinsamen Angebot CLARIAH-DE verbindet sich mit Herausforderungen auf (mindestens) drei verschiedenen Ebenen:

- Strukturelle Aspekte, bspw. bestehende organisatorische Strukturen oder Verpflichtungen Dritten gegenüber, die den Gestaltungsspielraum begrenzen und die globale vs. lokale Perspektiven von Forschungsinfrastrukturen (und Forschungsdatenmanagement) zu berücksichtigen haben. Im Fall von CLARIAH-DE sind dies bspw. die ERIC- oder NFDI-Ebenen.
- Infrastrukturelle Aspekte, bspw. inkompatible technische Komponenten oder unterschiedliche Verfahrensweisen und Methoden beim Umgang mit Forschungsdaten, wobei generische und fachspezifische Ausprägungen unterschiedlich komplex umzusetzen sind.
- Kulturelle Aspekte, bspw. die Art und Weise wie miteinander kommuniziert wird und wie Entscheidungen getroffen werden.

Im Folgenden stehen die infrastrukturellen Aspekte im Vordergrund. Die durch CLARIAH-DE adressierten Infrastrukturbereiche lassen sich in drei Kategorien mit jeweils spezifischen Charakteristika (siehe Abbildung 1) einteilen.

	Integrationsebenen			
	Komplexität / Aufwand	Standardi-sierung	Sichtbarkeit / Nutzenerwartung	Externe Abhängigkeiten
Basisinfrastruktur-Komponenten	standardisierte Komponenten mit geringerem Aufwand zusammen-führbar (AAI, Monitoring)	häufig Einsatz standardi-sierter, marktgängiger Lösungen	gering, selbst substanzieller Nutzen für die Forschenden wird nicht unbedingt mit der Forschungs-infrastruktur verbunden	interne Abhängigkeiten gering; Abhängigkeit von externen Dienstleistern (“marktgängige” Standardl-ösungen) stärker
Anwendungen	hoch; je “forschungs-näher”, desto komplexer, da forschungs-spezifische Anforderungen	Standardi-sierung bzgl. Forschungs-daten möglich, Werkzeug-ebene jedoch häufig mit fach-spezifischen Anpassungen	hoch; Nutzende erhoffen sich eine Verbesserung durch neue Werkzeuge	fallbezogen; Werkzeuge der einen Fachdomäne auch in der anderen Fachdomäne einsetzbar oder anpassbar
(Forschungs-) Daten	hoch, sofern die Daten abhängig von spezifischen Forschungsfragen erhoben wurden	Standardi-sierung häufig in einem Zielkonflikt mit Granularität	hohe Erwartungshaltung insb. in SSH Forschungs-kontexten	i.d.R. abhängig von der Nutzung angepasster, fachspezifischer Werkzeuge oder Standards

Abb. 1. Infrastrukturelle Integrationsebenen der Forschungsinfrastruktur CLARIAH-DE.

Ausgehend von den spezifischen Charakteristika können Entscheidungen zur Zusammenführung, Anpassung oder zum Beibehalt ausgewählter Komponenten getroffen werden, die ein angemessenes Verhältnis von Aufwand und Nutzen aufweisen. Der Nutzen kann sich auf der Seite der Infrastrukturbetreibenden wie auch auf Seiten der Forschenden materialisieren. Die Aufwand-Nutzen-Abwägung muss in die Entscheidungsfindung einfließen, wobei für eine Forschungsinfrastruktur die Nutzenden – in unserem Fall Forschende aus den Geistes-, Kultur- und Sozialwissenschaften – im Vordergrund stehen. Bei der operativen Umsetzung wurde in der Regel auf den gleichen Standardprozess zurückgegriffen: (1) auf Basis einer konkreten Zielspezifikation wurde durch kleine Arbeitsgruppen mit geringem organisatorischem Overhead Lösungen entwickelt und (2) die

konkreten Implementationsergebnisse in regelmäßigen Treffen mit dem Projekt abgestimmt bzw. diskutiert.

Als eine zentrale Schlussfolgerung dieser Arbeiten ergibt sich: je fachspezifischer Anforderungen, und damit Komponenten, sind, desto komplexer eine Zusammenführung. Dabei sind zukünftige Kontexte ebenfalls zu berücksichtigen, bspw. die NFDI oder EOSC, und gefällte Entscheidungen gegenüber der Community plausibel zu begründen.

Ausgehend von diesen drei Integrationsebenen werden im Folgenden konkrete Beispiele vorgestellt.

2 Basisinfrastruktur

In CLARIAH-DE wurden unter anderem das Monitoring, die Authentifizierungs- und Authorisierungsinfrastrukturen (AAI) sowie der Helpdesk integriert. Für Basisdienste ist die Verwendung einer überschaubaren Anzahl von Standardwerkzeugen üblich. Abgesehen von einer besseren Usability für die Nutzenden (Einheitlichkeit), verringert dies den Aufwand technischer Integration und verschiebt Integrationsarbeit in das Konfigurationsmanagement. Die Querbezüge zu kulturellen und strukturellen Integrationsaspekten lassen sich allerdings auch hier nicht ignorieren. Deutlich wird dies zum Beispiel im Bereich Monitoring, wenn unterschiedliche Herangehensweise an das Thema Diensteverfügbarkeit in einer einheitlichen Plattform berücksichtigt werden müssen[1]. Im Gegensatz dazu sind geringere Aufwände zu erwarten, wenn Dienste hinsichtlich ihrer Struktur und Nutzung weitgehend identisch bzw. komplementär aufgebaut/konfiguriert sind[2]. Querbezüge, die direkt in den Forschungs- und Arbeitsbereich der Nutzenden wirken, bspw. mit Blick auf das Forschungsdatenmanagement, sind bei den in CLARIAH-DE integrierten Basisinfrastrukturkomponenten weniger relevant.

3 Anwendungen

Integrationsbemühungen auf Anwendungsebene erfordern einen differenzierten Blick auf die Komponenten und deren Einsatzradius: Während die Erwartungshaltung seitens der Nutzenden höher ist als bei

[1] Buddenheim, 2021.
[2] Sambale et al. 2019.

der Basisinfrastruktur, setzt eine erfolgreiche Integration ein gewisses Maß an Kompatibilität beziehungsweise Anpassbarkeit der jeweiligen Werkzeuge voraus, um diese auch für alternative Domänen oder Anwendungsfälle zu ertüchtigen. Ein erfolgreiches Beispiel in CLARIAH-DE ist u.a. die Integration des Language Resource Switchboards von CLARIN in das TextGrid Repository[3] (siehe Abbildung 2)[4], die im Hintergrund die Verknüpfung von anderen Basisinfrastrukturkomponenten – hier AAI – voraussetzt.

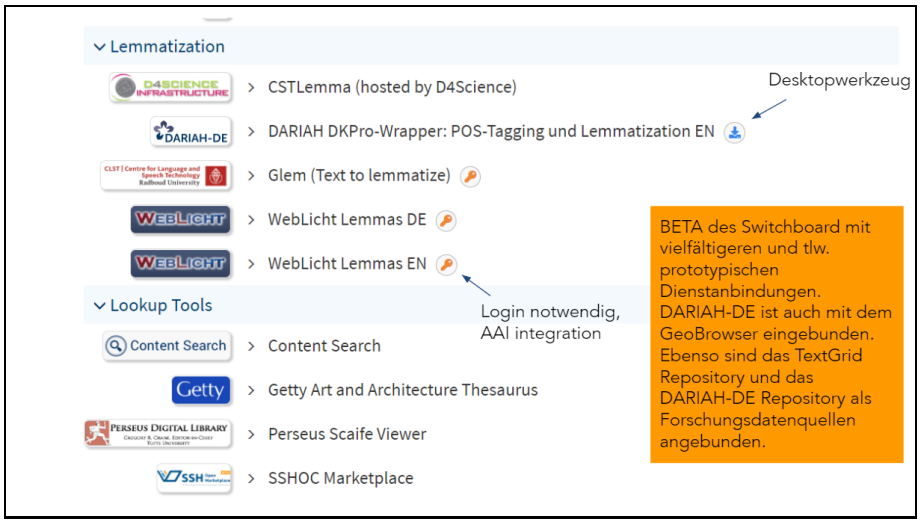

Abb. 2. Umsetzung auf Anwendungsebene: Verknüpfung von Werkzeugen und Diensten im Switchboard (Beta, https://beta-switchboard.clarin.eu/tools).

Auch für die Umsetzung des CLARIAH-DE Tutorial Finders (siehe Abbildung 3) finden aus beiden Initiativen hervorgegangene Komponenten eine integrative Anwendung. Mit Blick auf Medientypen, Zugriffsschnittstellen und konkrete Inhalte stark heterogene Lehr- und Lernmaterialien werden flexibel auf Basis der DARIAH-DE Datenföderationsarchitektur registriert, modelliert und integriert[5]. Sprachanalytische Funktionalität bspw. zur Geo-referenzierung und Textklassifikation wird durch die Integration von CLARIN Diensten bereitgestellt.

[3] Exemplarisch hier nachgewiesen: https://textgridrep.org/browse/vqmz_0 (Switchboard-Link in der Rubrik “Werkzeuge” aufrufen.).
[4] Weimer, 2021.
[5] Gradl, Jegan 2021.

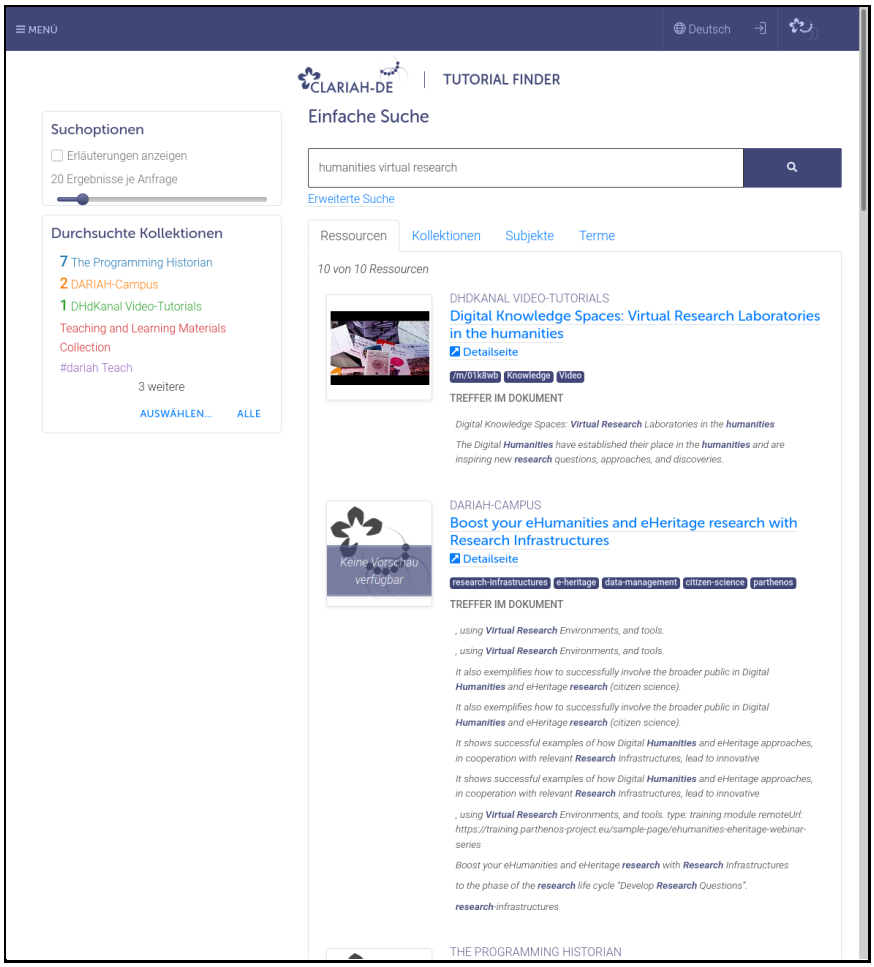

Abb. 3. Prototyp des CLARIAH-DE Tutorial Finder als integratives Anwendungsbeispiel.

4 Forschungsdaten

Der Bereich Forschungsdaten/Suchlogiken war in CLARIAH-DE mit vergleichsweise hohen Erwartungen durch die Nutzenden verbunden. Vereinfacht ist es der Wunsch nach einem zentralen Einstiegspunkt für die verschiedenen Suchräume von CLARIN und DARIAH, konkret:

- CLARIN: Virtual Language Observatory (VLO): Metadaten zur Beschreibungen von Text- und Sprachkorpora, lexikalischen Ressourcen, über 1,2 Mio. Einträge, Lucene-basiert; Federated Content Search (FCS): Schwerpunkt auf Volltexten und Korpora, 4.000 Kollektionen, FCS Query Language; Component Metadata

Initiative (CMDI) mit über 180 Metadaten-Schemata, Metadaten-Aufbereitung[6],

- DARIAH: Generic Search: 47 Kollektionen mit über 1,2 Mio. Einträgen, Kollektionen recht heterogener, vorwiegend textueller Forschungsdaten; DARIAH Collection Description Data Model (DCDDM), Elasticsearch (DCsimple für Facettierung),

ohne dabei deren Spezifika zu verlieren.

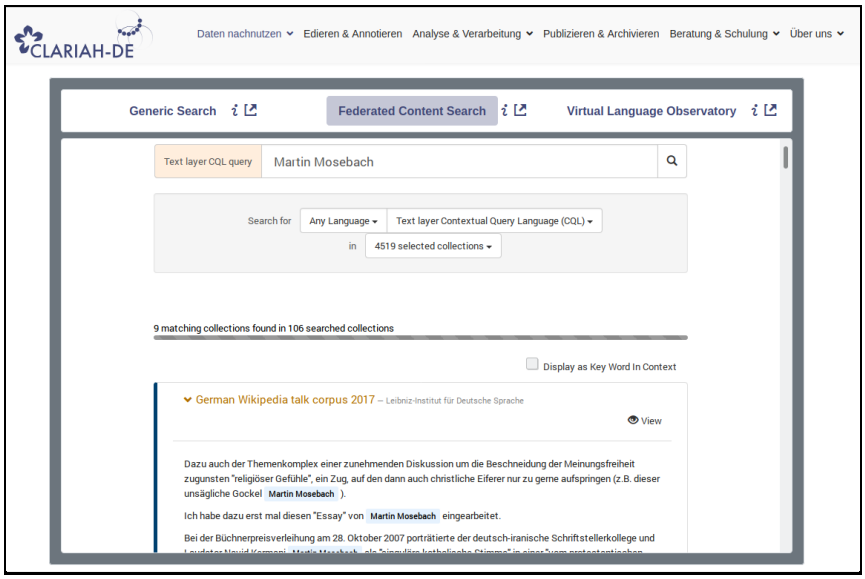

Abb. 4. Beispiel für ein eingebundenes CLARIN Zentrum, hier ein IDS-Korpus in der Federated Content Search; eingebettet in die integrierte CLARIAH-DE-Suche.

Der infrastrukturelle Rahmen dieser Suchen ist auf beiden Seiten teilweise vergleichbar aufgebaut, d.h. es gibt Komponenten für Mapping in unterschiedliche Schemata, Verfahren zur Registrierung von Kollektionen und Endpunkten sowie Anwendungen zur Verarbeitung oder Visualisierung von Forschungsdaten. Die tiefere Integration in die europäische Ebene (ERIC) bei CLARIN ist auffällig. Auf DARIAH-Seite liegt eine Herausforderung in der anderen Organisationsweise: anders als bei CLARIN sind die Forschungsdaten nicht in Zentren organisiert bzw. in der Regel sind Einzelforschende die Urheber.

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

Folgende Erfahrungen stechen für den Bereich Forschungsdaten/Suchlogiken heraus[7]:

- vergleichsweise hoher Integrationsaufwand, da unterschiedliche Zielgruppen und Informationsbedürfnisse in einer funktionalen Verkapselung der Suchdienste mit jeweils unterschiedlichen inhaltlichen Schwerpunkten resultieren,
- Maßnahmen: einfache Angleichung verwendeter Style-Informationen bis hin zu komplexen und aufwendigen Anpassungen verwendeter Datenmodelle und Anfragesprachen fließen in die Aufwand-Nutzen-Abwägung ein,
- Nutzerperspektive muss im Zentrum stehen, so dass als Lösung auch eine "Benutzerschnittstelle als Fassade vor einem potenziell hochintegrierten Datenbestand" das Mittel der Wahl sein kann,
- integrierte Zugänglichkeit aller Bestände an einer Stelle – wie in der CLARIAH-DE-Suche (Abbildung 4) prototypisch gezeigt – ist wichtig, um Anschlussfähigkeit zu anderen Akteuren (bspw. EOSC oder NFDI) zu gewährleisten; ein derartiger Zugangspunkt funktioniert dabei als technische und "organisatorische" Einladung an andere Forschungs-infrastrukturen.

Bibliografie

Buddenbohm, S., Das Monitoring von CLARIAH-DE, *DHdBlog*, abgerufen 12.04.2021, https://dhd-blog.org/?p=14881.

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. *DARIAH-DE Working Papers Nr. 41*. Göttingen: DARIAH-DE, 2021. URN: urn:nbn:de:gbv:7-dariah-2021-1-9.

Gradl, T., Jegan, R., Nachnutzung Git-basierter Sammlungen im Rahmen der Infrastrukturdienste von CLARIAH-DE. *DARIAH DE Working Papers Nr. 42*. Göttingen: DARIAH-DE, 2021. URN: urn:nbn:de:gbv:7-dariah-2021-2-5.

Sambale, H., Hedeland, H., Pirinen, T., User Support for the Digital Humanities. *Selected Papers from the CLARIN Annual*

[7] Eckart et al. 2021.

Conference 2019. Linköping Electronic Conference Proceedings 172:14, s. 119-125. https://doi.org/10.3384/ecp2020172014.

Weimer, L., Die Einbindung externer Werkzeuge in das TextGrid Repository, *DHdBlog*, abgerufen 12.04.2021, https://dhd blog.org/?p=14999.