Föderierte Inhaltssuche auf Forschungsdaten

AG Federated Content Search (FCS)

Föderiertes Informationssystem

- Zugriff auf mehrere autonome Informationsquellen, ohne dass deren Daten kopiert werden
- Zusammenschluss von einzelnen Systemen, die ihre jeweilige Selbstständigkeit bewahren
- Der Anwender eines föderierten Informationssystems kann auf die Daten so zugreifen, als habe er es mit einem einzigen System zu tun

Beispiel: Inhaltssuche vs. Metadatensuche

Suche in “Inhalten”

*„**Ich suche** in dieser Ressource **konkrete Belegstellen** in denen eine flektierte Form von „integrieren“ auf das Nomen ‚Infrastruktur‘ folgt!“*

Suche in “Metadaten”

*„**Ich suche ein Textkorpus** bestehend aus Zeitungstexten in deutscher Sprache das unter einer Creative Commons Lizenz verfügbar ist!“*

Mögliche Use-Cases

„Ich benötige einen schnellen Überblick über den Inhalt verteilt vorliegender Ressourcen!“

„Welche Nomen wurden in den letzten 5 Jahren in wissenschaftlichen Publikationen vermehrt verwendet?“

„Ich möchte Wortfeld-Angaben verschiedener Institutionen über eine API in meiner Anwendung einbinden!“

„Ich suche Bezeichnungen für Lebensmittel, die ausschließlich im süddeutschen Raum Verwendung finden!“

„Welche Flexionsvarianten dieses Wortes werden in verschiedenen Quellen wie häufig verwendet?“

Ziele und aktuelle Arbeiten in den Datendomänen

Sammlungen

Herausforderungen

- Vielzahl von Datenzentren
- Hoher Umfang und Heterogenität der Daten
  - Große und unstrukturierte Datenmengen
  - Verschiedene Datentypen
- Bestände mit urheber-, lizenz- oder datenschutzrechtlichen Einschränkungen
  - Daten, die nicht verarbeitet werden dürfen
  - Durchsuchbare Daten, deren Ergebnis nur eingeschränkt verfügbar ist

Ziele

- Performante und skalierbare Suche über große, verteilte Datenbestände
- Facettierung der Suche, z.B. anhand von Metadaten
- Transparente Suche auf zugriffsgeschützten Ressourcen

Lexikalische Ressourcen

Ziele

- Aufbau einer dezentralen Plattform für Wörterbücher
- Schnittstellen und eine gemeinsame Umgebung für den Zugriff auf verteilte lexikalische Dateneinträge
- alle beteiligten Institutionen stellen Ressourcen in dieser gemeinsamen Umgebung bereit

Aktueller Arbeitsstand

- Erweiterung der *CLARIN FCS* um Funktionalitäten für den Zugriff auf lexikalische Ressourcen
- Arbeiten an gemeinsamer Anfragesprache auf Basis der *Contextual Query Language* (CQL)
- Aufbau und inkrementelle Entwicklung von Schnittstellen bei den beteiligten Institutionen
- erster Prototyp für gemeinsame Anfragen verfügbar (aktuell über 40 lexikalische Ressourcen integriert)

Editionen

Status-Quo

Editionen sind hochgradig heterogene Ressourcen,

- in denen Sprache und Sprachstand (meist) genau erfasst werden
- deren Richtlinien der Transkription bzw. Texterfassung individuell geprägt sind
- die verschiedenste Modelle der Textauszeichnung (teilweise für gleiche Phänomene) anwenden

Fazit

- kaum effiziente Suche auf heterogener Datenbasis möglich
- Perspektive: Angleichung von Editionen auf Basis der Metadaten -> Entwicklung einer Registry (siehe Poster der AG)

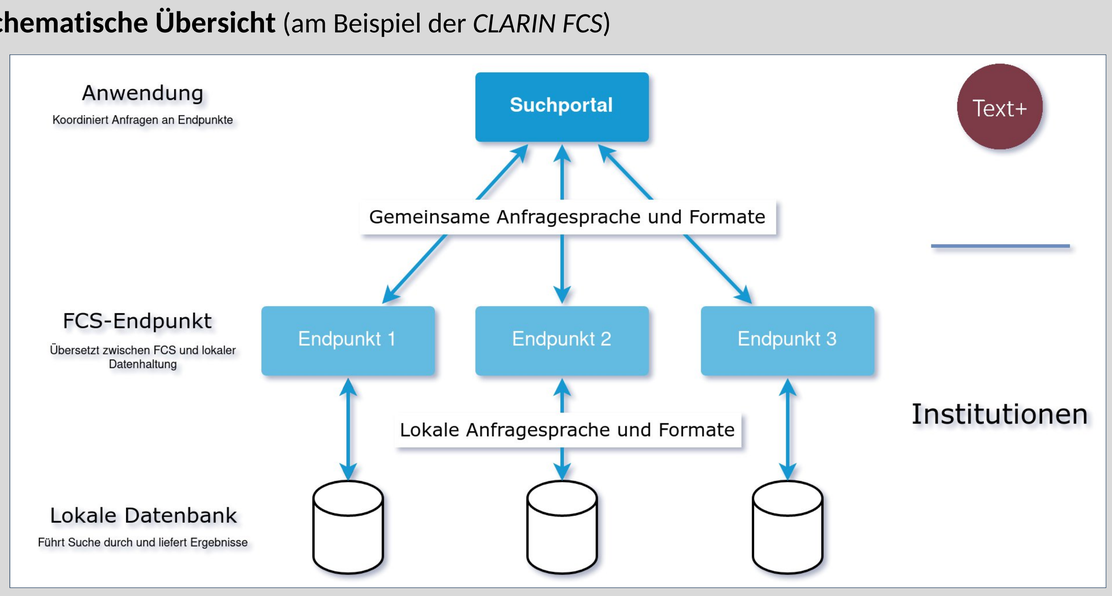
Umsetzung und Integration in die Infrastruktur

Technische Implementierung

Grundidee

- Datenbestände der beteiligten Institutionen werden über den Aufbau dezentraler *Endpunkte* aus der gemeinsamen Infrastruktur heraus abfragbar gemacht
- Jeder Endpunkt bildet die Schnittstelle zwischen der gemeinsamen Infrastruktur (Text+) und der lokalen Infrastruktur einer Institution
- Ein Endpunkt bildet gemeinsame Anfragesprachen auf die lokalen Gegebenheiten ab und liefert Antworten in standardisierten Formaten

Schematische Übersicht (am Beispiel der *CLARIN FCS*)



Weitere Themen

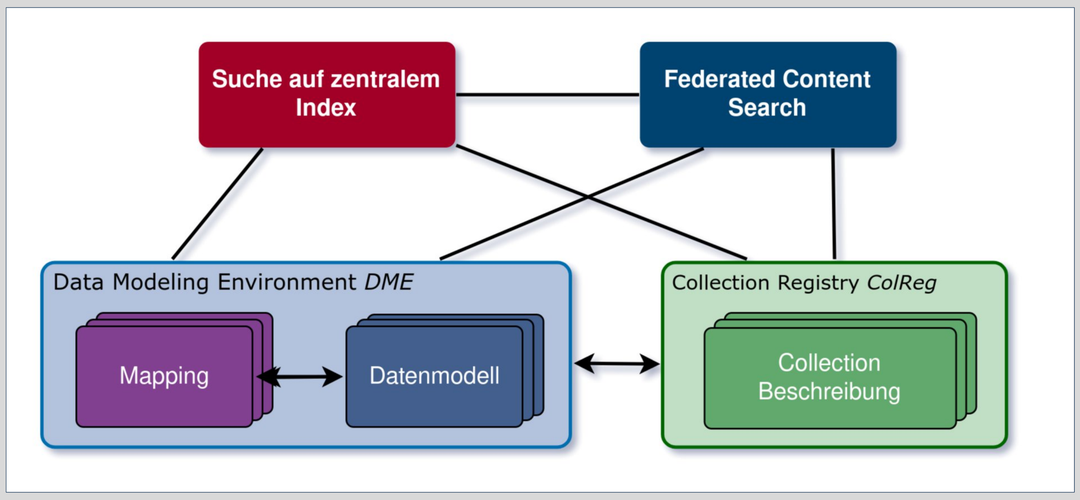
- Gemeinsame Anfragesprache
- Gemeinsame Datenformate
- Standardisierung
- Zentrales Suchportal
- Absicherung von Zugriffen

Einbindung in die Text+ Suchinfrastruktur

Grundidee

- Einbettung in bzw. Verzahnung mit bestehenden Infrastrukturkomponenten
- nahtloser Übergang der Suche von Metadaten ➔ Inhalten
- Facettierung von Suche
  - d.h. Fokussierung der Inhaltssuche durch Beschränkung (der Suche) auf Ressourcen deren Metadaten bestimmten Kriterien entsprechen
  - Beispiele: Filterung nach Sprache, Zeit(epoche), Autor oder Sammlung, Ressourcentyp, etc.

Schematische Übersicht (technischer Kontext in Text+)



Weitere technische Aspekte

- AAI
  - Beschränkung des Zugriffs auf bestimmte Ressourcen aus Urheber-, Lizenz- oder Datenschutzrechtlichen Gründen
  - Nutzung von föderierten Logins, z.B. DFN-AAI mit Shibboleth/SAML, zur Authentifizierung und Autorisierung
- Referenzierbarkeit über Persistente Identifikatoren (PIDs)
- Usability („One face to the customer“)

Deutsche Forschungsgemeinschaft

Das NFDI-Konsortium Text+ wird gefördert durch die Deutsche Forschungsgemeinschaft (DFG) unter der Projektnummer 460033370.

Text+ ist Teil der Nationale Forschungsdaten Infrastruktur

Text+ Plenary 2022, 12./13. September 2022, Mannheim