Gradl, Tobias; Henrich, Andreas

Die DARIAH-DE-Föderationsarchitektur : Datenintegration im Spannungsfeld forschungsspezifischer und domänenübergreifender Anforderungen

Tobias Gradl* und Andreas Henrich

Die DARIAH-DE-Föderationsarchitektur – Datenintegration im Spannungsfeld forschungsspezifischer und domänenübergreifender Anforderungen

DOI 10.1515/bfp-2016-0027

Zusammenfassung: Dieser Beitrag gibt einen Überblick über die Konzepte und Komponenten der DARIAH-DE-Föderationsarchitektur. Nach einer Abgrenzung grundlegender Anforderungen forschungsspezifischer und domänenübergreifender Integration wird zwischen dem Erstellungs- und Verwendungskontext von Forschungsdaten unterschieden und verdeutlicht, wie die Föderationsarchitektur bei deren Modellierung unterstützt. Die Anwendung im Rahmen der generischen Suche zeigt abschließend, wie in dieser Umgebung generische und forschungsspezifische Anfragen unterstützt werden können.

Schlüsselwörter: Forschungsdaten; Integration; Föderation; generische Suche; DARIAH-DE

The DARIAH-DE Federation Infrastructure – Data Integration between the Poles of Research Specific and Cross-domain Requirements

Abstract: This contribution presents an overview of the concepts and components of the DARIAH-DE federation architecture. After a distinction between requirements of research-specific and cross-domain integration, the creation context and application context of research data are differentiated along with the facilities for their explication within the research architecture, which is shown to support generic and specific queries within the generic search.

Keywords: Research data; integration; federation; generic search; DARIAH-DE

Konzepte und Dienste zur sammlungsübergreifenden Recherche geisteswissenschaftlicher Forschungsdaten und deren Analyse stehen im Trend der Forschung in den Digital Humanities. Aktuelle Arbeiten können dabei durchaus auf bestehenden Lösungen aufsetzen. Ein Interesse an und erste Untersuchungen zu Suchmöglichkeiten über Sammlungsgrenzen hinweg entstanden wenige Jahre nach ersten Installationen Digitaler Bibliotheken bereits vor der Jahrtausendwende. Im Jahr 2002 noch häufig als diskutierter, aber selten umgesetzter Traum bezeichnet,[1] stellt sich auch heute die Frage nach den Erfolgskriterien solcher übergreifenden Suchmöglichkeiten. Trotz aller Fortschritte in den Bereichen der Digitalisierung, digitalen Publikation und Konsolidierung von Ressourcen ist bereits die Vernetzung digitaler Bibliotheken kein triviales Problem. Durch eine Ausweitung des Sammlungsbegriffs auf weitere Institutionen, wie Museen und Archive, und die damit verbundene Generalisierung der betrachteten Ressourcen, Anwender und Informationsbedürfnisse wird die Beantwortung der Frage nach den Anforderungen und Erfolgskriterien übergreifender Such- und Analysemöglichkeiten – auf der abstrakten Ebene der Digital Humanities – weiter erschwert.

Nach einer Einführung in das grundlegende Problem der Spezifität von Daten, bieten wir in diesem Beitrag einen Überblick über das Konzept der DARIAH-DE-Föderationsarchitektur. Anhand einfacher Beispiele möchten wir dabei insbesondere skizzieren, wie die Komponenten für eine Modellierung von Daten und ihrer Zusammenhänge sowohl für forschungsspezifische Anforderungen, als auch für die Umsetzung allgemeiner, übergreifender Sichten herangezogen werden können. Interessiere Leser bitten wir, die Anwendbarkeit des Konzepts auf eigene Integrationsfragen auch selbst zu hinterfragen.[2]

***Kontaktperson: Tobias Gradl,** tobias.gradl@uni-bamberg.de
Andreas Henrich, andreas.henrich@uni-bamberg.de

1 Besser (2007).

2 Selbstverständlich freuen wir uns auch über neue Anforderungen an unsere Architektur und Dienste.

1 Datenstrukturen und Integrationsanforderungen

Ansätze zur Integration digitaler Sammlungen profitieren heute von den Bestrebungen nach Interoperabilität und Standardisierung, die häufig aus dem Bereich der Bibliotheken hervorgegangen sind. So wird der Zugriff auf digitale Sammlungen oft auf Basis standardisierter Schnittstellen wie dem Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[3] oder Z39.50[4] ermöglicht. Die Heterogenität der Daten wird vermindert durch die Entwicklung und Anwendung standardisierter Metadatenschemata. Diese reichen von einfachen, übergreifenden Standards wie dem Dublin Core Metadata Element Set (DC)[5] mit seinen 15 grundlegenden, nicht weiter eingeschränkten Elementen hin zu komplexen, fachspezifischen Strukturen, wie dem CIDOC Conceptual Reference Model (CRM)[6] zur Beschreibung von Konzepten und Beziehungen aus den verschiedensten Bereichen kulturellen Erbes oder auch den P5 Guidelines[7] der Text Encoding Initiative (TEI) mit ihren Schemata und Profilen zur Kodierung, Edition, Annotation und dem Austausch von Texten. Insbesondere bei komplexen, fach- oder medienspezifischen Schemata verschwimmt, die Grenze zwischen Daten und Metadaten. Ist bei DC ein Link auf einen beschriebenen Text erforderlich, so bietet beispielsweise TEI P5 Möglichkeiten, diesen direkt in einer entsprechend strukturierten Datei zu encodieren.

Untersuchungen[8] zeigen jedoch, dass lediglich DC als Standard breite Anwendung gefunden hat. Komplexere Standards werden dagegen entweder kaum verwendet oder aber entsprechend ausgezeichnete Daten werden nicht publiziert. Anstelle der Standards spielen in der Praxis oft originär in den einzelnen Sammlungen selbst entwickelte Datenmodelle eine größere Rolle. Während eine Überführung von Daten in standardisierte Formate oft mit hohen Aufwänden (und innerhalb der Sammlung mit nur geringem Nutzen) verbunden ist, sind Organisationen oftmals bereit, eine XML-basierte Repräsentation ihrer originären Daten über standardisierte Schnittstellen anzubieten. Um neben Standards auch sammlungsspezifische Formate unterstützen zu können und insbesondere auch, um sowohl domänenübergreifende, als auch forschungsspezifische Sichten auf Daten anbieten zu können, implementiert die DARIAH-DE-Föderationsarchitektur im Kern zwei wesentliche Eigenschaften:

- Unterschieden werden der *Erstellungskontext* zur Modellierung sammlungsspezifischer Datenmodelle sowie der *Verwendungskontext* zur Abbildung heterogener Quellen auf ein geeignetes Zielformat.
- Dieses Zielformat ist frei wähl- und modellierbar, wodurch eine größtmögliche Flexibilität im Hinblick auf die unterstützten Informationsbedürfnisse und Anwendungsfälle angeboten werden kann (vgl. Abb. 1): von der spezifischen Betrachtung einzelner oder weniger Sammlungen bis hin zu breiten Suchen auf Basis einfacher Schemata wie DC.

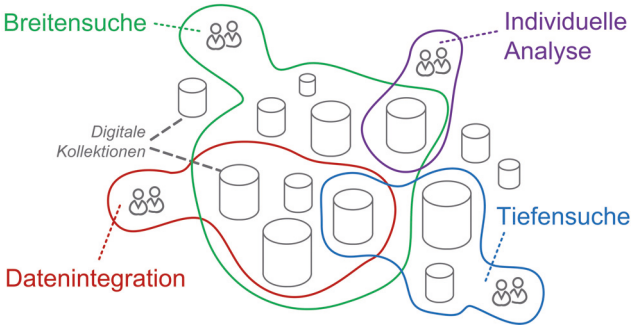

Abb. 1: Mögliche Anwendungsfälle der Integration digitaler Sammlungen

2 Beschreibung des Erstellungskontexts

Generative Aktivitäten an Forschungsdaten, insbesondere deren Erstellung, Publikation und Auszeichnung durch Metadaten, finden unter bestimmten Rahmenbedingungen statt. Diese sind definiert durch implizites Hintergrundwissen zur betrachteten Forschungsfrage, einer akademischen Umgebung und zu der relevanten Kollektion. Im Vergleich zu traditionellen Medien hat dieser Erstellungskontext eine größere Bedeutung für die Interpretation digitaler Daten, auch weil neben dem eigentlichen Inhalt technische Ebenen eingeführt werden, wie das Encoding der Daten oder – insbesondere bei binären/proprietären Formaten – der Verarbeitungssoftware. Wurden Daten beispielsweise mithilfe einer Softwareanwendung erzeugt, die nun nicht mehr beschafft werden kann oder aber auf aktuellen Rechnern nicht lauffähig ist, könnten technische Bedingungen des Erstellungskontexts ggf. nicht mehr reproduziert werden und Daten wären nicht darstellbar. Eine Grundvoraussetzung für die Interpretierbarkeit, Nachnutzbarkeit und Interoperabilität von For-

3 http://www.openarchives.org/pmh/.
4 http://www.loc.gov/z3950/agency/document.html.
5 http://dublincore.org/documents/dces/.
6 http://www.cidoc-crm.org/.
7 http://www.tei-c.org/Guidelines/P5/.
8 Polfreman (2005) und Vierkant (2013).

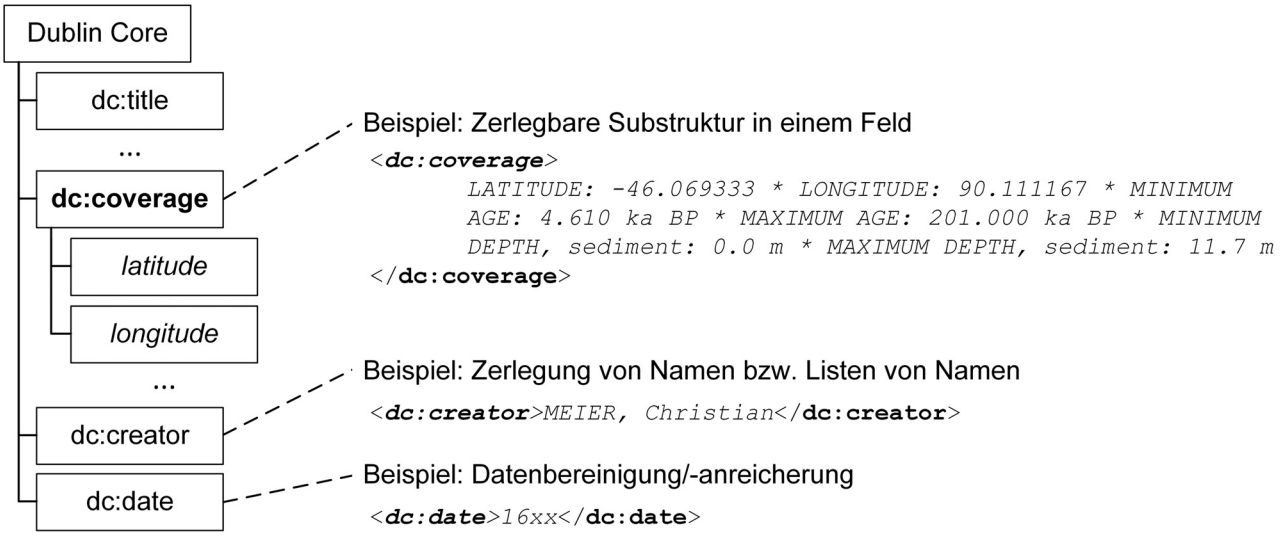

Abb. 2: Beispiele anreicherbarer Strukturelemente

schungsdaten besteht somit in der Verwendung offener Standards. In ihren Richtlinien zur Bewahrung des digitalen Erbes weist beispielsweise auch die UNESCO darauf hin, dass quelloffene, nicht-proprietäre Standards ein späteres Auffinden bzw. die Entwicklung notwendiger Werkzeuge für den Zugriff auf relevante Daten erleichtern.[9]

Auch die DARIAH-DE-Föderationsarchitektur setzt auf die Anwendung standardisierter technischer Protokolle, definiert jedoch keine strukturellen, syntaktischen oder semantischen Restriktionen.[10] Konkret bedeutet dies, dass Daten über standardisierte Schnittstellen oder als Datei in einem unterstützten Format, wie der Extensible Markup Language (XML), JavaScript Object Notation (JSON), Comma-separated values (CSV) oder aber auch als einfacher Volltext vorliegen müssen. Sämtliche über OAI-PMH zugreifbaren digitalen Sammlungen erfüllen so beispielsweise aufgrund der Verwendung von XML zur Strukturierung der Daten bereits diese Anforderungen. Die Komponenten der Föderationsarchitektur[11] bieten dann neben der Registrierung und Beschreibung von Kollektionen insbesondere Möglichkeiten zur Modellierung der für eine korrekte Interpretation notwendigen Eigenschaften des Erstellungskontexts. Ein einfaches Beispiel: Ein Eintrag 1299 könnte in einem Feld *date* auch ohne weiteres Wissen leicht als Jahreszahl festgemacht werden. Beinhaltet der Erstellungskontext aber die Anwendung eines aus einer abstrakt-deutschsprachigen Perspektive untypischen Kalenders, so kann die Explikation dieses Wissens eine korrekte Interpretation der Jahreszahl in anderen Kontexten erleichtern. Würde beispielsweise der islamische Kalender im Erstellungskontext des Datums Anwendung finden, entspricht das Jahr 1299 je nach Monat dem Jahr 1881 oder 1882 des gregorianischen Kalenders. Die Oberfläche der DARIAH-DE Schema Registry umfasst für dieses Beispiel die Möglichkeit, ein neues Strukturfeld *calendar* anzulegen oder aber ein zusätzliches Feld *gregorian_date* zu definieren, welches nach benutzerdefinierten Regeln berechnet wird.

Abb. 2 zeigt weitere Beispiele auf Basis von DC, bei denen Hintergrundwissen modelliert werden kann, um die maschinelle Interpretation der Daten zu erleichtern. Entnommen aus einem Datensatz von Pangaea[12] besteht der Inhalt des Feldes *coverage* aus der Aneinanderreihung einzelner Schlüssel/Wert-Paare, die sich einem Menschen auf dem ersten Blick als solche zu erkennen geben. Dadurch, dass diese Substruktur jedoch nicht mithilfe von XML Markup dargestellt wird, kann eine maschinelle Verarbeitung diesen Inhalt ohne die Anwendung spezifischer Regeln zunächst nur als Volltext erkennen. Die Verarbeitung von Namensbestandteilen im Feld *creator* sowie unvollständige bzw. ungenaue Angaben im Feld *date* können ebenso mithilfe von Verarbeitungsregeln extrahiert, bereinigt oder angereichert werden.

Der Bildschirmausschnitt in Abb. 3 zeigt neben dem eigentlichen Schemaeditor der DARIAH-DE Schema Registry auch die Ergebnisse der beispielhaften Transformation eines übergebenen Datensatzes. Die für *creator* und *subjectList* hinterlegten Regeln zur Interpretation und Weiter-

9 UNESCO (2003, 58 f.)
10 Gradl und Henrich (2014, 3 f.).
11 Gradl, Henrich und Plutte (2015, Abschnitt 3).
12 http://doi.pangaea.de/10.1594/PANGAEA.51915.

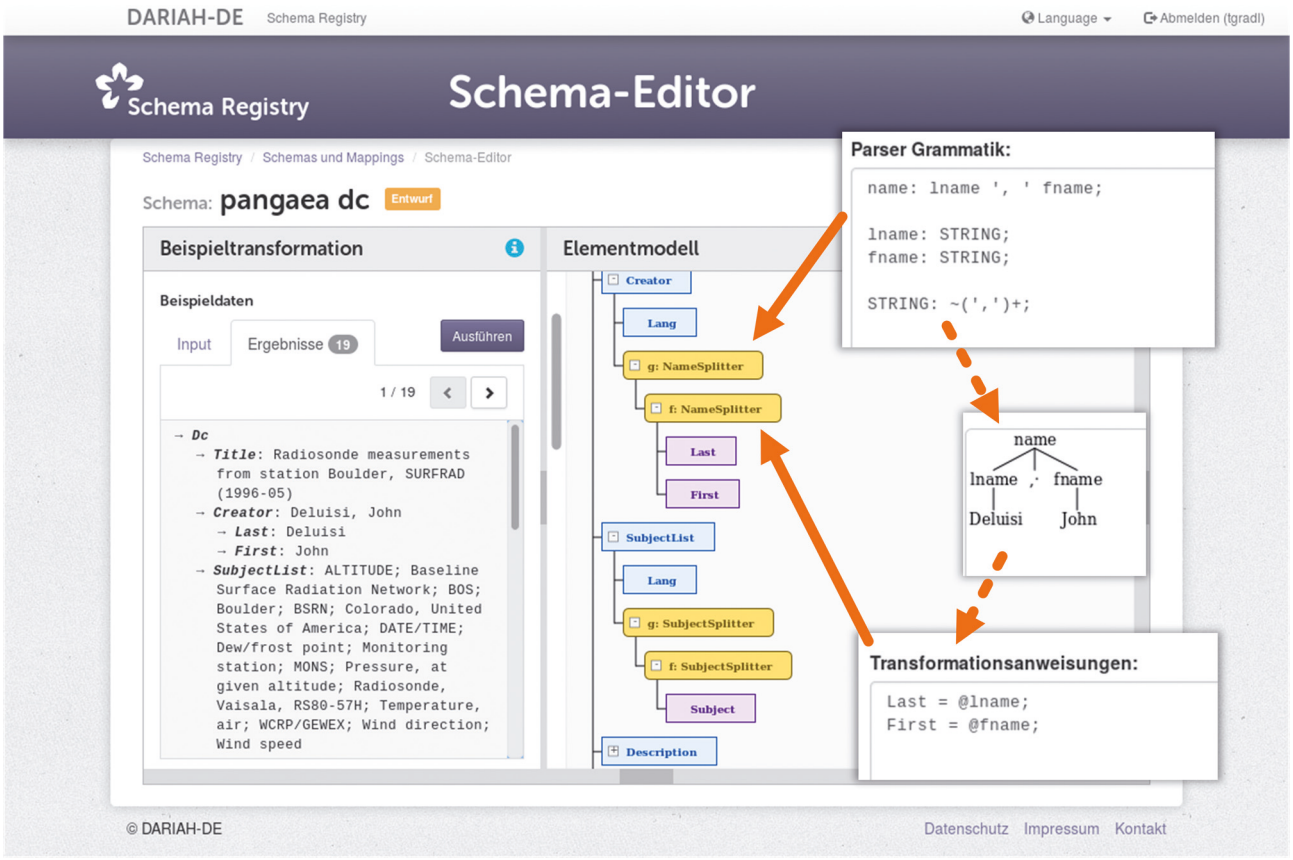

Abb. 3: Schema-Editor im Rahmen der DARIAH-DE Schema Registry

verarbeitung der Daten führen im Beispiel zur Generierung von untergeordneten Strukturelementen und damit zur Anreicherung des Datensatzes auf Basis der hinterlegten Regeln. Das Ergebnis im Beispiel (Abb. 3, links) zeigt, dass neben den zusätzlichen Daten auch die ursprünglichen Daten stets erhalten werden – so beispielsweise die Aneinanderreihung von Schlüsselworten im originären Feld *subject*. In der Abbildung wird auch das zugrundeliegende Verarbeitungskonzept deutlich:[13] Eine kontextfreie Grammatik (context free grammar, CFG) wird durch den Anwender formuliert und definiert eine Sprache zur Analyse beinhalteter Daten. Blätter in einem durch Anwendung der Grammatik erzeugten Syntaxbaum sind gezielt adressierbar, wodurch Transformationsanweisungen im zweiten Schritt auf eine verfeinerte Version der ursprünglichen Daten zurückgreifen können. Werden im Beispiel nur einfache Zuweisungen als Transformationsanweisungen dargestellt, so sind an dieser Stelle auch komplexere Befehle möglich, wie die Ansteuerung computerlinguistischer Verfahren oder externer Webschnittstellen.

3 Transformation in den Verwendungskontext

Ausgehend von den Daten und den um relevantes Hintergrundwissen ergänzten Strukturinformationen wird mit dem Aspekt der Datentransformation eine Umwandlung von Daten in einen definierten Verwendungskontext erreicht. Verwendungskontexte reichen von einfachen, übergreifenden Informationsbedürfnissen, wie der Suche nach Werken eines Autors im Rahmen der generischen Suche hin zu den spezifischen Gegebenheiten und Fragen konkreter Forschungsprojekte. Mit der DARIAH-DE-Föderationsarchitektur wird nun eine Infrastruktur angeboten, die Transformations- und Integrationsanforderungen generisch für verschiedenartige Verwendungskontexte unterstützt. So kann ein Mapping von Daten mit einfachen Schemata wie DC zur integrierten Darstellung einer breiten, disziplinübergreifenden Datenbasis verwendet werden. Die Assoziation von Daten mit projekt- oder disziplinspezifischen Datenmodellen unterstützt Forschende bei der Einrichtung integrierter Sichten auf heterogene Daten, bei denen jedoch die für die Forschungsfrage relevanten Elemente erhalten werden bzw. in eine angereicherte Form

13 Vgl. insbesondere Gradl und Henrich (2014) und Gradl und Henrich (2016).

überführt werden können. Auch im Hinblick auf die Integration der Daten abstrahiert die Föderationsarchitektur von technischen Problemen und lenkt den Fokus des Anwenders auf die Modellierung inhaltsbezogener Aspekte der Integration.

Aufbauend auf dem eingeführten Beispiel der im Feld *coverage* eingebetteten Substruktur, könnte eine beispielhafte Weiterverwendung der Daten eine Transformation in die Keyhole Markup Language (KML)[14] vorsehen. Die KML ist eine vom Open Geospacial Consortium standardisierte Sprache für die Auszeichnung von Geodaten und findet Anwendung in unterschiedlichen Kontexten und Applikationen. Sie eignet sich insbesondere, um eine harmonisierte Darstellung von Geodaten aus heterogenen Quellen zu erreichen. In KML integrierte Daten können leicht im Rahmen geotemporaler Analysen und Visualisierungen z. B. auch im Rahmen des DARIAH-DE Geobrowsers[15] genutzt werden. Der folgende vereinfachte Ausschnitt zeigt die Definition des Feldes *coverage* als kontextfreie Grammatik. Obwohl ggf. nicht ohne weitere Erklärung verständlich, vermittelt die Grammatik zumindest einen Eindruck über die im Gegensatz zu einer vollständigen und spezifischen Implementierung geringen Aufwände für die Spezifikation solcher inhaltsbasierter Regeln.

Tab. 1: Vereinfachte Grammatik zur Spezifikation des Pangaea-coverage-Feldes

```
substruct     : subelem+;
subelem       : (longitude | latitude | start | end | minDepth | max-
                Depth | otherElem)
                SEPARATOR?;
longitude     : 'LONGITUDE' ': ' value;
latitude      : 'LATITUDE' ': ' value;
start         : 'DATE/TIME START' ': ' value;
end           : 'DATE/TIME END' ': ' value;
minDepth      : 'MINIMUM DEPTH, sediment/rock' ': ' value;
maxDepth      : 'MAXIMUM DEPTH, sediment/rock' ': ' value;
otherElem     : key ': ' value;
key           : ID;
value         : DATE
              | ID;

ID            : ~(' '|':'|'*') ~(':'|'*')+ ~(' '|':'|'*');
DATE          : YEAR '-' MONTH '-' DAY 'T' HOUR ':' MIN ':' SEC;
SEPARATOR     : ' '? '*' ' '?;
fragment YEAR      : [1–2][0–9][0–9][0–9];
fragment MONTH     : [0–1][0–9];
fragment DAY       : [0–3][0–9];
fragment HOUR      : [0–2][0–9];
fragment MIN       : [0–6][0–9];
fragment SEC       : [0–6][0–9];

WS            : [\t\r\n]+ -> skip;
```

Bei einer Spezifikation der Substruktur im Feld *coverage* und der anschließenden Einrichtung entsprechender Subelemente im Rahmen der Modellierung des Erstellungskontexts reicht die einfache Assoziation der Felder (siehe Abb. 4) in Quell- und Zielschemata aus, um eine Transformation in den Zielkontext zu erreichen.

In der Abb. 4 finden sich auch die Transformationsanweisungen, die zu der Generierung der Subelemente *Long*, *Lat* und *City* führen. So wird die nächstgelegene Stadt mithilfe der Nutzung implementierter Geofunktionalität des zugrundeliegenden Transformationsframeworks auf Basis der extrahierten Koordinaten ermittelt.

Grundsätzlich könnte auch argumentiert werden, dass die Anreicherung des Datensatzes um die Angabe der nächsten Stadt nicht vom Erstellungskontext, sondern der konkreten Verwendung der Daten abhängig ist. Neben dem Vorteil, dass eine einmalige Modellierung im Erstellungskontext eine mehrfache Modellierung an unterschiedlichen Mappings ersetzen kann, zeigen wir im folgenden Abschnitt, welchen Einfluss die Entscheidung über Erstellungs- oder Verwendungskontext auf das konkrete Laufzeitverhalten aufsetzender Dienste hat.

4 Implikationen für die generische Suche

Abb. 5 zeigt die Zusammenhänge zwischen den für die DARIAH Infrastruktur zugänglichen Kollektionen, den Registries der DARIAH-DE-Föderationsschicht und der generischen Suche.

Aus dieser Übersicht geht insbesondere auch hervor, dass für die generische Suche zwei klar abgegrenzte Phasen unterschieden werden müssen, die sich jeweils konkret den diskutierten Kontexten widmen.

– Indexierung: Für den Zugriff auf Kollektionen und die darin verwalteten Daten interagiert die generische Suche zunächst mit der Collection Registry, um Informationen über verfügbare Kollektionen und deren Schnittstellen zu erhalten. Mithilfe der Schema Registry erhält die generische Suche die um Hintergründe des Erstellungskontexts erweiterten Schemata zur Interpretation, Verarbeitung und Anreicherung der Daten.

14 http://www.opengeospatial.org/standards/kml.
15 http://geobrowser.de.dariah.eu/.

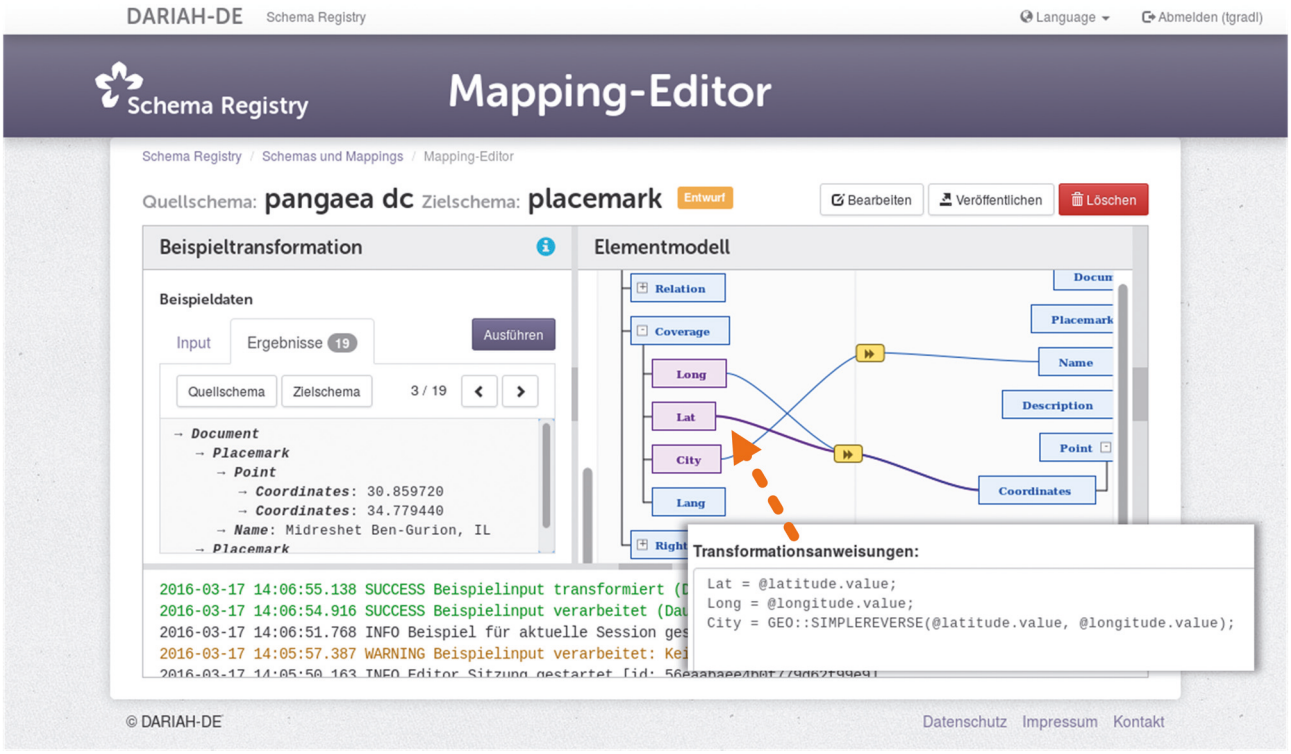

Abb. 4: Mapping-Editor im Rahmen der DARIAH-DE Schema Registry

– Anfragebearbeitung: Für eine Suchanfrage mit den konkret spezifizierten Suchfacetten selektiert die generische Suche relevante Schemata und wertet deren Zusammenhänge aus. Werden durch den Anwender Kollektionen für eine Suche selektiert, die in Bezug auf ihre Schemata einen engen semantischen Zusammenhang aufweisen, können dynamisch feinere Suchfacetten angeboten werden. Soll eine Suche dagegen über eine breite Menge von Kollektionen durchgeführt werden, stehen nur wenige, oft verwendete Facetten (Autor, Schlüsselworte etc.) zur Verfügung.

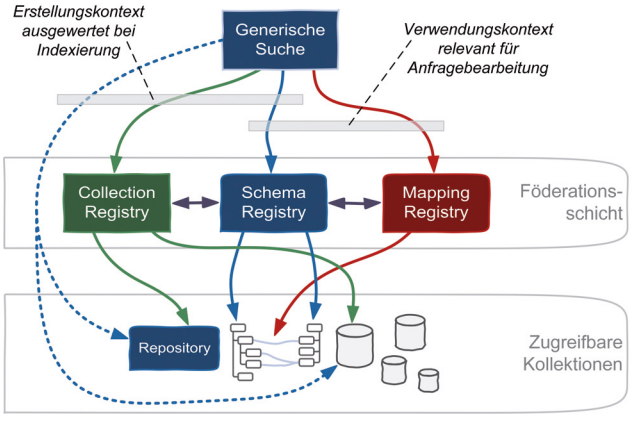

Abb. 5: Zusammenhang zwischen betrachteten DARIAH-DE Komponenten

Neben der unterschiedlichen Semantik der Modellierung des Erstellung- und Verwendungskontext, hat die Unterscheidung zwischen diesen Kontexten auch Einfluss auf die technische Verarbeitung der Daten in der generischen Suche. Sämtliche Verarbeitungsanweisungen des Erstellungskontexts, also grammatikalische Regeln und Transformationsregeln, die zu einer verwendungsagnostischen Erweiterung eines Schemas führen, werden zum Zeitpunkt der Indexierung der Daten ausgewertet. Der Grund hierfür besteht darin, dass Daten im Rahmen der generischen Suche genau einmal mit dem jeweils maximal möglichen Grad integrierter Semantik gespeichert werden. Inwiefern diese Semantik genutzt werden kann, entscheidet sich dann jeweils zur Anfragezeit. Der Vorteil einer Auswertung von Regeln zum Zeitpunkt der Indexierung besteht darin, dass diese unabhängig von konkreten Anfragen in Form eines Hintergrundprozesses der generischen Suche ausgeführt werden kann und keine zusätzliche Berechnung zum Anfragezeitpunkt benötigt.

Die Erkennung der nächstgelegenen Stadt im vergangenen Abschnitt (vgl. Abb. 4) bedingt die Abfrage einer Datenbank ausgehend von ermittelten Koordinaten jedes einzelnen Eintrags. Bei einer Menge von mehreren 100 000 Datensätzen würde eine Ermittlung der Stadt zum Zeitpunkt der Anfragebearbeitung mehrere Sekunden bis Minuten in Anspruch nehmen. Ein komplexeres Beispiel findet sich im Rahmen des DARIAH-DE-Use-Cases „Bio-

graphien".[16] Hier werden unter anderem auf Basis von Wikipedia biographische Profile zu Personen erstellt. Mithilfe grammatikalischer Regeln werden einzelne Wikipedia-Seiten in Inhalts- und Strukturbestandteile zerlegt. Weitere Regeln spezifizieren biographisch relevante Inhaltsbereiche (z. B. Blöcke mit passenden Überschriften wie „Leben" und „Herkunft"). Durch Anwendung von Transformationsregeln werden Techniken des Natural Language Processing (NLP) auf die relevanten Bereiche angewendet, Quadrupel bestehend aus <Person, Ort, Zeit, Ereignis> extrahiert und in ein zentrales, strukturiertes Schema für biographische Daten überführt. Die Anwendung der NLP-Techniken dauert je nach Umfang des Artikels mehrere Sekunden bis Minuten und kann bei derzeit etwa 500 000 Volltext-Einträgen[17] zu Personen nicht im Rahmen von Anfragen ausgeführt werden, sondern muss bei der Indexierung erfolgen.

5 Ausblick

Für diesen Beitrag wurden bewusst einfache Beispiele gewählt, die eine Übertragung auf eigene Daten der Leser erleichtern sollen. Mit der Extraktion bestimmter semantischer Daten aus den Volltexten der Wikipedia wurde ein komplexerer Anwendungsfall angesprochen. Hervorzuheben ist in diesem Fall vor allem, dass die im Volltext von Wikipedia-Einträgen vorkommenden Substrukturen grammatikalisch separiert und explizit ausgenutzt werden. Einmal werden Überschriften nach benutzerdefinierten Regeln daraufhin untersucht, ob diese biographisch relevante Texte beinhalten können. Außerdem werden interne Links, beispielsweise zu Personen und Orten, als definierte Eingangswerte für die spätere NLP-Verarbeitung genutzt, um z. B. bei der Disambiguierung von Begriffen zu unterstützen.

Inwiefern das DARIAH-DE-Föderationskonzept die selbst gesetzten Erfolgskriterien nach einer generischen Unterstützung übergreifender *und* spezifischer Integrationsbedürfnisse erfüllt, werden die verschiedenen geisteswissenschaftlichen Anwendungsfälle zeigen, an denen die Föderationsarchitektur unterstützend mitwirken darf. Erste untersuchte Daten und Informationsbedürfnisse – beispielsweise im Rahmen der biographischen Analysen – deuten die Ausdrucksmächtigkeit und Flexibilität des Konzepts an. Die Unterstützung weiterer spannender geisteswissenschaftliche Forschungsfragen steht in den Startlöchern.

Literaturverzeichnis

Besser, Howard (2007): The Next Stage: Moving from Isolated Digital Collections to Interoperable Digital Libraries. In: *First Monday*, 7(6). doi: 10.5210/fm.v7i6.958.

Gradl, Tobias; Henrich, Andreas (2014): A novel approach for a reusable federation of research data within the arts and humanities. In: *Digital Humanities*, 382–84. Verfügbar unter http://dharchive.org/paper/DH2014/Paper-779.xml.

Gradl, Tobias; Henrich, Andreas (2016): Nutzung und Kombination von Daten aus strukturierten und unstrukturierten Quellen zur Identifikation transnationaler Lebensläufe. Tagung des Verbands „Digital Humanities im deutschsprachigen Raum e. V." 7.–12. März 2016. Leipzig.

Gradl, Tobias; Henrich, Andreas; Plutte, Christoph (2015): Heterogene Daten in den Digital Humanities: Eine Architektur zur forschungsorientierten Föderation von Kollektionen. In: *Grenzen und Möglichkeiten der Digital Humanities*, Hg. v. Constanze Baum und Thomas Stäcker. Wolfenbüttel: HAB – Herzog August Bibliothek. Verfügbar unter http://zfdg.de/sb001_020.

Polfreman, Malcolm (2005): Commonly-used metadata formats in the Arts and Humanities. verfügbar unter http://www.ahds.ac.uk/metadata/arts-humanities-metadata-formats.htm.

UNESCO (2003): Guidelines for the preservation of digital heritage. Verfügbar unter http://portal.unesco.org/ci/en/ev.php-URL_ID=13271.

Vierkant, Paul (2013): Leuchttürme der deutschen Repositorienlandschaft. Verfügbar unter http://de.slideshare.net/paulvierkant/leuchttrme-der-deutschen-repositorienlandschaft.

Tobias Gradl
An der Weberei 5
D-96047 Bamberg
tobias.gradl@uni-bamberg.de

Andreas Henrich
An der Weberei 5
D-96047 Bamberg
andreas.henrich@uni-bamberg.de

16 Vgl. Gradl und Henrich (2016).
17 http://search.de.dariah.eu/cosmotool.