Andreas Henrich und Tobias Gradl

INTEGRATION VON FORSCHUNGSDATEN

Wie können Forschungsinfrastrukturen helfen?

1. Problem

Methoden der kultur- und geisteswissenschaftlichen Forschung sind so vielschichtig und spezifisch wie die Disziplinen und Forschungsfragen selbst. Wenn über Möglichkeiten zur Unterstützung durch generische Forschungsinfrastrukturen diskutiert wird, muss deshalb die Frage nach der Sinnhaftigkeit und dem Nutzen von Infrastrukturprojekten in den Kultur- und Geisteswissenschaften berechtigt sein. Zur gedanklichen Auseinandersetzung schadet es dabei nicht, eine Analogie mit den Naturwissenschaften zu wagen: Könnten Forschende der theoretischen Physik prinzipiell autark mit Stift und Papier arbeiten und dabei herausragende Forschungsergebnisse hervorbringen, so gilt dies trotz aller Verschiedenartigkeit in ihren Methoden auch für Forschende in den klassischen Geisteswissenschaften. Im Gegensatz zu ihrem theoretischen Pendant ist die Experimentalphysik heute aber weitgehend von infrastrukturellen Komponenten und Institutionen abhängig, um Einrichtung und Wartung von einer tatsächlichen Nutzung zu entkoppeln und einen Teilchenbeschleuniger oder ein Radioteleskop finanzieren und für Forschende unterschiedlicher organisatorischer Zugehörigkeit bereitstellen zu können. Auch wenn die Analogie insofern hinkt, dass finanzielle Aufwände in den Digital Humanities kaum mit denen der Experimentalphysik vergleichbar sind, so sind Vorteile einer Spezialisierung erkennbar. Auch hier könnten die oft kleinteilig geförderten Forschungsprojekte nachhaltig von Lösungen im Umfeld von Langzeitarchivierung, Nachnutzbarkeit und rechtlichen Fragestellungen profitieren. Und wie auch die Digital Humanities ihren Nutzen und ihre wissenschaftliche Berechtigung gegenüber klassischen Geisteswissenschaften rechtfertigen müssen, so wird oftmals auch die Experimentalphysik als bloßes Werkzeug zur Überprüfung entwickelter Theorien verkannt – bis einmal mehr ein überraschendes Versuchsergebnis das theoretische Gebäude seiner Wissenschaft sprengt und nach neuen Antworten verlangt.

DARIAH-DE ist eine Forschungsinfrastruktur für die Kultur- und Geisteswissenschaften[1] und fokussiert neben einer eigenen inhaltlichen Forschung insbesondere die Unterstützung geisteswissenschaftlicher Forschungsprojekte. Die Übersichtskarte von DARI-

AH-DE (Abb. 1) zeigt die zum Teil ineinander übergehenden Betrachtungsebenen von Lehre, Forschung und technischer Infrastruktur. Im Kasten „Software Hosting Services" finden sich Softwarekomponenten in folgenden Bereichen:

- *technische Infrastruktur*: Basisdienste, operative IT-Dienste und Hosting zur Kollaboration.
- *Forschung*: Generische Dienste unterstützen hier fachübergreifend auftretende, inhaltliche Bedürfnisse. Demonstratoren und fachwissenschaftliche Dienste dienen der Evaluation der eigenen Infrastruktur von DARIAH-DE und implementieren darüber hinaus konkrete fachwissenschaftliche Anwendungsfälle.[2]

Bei aller Grundsätzlichkeit und Ganzheitlichkeit in der Ausrichtung von DARIAH-DE widmet sich dieser Artikel insbesondere den Möglichkeiten zur Zusammenführung heterogener Datenbestände mit Hilfe der infrastrukturellen Komponenten von DARIAH-DE. Nach einer Diskussion von Begriffen und Konzepten im Kontext von Forschungsdaten, Daten und Metadaten soll in den nachfolgenden Abschnitten insbesondere ein Einblick in Ideen und Dienste der DARIAH-DE Föderationsarchitektur gegeben werden.

2. Forschungsdaten

Forschungsdaten entstehen durch die Anwendung fachwissenschaftlicher Methoden und bilden das Ergebnis von Forschungsprozessen. Sie werden als Ausdruck eines spezifischen Erstellungskontextes erarbeitet und sind definiert durch komplexe Rahmenbedingungen wie Disziplin, Projekt und Forschungsfrage. Aus einem abstrakten Blickwinkel heraus erscheinen Forschungsdaten aufgrund der Vielschichtigkeit der wissenschaftlichen Kontexte oft derart spezifisch und heterogen, dass eine disziplinübergreifende, integrative Betrachtung zunächst kaum zielführend scheint. Und dennoch können bei aller Diversität kultur- und geisteswissenschaftlicher Kontexte wesentliche Eigenschaften identifiziert werden, anhand derer eine Klassifikation von Forschungsdaten ermöglicht werden kann. Diese sind beispielsweise:

- *Medienformat*: unstrukturierte Volltexte, Annotationen, Fotografien, Scans, Audio und Videodateien etc.
- *Provenienz*: Hinweise auf die kontextuelle Herkunft von Daten erleichtern beziehungsweise ermöglichen oft erst eine korrekte Interpretation von Forschungsdaten
- *Granularität*: Forschungsdaten können zum Beispiel auf der Ebene von Sammlungen, beinhalteten Objekten oder – bei zusammengesetzten Objekten – Teilobjekten und Facetten erarbeitet werden und gegebenenfalls diese Zusammenhänge auch widerspiegeln
- *Sprache*: Hierbei klassifiziert insbesondere die zur Erstellung von Forschungsdaten verwendete, natürliche Sprache, darüber hinaus auch technische Sprachen zur Speicherung und Übermittlung von Forschungsdaten, wie zum Beispiel die Extensible Markup Language (XML).[3]

Anhand solcher Merkmale gebildete Klassen von Forschungsdaten weisen gemeinsame Eigenschaften auf, die zumeist eine generische Unterstützung auf technischer Ebene (zum Beispiel Verarbeitungs- und Betrachtungssoftware) und darüber hinaus oft auch auf einer inhaltlichen Ebene erlauben. Forschungsdaten mit vergleichbaren Erstellungskontexten und Merkmalen werden gemeinsam in thematischen Kollektionen archiviert und über diese bereitgestellt – vergleichbar mit Abteilungen oder Katalogen in klassischen Bibliotheken und Sammlungen.

2.1 Thematische Kollektionen

Die Nachnutzung digitaler Forschungsdaten zur Anwendung unterschiedlicher Methoden desselben Fachs, in interdisziplinären Kontexten oder generisch im Rahmen von Forschungsinfrastrukturen, stützt sich gerade auf die teilweise Homogenität von Forschungsdaten in digitalen Kollektionen. Diese ermöglicht die Reduzierung der Komplexität von Daten im Hinblick auf fachliche und fachübergreifende Eigenschaften und erleichtert dadurch eine spätere Nachnutzung. Viele der für Forscherinnen und Forscher zur Verfügung stehenden digitalen Kollektionen weisen eine inhärente, fachunabhängige Homogenität auf. Dies äußert sich darin, dass Dokumente über eine einheitliche technische Schnittstelle angeboten werden. Die derzeit wohl bedeutendste und am häufigsten implementierte Schnittstelle im Bereich der Kultur- und Geisteswissenschaften beschreibt das „Protocol for Metadata Harvesting der Open Archives Initiative" (OAI-PMH).[4] Zudem folgen Daten oft einheitlichen technischen Spezifikationen wie XML und weisen strukturelle Ähnlichkeiten auf – im Fall von XML-Dokumenten spezifiziert beispielsweise durch XML-Schemata.[5] Neben gemeinsamen technischen Eigenschaften von Forschungsdaten innerhalb einer Kollektion zeigen diese typischerweise auch inhaltliche Überschneidungen zum Beispiel in Bezug auf betrachtete Personen, Zeiträume oder Konzepte. Die Identifizierung solcher inhaltlichen Eigenschaften durch infrastrukturelle Komponenten ermöglicht dann beispielsweise eine Einschätzung der Relevanz einzelner digitaler Kollektionen für eine Forschungsfrage.

Während der Begriff der „digitalen Kollektion" im Bereich der Digital Humanities weitläufige Verwendung findet, eignet sich die Beschreibung der „thematischen Kollektion" durch John Unsworth, um zentrale Eigenschaften zusammenzufassen. Demnach sind (digitale) „thematische Kollektionen" nach Unsworth:[6]

- elektronisch
- heterogen in Bezug auf Datentypen
- reichhaltig, dabei thematisch zusammenhängend
- strukturiert, dabei aber erweiterbar
- entwickelt, um Forschung zu unterstützen
- von einer oder mehreren Autorinnen beziehungsweise Autoren verfasst
- interdisziplinär
- Sammlungen digitaler, primärer Forschungsobjekte.

2.1.1 Exkurs: epidat

Als Beispiel einer thematischen Kollektion dient an dieser Stelle die Datenbank „epidat" des Salomon Ludwig Steinheim-Instituts für deutsch-jüdische Geschichte an der Universität Duisburg-Essen. Mit Hilfe von epidat werden wertvolle Forschungsdaten der jüdischen Grabsteinepigraphik nachnutzbar zusammengefasst, die oft im Rahmen kleinerer Forschungsprojekte entstehen und die ohne ein derartiges System wohl ohne Möglichkeiten einer Nachnutzung geographisch verteilt erhoben und abgelegt würden. Die Datenbank aggregiert derzeit 174 digitale Editionen mit 31.154 Grabinschriften und insgesamt 61.168 Bilddateien und stellt diese für eine Nachnutzung bereit.[7]

Ein solches, über epidat zugängliches Beispielprojekt ist das im Auftrag der Israelitischen Kultusgemeinde Bayreuth durchgeführte Projekt zur Dokumentation des jüdischen Friedhofs in Bayreuth. Dieses resultierte bislang in der Erfassung von Forschungsdaten zu 957 dokumentierten Grabsteinen sowie Daten auf der übergeordneten Ebene des Friedhofs selbst.[8] Letztere führten dabei insbesondere zu einer detaillierten Beschreibung der Lage der dokumentierten Grabsteine (Abb. 2) und Friedhofs-Metadaten im TEI P5 Format.[9]

Die Forschungsdaten zu Grabsteinen selbst beinhalten neben der fotografischen Dokumentation insbesondere textuelle Abbildungen der Inschriften sowie deren Übersetzung im Rahmen von interoperablen und nachnutzbaren TEI P5- beziehungsweise epiDoc-Dokumenten[10] (Abb. 3). Letzteres ist ein auf TEI P5 basierender Standard für die Edition epigraphischer Dokumente.

Epidat qualifiziert sich dabei nach den Kriterien von Unsworth als idealtypische, thematische Kollektion: Epidat ist eine *elektronische* Kollektion und beinhaltet mit TEI-Dokumenten, Lageplänen und Fotografien *heterogene Datentypen*. Die Primärdaten, transkribierte hebräische Handschriften und deren Übersetzung ins Deutsche, sind Facetten der *reichhaltigen* Forschungsdaten in epidat. Eine *thematische Kohärenz* ist ebenso erkennbar wie auch die Ausrichtung auf die *Unterstützung interdisziplinärer epigraphischer und judaistischer Forschung*. Eine *Erweiterbarkeit* von epidat ist dadurch gegeben, dass fachwissenschaftlich relevante Aspekte aufgrund der Erweiterbarkeit von TEI-Profilen problemlos ergänzt werden können. Neu erfasste Friedhöfe und Grabsteine erweitern epidat dagegen mengenmäßig.

Neben der primären Aufgabe der nachhaltigen Dokumentation und damit des Erhalts kulturellen Erbes ergibt sich aus der offenen und interoperablen Form der in epidat abgelegten Forschungsdaten die direkte Möglichkeit einer Nachnutzung. So können die durch lokale Förderung und Forschung angesammelten Daten – entsprechende Rechte vorausgesetzt – jederzeit in einer übergreifenden Form bereitgestellt und für weiterführende Forschung angeboten werden.

2.2 Metadaten sind Modelle

Wenngleich thematische Kollektionen intuitiv und auch nach Definition von Unsworth primäre Forschungsobjekte enthalten, werden über die Schnittstellen dieser Kollektionen

zunächst meist Metadaten angeboten. Diese können einerseits einen Verweis auf das digitale Primärobjekt enthalten, andererseits werden im Fall besonders umfangreicher Metadaten, wie TEI-basierter Dokumente, Aspekte des Primärobjektes übernommen. Eine trennscharfe Unterscheidung zwischen Primär- und Sekundärdaten (Metadaten) ist daher oft nicht möglich und gegebenenfalls auch nicht notwendig, da sich auch Metadaten je nach Forschungsfrage und Reichhaltigkeit selbst oftmals als Forschungsdaten per se erweisen können. Mit zunehmender Strukturiertheit, Detailliertheit und semantischer Reichhaltigkeit von Metadaten werden Aussagen über das Forschungsobjekt auf Basis der Metadaten ermöglicht: Wird etwa die Inschrift eines jüdischen Grabsteins in ausreichender Detailliertheit erfasst und übersetzt, können Suchanfragen in deutscher Sprache an diese Metadaten ausgeführt werden und relevante Primärdaten liefern.

Bei allen Möglichkeiten der semantischen Ausgestaltung von Metadaten ist ein Verständnis für deren Modellhaftigkeit von essenzieller Bedeutung für den zielführenden Umgang mit digitalen Daten: Nach Stachowiak[11] ist ein Modell eine unvollständige Abbildung von etwas *Komplexerem*. Ein Modell wird dabei im Wesentlichen aus einem bestimmten Blickwinkel erzeugt, verfolgt einen bestimmten Zweck und vernachlässigt gegebenenfalls Aspekte eines modellierten Objektes, die im Kontext der verfolgten Zielsetzung irrelevant sind.

Beschreibungen von Forschungsdaten erfüllen die Kriterien eines Modells nach Stachowiak: Sie sind *Abbildungen* von Primärobjekten, *verkürzen* dieses um Aspekte und Merkmale, die für den Erfassungskontext irrelevant sind und unterliegen einem *Pragmatismus* zu deren Optimierung auf eine bestimmte Zielgruppe, Disziplin oder Forschungsfrage.

2.3 Unterstützungspotentiale

Die Bereitstellung von Forschungsdaten zur Nachnutzung ist wichtig und sinnvoll. Kritisch für den Erfolg im Umgang mit digitalen Daten und Metadaten ist dabei das Bewusstsein, dass Forschungsdaten immer zielgerichtete Abbildungen beziehungsweise Modelle des Forschungsobjektes – und somit eine absichtliche Vereinfachung – darstellen. Durch die Bereitstellung von Daten können Forschungsinfrastrukturen die Auffindbarkeit und Verarbeitbarkeit von Daten und assoziierten Objekten herstellen und Forschende so beispielsweise auf potenziell relevante Daten und Metadaten und – über transparente Provenienz- und Kontextinformation – Objekte hinweisen.

Mit Blick auf die Nachnutzung kontextspezifisch erhobener Daten leisten Forschungsinfrastrukturen einen Beitrag zur Auffindbarkeit und weiteren Analyse. Für die Verarbeitung und Nachnutzung von Forschungsdaten sind dabei jedoch fast immer Schritte notwendig, um Daten aus ihrem originären Kontext in einen neuen Verwendungskontext zu überführen. Hierfür können zwei wesentliche Strategien unterschieden werden:

2.3.1 Harmonisierung von Daten

Sind Zielgruppe und Zweck der Nachnutzung von Daten bekannt, so kann ein zentrales Integrationsmodell definiert werden. In die Nachnutzung aufzunehmende Forschungsdaten werden mit diesem Modell in Beziehung gesetzt, gegebenenfalls angereichert und transformiert. Daten können je nach Verwendungskontext unter einem sehr einfachen Schema wie Dublin Core[12] vereinheitlicht werden, um wie im Fall von OAIster[13] eine möglichst große Zahl digitaler Kollektionen generisch zu integrieren. Das Integrationsmodell von Europeana[14] fokussiert die Aggregation und Bereitstellung digitaler Objekte kulturellen Erbes und verkürzt bei der Integration weitestgehend um fachspezifische Aspekte, die für die generische Auffindbarkeit und Präsentation der Objekte irrelevant sind.

2.3.2 Integration impliziten Wissens

Im Fall von DARIAH-DE ist aufgrund der Breite des Anwendungskontextes kein Integrationsmodell definierbar, das sowohl eine breit integrierende Ansicht über fachspezifische Forschungsdaten erlaubt als auch sämtliche relevanten Aspekte der Daten für potenzielle Zielgruppen und Forschungsfragen erhalten kann und dabei skalierbar und verwaltbar bleibt. Für die Konzepte und Komponenten der DARIAH-DE-Föderationsarchitektur wurde aus diesem Grund eine neuartige Methode zur fachspezifischen Beschreibung von Daten entwickelt, die es ermöglicht, den Erstellungskontext von Daten verwendungsneutral zu beschreiben und dadurch Hintergrundwissen zur originären Disziplin, Forschungsfrage und Kollektion zu explizieren. Die dadurch entstehenden, semantisch reicheren Varianten sind in Bezug auf originäre Daten und Metadaten verlustfrei und stehen für eine Assoziation mit fachspezifischen, interdisziplinären oder generischen Verwendungskontexten zur Verfügung.

Im Folgenden wird dieser Artikel nun wesentliche Ideen und Komponenten der DARIAH-DE-Föderationsinfrastruktur – und damit auch die Strategie der *Integration impliziten Wissens* – vorstellen. Für die einfachere und im Fall von a-priori definierbaren Verwendungskontexten schneller anwendbare Harmonisierung sei neben den Projekten von Europeana und OAIster auf die Literatur um das klassische Integrationskonzept globaler Schemata verwiesen.[15]

3. Forschungsinfrastrukturen

Wenngleich epidat als fachspezifische Datenbank primär der Unterstützung von Fragen im Bereich judaistischer Forschung gewidmet ist, so kann epidat als infrastrukturelle Komponente verstanden werden, welche autonom und geographisch verteilt erhobene Forschungsdaten zusammenführt. Für die Fachwissenschaft wurde mit epidat eine Plattform geschaffen, die eine Auffindbarkeit und fachspezifische Nachnutzung der Daten erleichtert.

Für die Fotografien der Grabsteine und insbesondere auch die akribisch erarbeiteten Informationen zu Symbolik, Epigrafik etc. ist jedoch auch eine weiterführende und interdisziplinäre Fortsetzung und Nachnutzung denkbar. Informationen beispielsweise zu den Verwitterungszuständen der Grabsteine und naher Bauwerke könnten sich für die kooperative Bearbeitung von Fragen judaistischer und kunsthistorischer Forschung qualifizieren.

Die Existenz geeigneter thematischer Kollektionen vorausgesetzt, treten Möglichkeiten zu einer interdisziplinären Kooperation auf Basis digitaler Forschungsdaten in der wissenschaftlichen Praxis häufig auf. Ohne direkte Förderung scheitern solche Vorhaben jedoch oft an technischen Hürden, die ohne zur Verfügung stehende Ressourcen beziehungsweise Werkzeuge nicht überwindbar sind. Aber auch geförderte Projekte stehen häufig vor den technischen Problemen des Zugriffs auf Kollektionen und der Zusammenführung von Daten, die vor einer eigentlichen inhaltlichen Bearbeitung zu lösen sind. Dieser Problematik versucht die Forschungsinfrastruktur von DARIAH-DE zu begegnen, indem insbesondere technische Aspekte des Zugriffs, der Verarbeitung und Transformation generisch gelöst werden – ohne Daten hierfür inhaltlich beeinträchtigen oder verkürzen zu müssen.

3.1 Anwendungsfälle

Eine wesentliche Zielsetzung der DARIAH-DE Föderationsarchitektur besteht in deren flexibler Anpassbarkeit an die Bedürfnisse konkreter Anwendungsfälle. In Abbildung 4 werden einige dieser Anwendungsfälle grob skizziert (Abb. 4):

- *Breitensuche*: generische Suche mit geringer semantischer Tiefe über die Inhalte einer Vielzahl heterogener Kollektionen
- *Tiefensuche*: Suche über thematisch zusammenhängende Kollektionen mit verfeinerten Möglichkeiten zur Navigation, Facettierung etc.
- *Datenintegration*: Zusammenführung heterogener Kollektionen unter einer gemeinsamen Integrationsstruktur; die strukturelle Tiefe der Integration variiert in Abhängigkeit von Forschungsfrage und thematischer Homogenität der zu integrierenden Kollektionen
- *Individuelle Analyse*: Bereitstellung eigener, lokaler Daten zur Kontextualisierung oder auch zur Anwendung von Analysemethoden, die über die DARIAH-DE Forschungsinfrastruktur zugänglich sind.

3.2 Semantische Cluster

Eine Besonderheit der DARIAH-DE-Föderationsarchitektur besteht in der Zielsetzung, eine möglichst große Zahl solcher Anwendungsfälle flexibel zu unterstützen. Dabei sollen breite und übergreifende Perspektiven ebenso modelliert werden können wie die spezifischen Verwendungskontexte aktueller Forschungsfragen. In der Praxis findet sich als typischer Startpunkt zur Modellierung und Zusammenführung von Forschungsdaten die Definition eines übergeordneten Integrationsmodells[16] – heterogene Daten sind dann mit

diesem Modell zu assoziieren. Der Erfolg komplexer Datenmodelle wie TEI zeigt, dass ein derartiger Ansatz durchaus funktionieren kann, allerdings auch bei einer Fokussierung auf bestimmte fachliche Domänen sehr schnell zu einer kaum beherrschbaren Komplexität führt, welche die Skalierbarkeit globaler Ansätze stark einschränkt.

Der Anwendungskontext von DARIAH-DE ist grundsätzlich nicht auf Wissenschaftszweige oder Disziplinen beschränkt, wodurch sich dieses Komplexitätsproblem weiter verstärkt. Als Antwort bietet DARIAH-DE einen im Wesentlichen zweistufigen Ansatz:

- Daten werden in ihrer originären Struktur und ihrem Erstellungskontext modelliert und verfeinert.
- Strukturen werden so miteinander assoziiert, wie dies für die Unterstützung von Forschungsfragen aus Sicht der Fachwissenschaftler erforderlich erscheint.

Aufgrund der fachlichen Dezentralität dieses Ansatzes entstehen Wissensinseln enger semantischer Kohärenz, die im Kontext von DARIAH-DE als semantische Cluster bezeichnet werden (Abb. 5). Solche Cluster entstehen dabei implizit im Rahmen aktiver Forschung oder in gezielten thematischen Integrationsprojekten: Fachwissenschaftler identifizieren zunächst Kollektionen, die sich als relevant für die Beantwortung einer Forschungsfrage erweisen könnten. Sofern Kollektionen mehrere Exportstrukturen unterstützen, ist die für einen fachwissenschaftlichen Kontext semantisch geeignetste Struktur festzulegen. Schließlich ist eine Assoziation der relevanten, lokalen Schemata zur Erstellung einer übergreifenden Sicht erforderlich. Dabei ist es prinzipiell unerheblich, ob die übergreifende Sicht aus der Menge der lokalen Schemata ausgewählt wird oder beispielsweise ein passender Standard oder eine Ontologie ausgewählt werden.

Abbildung 6 beschreibt die Modellierung eines Beispielclusters im Bereich biographischer Analysen (Abb. 6). Der Anwendungsfall ist motiviert durch historische Forschung am Institut für Europäische Geschichte in Mainz[17] und zielt auf die Zusammenführung biographischer Indizien aus unterschiedlichen strukturierten und unstrukturierten Quellen zur Kompilation transnationaler Mobilitätsprofile historischer Personen.[18] Insbesondere aufgrund der Notwendigkeit einer nachvollziehbaren Aufbereitung und Speicherung von Provenienz- und Konfidenzinformationen wurde eine eigene Struktur für biographische Profile entwickelt (S8 in Abb. 6), mit welcher relevante Aspekte lokaler Datenmodelle zu assoziieren sind. So wurde nach einer Beschreibung der intrinsischen Struktur biographischer Artikel in Wikipedia (S6) sowie der Definition von Regeln zur Erkennung von Entitäten und entsprechender Korrelationen ein Mapping auf das Integrationsmodell vorgenommen. Zunächst zu Testzwecken integriert, eignet sich Wikipedia aufgrund des Mengengerüsts insbesondere als Testkorpus für die Anwendung informatischer Methoden, die dann auch auf kleinere Korpora wie die Neue Deutsche Biographie (NDB, S9) angewendet werden können. Neben automatisiert gewonnenen Indizien aus unstrukturierten Artikeln können Daten aus strukturierten Quellen wie Wikidata etc. assoziiert und genutzt werden.

Auf Basis eng zusammenhängender semantischer Cluster können nun – wie auch in Abb. 5 angedeutet – durch die Assoziation der Cluster miteinander oder auch mit generi-

schen Schemata, wie zum Beispiel Dublin Core (im Beispiel etwa S10 in Abb. 5), interdisziplinäre Perspektiven erstellt werden. Aus den Assoziationen innerhalb der Cluster ergibt sich der besondere Vorteil, dass für die übergreifende Zusammenführung typischerweise nur die semantisch reichhaltigsten Strukturen beziehungsweise die gewählten Integrationsmodelle (S3, S5, S8 in Abb. 5 u. 6) zu assoziieren sind.

3.3 Komponenten

Die Konzepte der fachspezifischen Föderation von Daten und der semantischen Cluster werden im Rahmen der DARIAH-DE-Forschungsinfrastruktur mit Hilfe verschiedener Komponenten implementiert. Abbildung 7 zeigt das Zusammenspiel zwischen den Komponenten zur Bereitstellung von nutzerorientierten Diensten. Die Funktionalität zur Modellierung und Assoziation von Daten wird durch die sogenannte Föderationsschicht abgebildet (Abb. 7).[19]

3.3.1 Sammlungen

Die Collection Registry ist ein zentrales Verzeichnis zur Registrierung und Beschreibung von Sammlungen von Ressourcen. Sammlungen können selbst direkt Ressourcen oder weitere untergeordnete Teilsammlungen beinhalten und können sowohl physische als auch digitale Objekte oder nur Daten aggregieren. Die Sammlungsbeschreibungen decken neben Verschlagwortung, zeitlichen und geografischen Dimensionen auch Sammlungsformate und Informationen zur Datenpflege ab.

Abbildung 8 zeigt einen Bildschirmausschnitt des Sammlungseditors der Collection Registry (Abb. 8). Verschiedene Attribute erlauben die Beschreibung inhaltlicher Eigenschaften in Form von Freitexten, aber auch die Definition von Zugriffsmöglichkeiten, über die gegebenenfalls auch maschinell auf Daten der Kollektion zugegriffen werden kann. Neben der Modellierung des Konzepts der Teilsammlung können in der Collection Registry auch Akteure (Personen oder Organisationen) definiert und mit Kollektionen in Beziehung gesetzt werden. Die Daten in der Collection Registry selbst ergeben so ein komplexes Netzwerk aus assoziierten Kollektionen und Akteuren, welches für interessierte Benutzer, insbesondere aber auch für integrative Dienste zur Verfügung gestellt wird. Die generische Suche von DARIAH-DE holt beispielsweise in definierten zeitlichen Abständen Informationen über Kollektionen aus der Collection Registry ab und reagiert entsprechend auf Änderungen oder neue Einträge mit einem *Harvesting* angebotener Daten und deren Neuindexierung.

3.3.2 Datenmodelle

Während in der Collection Registry Metadaten auf einer aggregierenden Sammlungsebene beschrieben werden, verzeichnet das Data Modeling Environment (DME) Informationen

über die Struktur und Semantik der in Kollektionen enthaltenen Daten. Die grundlegende Zielsetzung des DME besteht dabei primär in der Definition und nachnutzbaren Modellierung des Erstellungskontexts von Daten. Ausgehend beispielsweise von einem XML-Schema wird ein Datenmodell angelegt, verfeinert und um Hintergrundwissen zum Beispiel zur Sammlung und Institution erweitert. Hierdurch wird insbesondere eine Nachnutzung von Daten außerhalb des originären Sammlungskontexts ermöglicht.

Als einfaches Beispiel für die Explikation eines Erstellungskontexts dient der Auszug eines Dublin Core-Dokuments in Abbildung 9. Das Attribut „xsi:schemaLocation" verweist auf ein XML-Schema von Dublin Core, welches im Wesentlichen die Existenz, nicht jedoch die Ausgestaltung der 15 Felder (Title, Creator, Date etc.) des Standards definiert (Abb. 9). Nach einem Import dieses XML-Schemas als Ausgangsbasis für das zu entwickelnde Datenmodell können nun Regeln zur Verfeinerung der Daten angewendet werden.

Für den einfachen Fall des dargestellten XML-Dokuments sind einige Regeln der beinhaltenden Kollektion für das menschliche Verständnis sofort erkennbar, die jedoch für eine maschinelle Verarbeitung zu spezifizieren sind:

- *Creator*: Es wird ein vollständiger Personenname nach der Regel Nachname, Vorname eingetragen. Die explizite Definition dieser Regel verbessert beispielsweise eine spätere Abfrage einer Datenbank für Personennormdaten.
- *Date*: Zeitangaben sind stark heterogen zwischen und teilweise auch innerhalb von Kollektionen. Während die Datumsangabe im Beispiel dem unkritischen Muster Jahr-Monat-Tag folgt, ist die Angabe eines maschinenverarbeitbaren Äquivalents beispielsweise zu einer Angabe „VD17"[20] unerlässlich, um spätere zeitbasierte Anfragen oder Auswertungen zu ermöglichen.
- *Subject*: Für den konkreten Beispielfall wären hier mehrere Subject-Felder mit jeweils einem Schlüsselbegriff wünschenswert. Für die vorliegende Aneinanderreihung innerhalb eines einzigen Feldes muss eine einfache Regel formuliert werden, die eine strukturell bessere Variante des Dokuments erzeugt.

Der Bildschirmausschnitt des Schemaeditors in Abbildung 10 zeigt auf der rechten Seite die visuelle Aufbereitung der Datenstruktur und deutet links das nach Anwendung passender Regeln gewonnene, angereicherte Dokument an (Abb. 10). Die in diesem Abschnitt diskutierten Felder und Verarbeitungsregeln sind dabei als einfache Beispiele zu verstehen. Ähnliche Regeln können auch für bedeutend komplexere Anwendungsfälle herangezogen werden. Für die Identifizierung und Zusammenführung biographischer Indizien wurde so beispielsweise die Erkennung, Auflösung und Disambiguierung von Entitäten in Form von Regeln im DME spezifiziert.

3.3.3 Mappings

Durch die Definition fallspezifischer Integrationsmodelle können Datenstrukturen miteinander assoziiert werden. Durch eine Formulierung von Transformationsregeln werden

Daten dabei so umgewandelt und integriert, wie sie für eine weiterführende Untersuchung benötigt werden.

Abbildung 11 zeigt einen Ausschnitt des Mappings zwischen den beiden Schemata oai_dc und OLAC-DcmiTerms (Abb. 11).[21] Die einfache Verbindung der Identifier- und MdSelfLink-Felder ist in der Ansicht farblich markiert und bildet ein Beispiel einer einfachen Wertkorrespondenz. Dies bedeutet, dass nach dieser Modellierung die Inhalte im Identifier-Feld von oai_dc als semantisch äquivalent zu Inhalten des MdSelfLink-Feldes von OLAC-DcmiTerms gelten. Über solch einfache Wertkorrespondenzen hinaus unterstützen Mappings im DME verschiedene Multiplizitäten (1 : N, N : 1, N : M). Zudem können auch für Mappings inhaltsbezogene Transformationsregeln formuliert werden. So könnte für eine Assoziation zwischen einem Creator-Feld in einem Schema A und einem Creator-FirstName-/Creator-LastName-Feldpaar in einem weiteren Schema B beispielsweise eine 1 : 2 Assoziation [Creator → Creator-FirstName, Creator-LastName] formuliert werden. Nach der Idee der Trennung von Erstellungs- und Verwendungskontext wäre in diesem konkreten Fall jedoch die Aufteilung des Creator-Feldes in Schema A zur Explikation des Erstellungskontexts und die direkte 2 : 2 Assoziation der Namenskomponenten zwischen den Schemata A und B zu bevorzugen.

4. Integrative Dienste

Entgegen der intuitiven Annahme, die Dienste der Föderationsschicht würden ausschließlich für eine Nachnutzung modellierter Konzepte in integrativen Diensten implementiert, erfüllen diese tatsächlich auch einen gewissen Selbstzweck. So wird durch die Registrierung und ausführliche Beschreibung einer Sammlung in der Collection Registry diese dokumentiert und kann durch interessierte Forscherinnen und Forscher einfacher aufgefunden werden. Die Modellierung von Daten bedingt dagegen eine intensive Auseinandersetzung mit unterschiedlichen Forschungskontexten und den Daten selbst, wodurch auch durch das DME eine Art Erschließung und Dokumentation digitaler Sammlungen erreicht wird. Die primäre Zielsetzung der Förderationsschicht besteht allerdings tatsächlich in der Bereitstellung von Daten und Funktionalität, um heterogene Daten in nutzerorientierten Diensten zusammenzuführen.

Neben verschiedenen assoziierten Projekten und Fallstudien, die auf der DARIAH-DE-Föderationsschicht aufbauen, implementiert DARIAH-DE selbst im Wesentlichen drei integrative Dienste, die im Folgenden kurz vorgestellt werden.

4.1 Generische Suche

Mit der generischen Suche wird im Rahmen von DARIAH-DE ein konkreter Anwendungsfall der Datenföderation umgesetzt. Hierbei werden Daten aus den in der Collection Registry verzeichneten Kollektionen nach den im DME definierten Datenmodellen verarbeitet und indexiert. Die Heterogenität der Ressourcen wird zum Zeitpunkt konkreter Suchanfragen, basierend auf der zu durchsuchenden Menge von Kollektionen, mit Hilfe

der im DME verfügbaren Mappings aufgelöst (Abb. 12). Dies bedeutet, dass logische Suchanfragen durch Anwender in einem beliebigen Schema formuliert werden können. Die auf den indexierten Daten tatsächlich ausgeführten Anfragen können dann anhand der in den Mappings formulierten Assoziationen und Regeln transformiert werden – insofern ein Mapping vom Anfrageschema auf das jeweils indexierte Schema vorliegt. Der besondere Charme dieses Ansatzes besteht nun darin, dass mit zunehmender Bildung semantischer Cluster und deren interdisziplinärer Vernetzung auch die Kohäsion der Daten in der generischen Suche steigt. Während also im Rahmen von Forschungsprojekten Daten und Metadaten so modelliert werden, dass eine Harmonisierung heterogener Daten nach den Anforderungen der Forschungsfrage erreicht werden kann, so steht dieses explizierte Wissen unmittelbar auch im Rahmen der generischen Suche zur Verfügung – ohne dass hierfür ein zusätzlicher Entwicklungsaufwand notwendig wäre.

4.2 Branded Searches

Über die Möglichkeit der einfachen Suche über die Daten verzeichneter Kollektionen hinaus bietet die generische Suche von DARIAH-DE Möglichkeiten zur Definition eigener Suchmaschinen, wie in Abbildung 13 angedeutet (Abb. 13).[22] Sämtliche Schritte, die für die Umsetzung einer solchen Suchmaschine notwendig sind, können ohne technischen Aufwand durch die alleinige Modellierung von Daten erreicht werden. Im Rahmen der generischen Suche stehen Möglichkeiten zur Filterung relevanter Kollektionen und Auswahl geeigneter Integrationsmodelle zur Verfügung. Eine Branded Search wird schließlich durch das Speichern dieser Festlegungen, einer visuellen Anpassung sowie der Auswahl eines gewünschten URL-Präfixes umgesetzt.

4.3 Cosmotool

Durch das Konzept der Branded Search können Suchprototypen auf Basis der generischen Suche einfach hergestellt werden. Über die Festlegung von Integrations- beziehungsweise Anfragemodellen und einigen visuellen Anpassungen hinaus sind die eingestellten Suchvarianten jedoch auf die Funktionalität der generischen Suche beschränkt.

Aber auch für die Entwicklung fachspezifischer Dienste kann auf die Funktionalität der Föderationsschicht zurückgegriffen werden, um von technischer (und bei gegebenen Schemata und Mappings von schematischer) Heterogenität von Daten abstrahieren zu können. Die Entwicklung der Anwendung muss sich demnach nicht mit unterschiedlichen Formaten, Zugriffsstrukturen etc. auseinandersetzen, sondern kann sich auf nutzerorientierte Aspekte konzentrieren.

Im Rahmen des DARIAH-DE Use-Case Biographien wird ein Prototyp entwickelt, der biographische Indizien aus unterschiedlichen Quellen zusammenführt und für Suchanfragen bereitgestellt (Abb. 14). Das so entwickelte Werkzeug kann dabei als logische Konsequenz einer Spezialisierung der generischen Suche interpretiert werden:

- die Sammlung von Datenquellen erfolgt in der DARIAH-DE Collection Registry,
- die Modellierung der Daten sowie deren Assoziation mit einem zentralen biographischen Schema erfolgt im DARIAH-DE DME,
- die Verarbeitung und Indexierung der Daten basiert auf funktionalen Komponenten der generischen Suche,
- die Analyse und Visualisierung wurde und wird dagegen spezifisch für den Anwendungsfall entwickelt und bildet den tatsächlichen Kern des CosmoTools.

Die integrierten biographischen Profile werden dabei anhand unterschiedlicher Visualisierungen geotemporal und inhaltlich aufgearbeitet. Die Darstellung eines biographischen Profils erfolgt in der aktuellen Version des Prototyps mit Hilfe von drei visuellen und interaktiven Komponenten (Abb. 15):

- Die Zeitleiste ordnet bekannte Ereignisse chronologisch entlang einer Zeitachse an und integriert dabei erste Indikatoren der Konfidenz und Provenienz. Die Einfärbung weist auf die Herkunft des Eintrags hin (grün für strukturierte Quelle, gelb für Extrakt aus unstrukturiertem Text). Die Distanz des Eintrags von der Zeitleiste deutet die Konfidenz eines Eintrags an: Direkte Ereignisse im Leben einer Person wie Geburt, Studium und Tod sind demnach als gesicherter (näher an der Linie) dargestellt als beispielsweise die Geburt eines Kindes. Kann dieses Ereignis als sicheres Indiz für den Aufenthaltsort der Mutter verwertet werden, so birgt selbiges Ereignis im Lebenslauf des Vaters eine zu berücksichtigende Unsicherheit.
- Im Bereich Ereignis-Details werden zur Verfügung stehende Informationen über die Fundstelle eines Ereignisses angezeigt. Das dargestellte Beispiel in Abbildung 15 weist auf eine identifizierte Korrelation von Person, Ort und Zeit in einem unstrukturierten Text hin. Anwender können anhand des relevanten Auszugs direkt prüfen, ob eine Korrelation richtig erkannt wurde, oder aber dem Verweis zu der relevanten Fundstelle folgen.
- Der Fokus des Anwendungskontextes des Prototyps liegt auf der Erkennung und Darstellung von Transnationalität und Mobilität in Biographien, weshalb die Kartendarstellung als dritte Visualisierungsform gewählt wurde. Diejenigen Ereignisse, zu denen ein Ort identifiziert und geographisch aufgelöst werden kann, werden in der Karte dargestellt und durch Pfade verbunden, wodurch Mobilität für den Anwender des Prototyps unmittelbar erkennbar ist.

Abbildung 16 zeigt eine weitere Visualisierung des Prototyps, mit dessen Hilfe insbesondere die Erkennung von Entitäten und Zusammenhängen aus unstrukturierten Daten nachvollzogen werden kann. Hierzu werden die zur Anwendung automatischer Verfahren verwendeten Texte gemeinsam mit den darin erkannten Ereignissen angezeigt. In den inhaltlich unveränderten Texten werden Bestandteile hervorgehoben, die zu der Extraktion eines biographischen Ereignisses geführt haben, sobald der Anwender mit dem Mauszeiger über das Ereignis navigiert (Abb. 16).

Bei der Umsetzung von CosmoTool besteht ein wesentliches Ziel in der Anwendung einer Kombination aus qualitativen und quantitativen Methoden zur Erschließung einer großen Menge verfügbarer biographischer Quellen. Eine weitere Zielsetzung besteht insbesondere auch darin, die inhärente Unsicherheit automatischer Methoden durch die Hervorhebung von Informationen über angewendete Methoden und die Provenienz von Daten transparent zu gestalten. Durch Interaktion mit den Anwendern können sowohl die biographischen Profile als auch die zu deren Zusammenstellung angewendeten Methoden positiv beeinflusst werden.

5. Ausblick

Nur auf den ersten Blick scheinen sich die Vielfältigkeit der kultur- und geisteswissenschaftlichen Forschungslandschaft und die Zielsetzungen generischer Forschungsinfrastrukturen zu widersprechen. Für viele Forschungsfragen kann die Nutzung entsprechender Infrastrukturen die Arbeit durch den strukturierten Zugang zu digitalen Quellen auf eine breitere Basis stellen. Hinzu kommen die Möglichkeiten zur Nutzung entsprechender von den Infrastrukturen angebotener Werkzeuge[23] sowie die einfache Möglichkeit, Forschungsdaten für die Nachnutzung bereit zu stellen. Dazu werden zum Beispiel in der DARIAH-DE-Föderationsarchitektur Konzepte und Komponenten entwickelt, die auf eine inhaltliche Forschungsunterstützung abzielen, indem technische Aspekte des Zugriffs, der Verarbeitung und der Transformation von Daten generisch gelöst werden, um den Fokus der Forschung auf diejenigen semantischen Fragestellungen zu lenken, für die fachwissenschaftliches Wissen erforderlich ist.

Neben diese direkte Unterstützung von Forschungsprojekten durch entsprechende Dienste treten weitere Leistungen der Forschungsinfrastrukturen auf mehreren Ebenen – seien es die Bemühungen zur curricularen Entwicklung der Digital Humanities, um wissenschaftliches Personal an der Schnittstelle zwischen den Geisteswissenschaften und der Informatik besser ausbilden zu können, oder aber sei es die Schärfung des Bewusstseins unter den Forscherinnen und Forschern, dass ihre Daten für eine potenzielle Nachnutzung relevant sein könnten und wie sie diese begünstigen könnten.