Andreas Henrich, Tobias Gradl, Robin Jegan

SUCHWERKZEUGE FÜR SAMMLUNGEN

HERAUSFORDERUNGEN, TRENDS UND STRATEGIEN

1. Einleitung

Sammlungen stellen wesentliche Ressourcen für die Forschung dar. Dabei kommt im Zuge der Digitalisierung dem Forschungsdatenmanagement eine wichtige Rolle zu. Ein Konsortium von Wissenschaftlerinnen und Wissenschaftlern und Organisationen hat diesbezüglich im März 2016 die »FAIR Guiding Principles for scientific data management and stewardship«[1] formuliert und veröffentlicht. Wissenschaftliche Daten sollen demnach auffindbar (Findable), zugänglich (Accessible), interoperabel (Interoperable) und wiederverwendbar (Re-usable) sein. Suchwerkzeuge leisten in diesem Kontext einen wesentlichen Beitrag zur *Auffindbarkeit*.

Um die Rolle der Suchwerkzeuge und die von ihnen zu erfüllenden Aufgaben besser zu verstehen, ist dabei zunächst ein Blick auf die Suche an sich notwendig. Offensichtlich gibt es viele verschiedene Arten von Informationsbedürfnissen und damit auch Arten der Suche. Marchionini unterscheidet hier »Lookup«, »Learn« und »Investigate«, wobei »Learn« und »Investigate« unter dem Begriff explorative Suche zusammengefasst werden können.[2] Während es beim Lookup eher um die Suche nach einfachen Fakten oder das Auffinden von Daten bzw. Dokumenten geht, deren Existenz man bereits kennt, dient die explorative Suche dem Wissenserwerb, der Interpretation, dem Vergleich oder auch der Wissensaggregation, der Analyse, dem Ausschluss sowie der Bewertung. Je nach Art der Suche stehen dabei im Hinblick auf eine Suchlösung die breite Abdeckung möglichst vieler Ressourcen oder die tiefe Adressierbarkeit von Ressourcen, die z. B. nach fachlichen Kriterien vorausgewählt wurden, im Vordergrund.

Der Forschungsverbund Marbach Weimar Wolfenbüttel (MWW), der in diesem Artikel als Beispiel betrachtet werden soll, umfasst Dutzende digitale

1 Mark D. Wilkinson et al.: The FAIR Guiding Principles for scientific data management and stewardship, in: Scientific Data 3 (2016), doi:10.1038/sdata.2016.18.

2 Gary Marchionini: Exploratory search: from finding to understanding, in: Communications of the ACM 49,4 (2006), S. 41-46, doi: https://doi.org/10.1145/1121949.1121979.

DOI https://doi.org/10.15499/kds-003-004 | CC BY-SA 4.0

Sammlungen, die Zugang zu europäischer und deutscher Literatur ermöglichen.[3] Die große Menge und Heterogenität der Daten in diesen Sammlungen führt jedoch zu Problemen und Einschränkungen, sobald sie mittels einer einzigen Anfrage in der Breite durchsucht werden sollen. Im vorliegenden Artikel werden wir verschiedene Lösungsansätze für diese häufig anzutreffende Konstellation skizzieren und vergleichen. Diese Ansätze unterscheiden sich im Grad der Integration der Daten. Um die Betrachtung zu konkretisieren, beschreiben wir eine von uns entwickelte Suchlösung – die MWW-Verbundsuche für den Forschungsverbund Marbach Weimar Wolfenbüttel – als föderiertes System mit einer »dezentralen Integration von Daten«.[4] Die Möglichkeiten, die durch einen föderierten Ansatz geschaffen werden, eröffnen hierbei neue Anwendungsszenarien über Sammlungsgrenzen hinweg.

Der Artikel ist dazu wie folgt gegliedert: Das zweite Kapitel ordnet einleitend die Suche in Sammlungen im Kontext der Digital Humanities ein. Das dritte Kapitel befasst sich mit der Beschreibung und Charakterisierung von verschiedenen Konzepten bzw. Architekturen für Suchlösungen. Im vierten Kapitel wird die generische Suche aus DARIAH-DE beschrieben, deren Anwendung im Kontext des Forschungsverbundes MWW Gegenstand von Kapitel fünf ist. Kapitel sechs geht zum Abschluss kurz auf weitere Nutzungsmöglichkeiten von generischen Suchlösungen ein.

2. Einordnung der Suche in Sammlungen

Digitale Sammlungen, das heißt Forschungsdaten im geisteswissenschaftlichen Rahmen, rücken in jüngerer Vergangenheit im Kontext der Digital Humanities verstärkt in den Vordergrund. Dadurch wächst auch der Bedarf an unterschiedlich ausgeprägten Suchmöglichkeiten über die heterogenen Daten in solchen Sammlungen. Die Sammlungen sind jedoch geprägt durch Heterogenität auf mehreren Ebenen, nicht nur auf syntaktischer, sondern auch auf struktureller und semantischer Ebene.[5]

3 https://vfr.mww-forschung.de/die-sammlungsuebersicht [zuletzt 14.12.2022].

4 Tobias Gradl, Andreas Henrich und Christoph Plutte: Heterogene Daten in den Digital Humanities. Eine Architektur zur forschungsorientierten Föderation von Kollektionen, in: Grenzen und Möglichkeiten der Digital Humanities: Sonderband der Zeitschrift für digitale Geisteswissenschaften 1 (2015), http://zfdg.de/sb001_020 [zuletzt 8.5.2019].

5 Amit P. Sheth und Vipul Kashyap: So far (schematically) yet so near (semantically), in: Proceedings of the IFIP WG 2.6 Database Semantics Conference on Interoperable Database Systems (DS-5) (1993), S. 283-312.

Die syntaktischen Unterschiede beziehen sich auf technische Eigenschaften der Daten, wie das Dateiformat oder die Zugriffsmechanismen der Sammlungen. Auf struktureller Ebene werden die schematischen Eigenschaften der Sammlungen behandelt, das heißt in welchen Metadatenformaten und Formaten Informationen zu den Sammlungen, beziehungsweise die Daten in den Sammlungen selbst, vorliegen. Semantische Unterschiede beziehungsweise die Verknüpfungen zwischen Daten und Sammlungen können oftmals nur mithilfe von Hintergrundwissen über die jeweiligen Sammlungen abgebildet werden.[6] Weiterhin ist die genaue Eingliederung von Sammlungen in einen größeren Suchkontext nur durch die Befragung von Fachexperten und deren detailliertes Wissen über die Eigenschaften, Inhalte, und den Aufbau der Sammlung möglich.

Die Zusammenführung heterogener Sammlungen ist im Kontext von Digital Libraries kein neuartiges Phänomen.[7] Jedoch stellen die konkrete Umsetzung und die Bereitstellung von Interoperabilität für diese Daten auch heute noch eine Herausforderung dar. Von besonderer Bedeutung ist dabei die Verwendung von etablierten Metadatenstandards wie etwa dem Dublin Core Metadata Element Set (DC)[8] mit 15 zentralen Elementen oder der Text Encoding Initiative (TEI)[9] für die Auszeichnung von Textdaten. Diese Standards und die darin verwendeten Schemata ermöglichen eine effizientere Verarbeitung, Analyse und Suche der Daten.

Aufbauend auf diesen Überlegungen ist nun die Suche in Sammlungen einzuordnen. Hierbei können die Inhalte der Sammlungen selbst sehr unterschiedlich sein – von Textdokumenten, über Bilder bis hin zu Videos oder Audiodateien. Während man bei (Text-)Dokumenten sicher auch Interesse an einer Suche in den Volltexten hat, sind Ansätze zur inhaltsbasierten Suche in anderen Medientypen bisher auf Nischen beschränkt.[10] Daher kommt der Suche in Metadaten eine wichtige Rolle zu. Eine übergreifende Suche

6 Tobias Gradl und Andreas Henrich: A novel approach for a reusable federation of research data within the arts and humanities, in: Digital Humanities 2014: conference abstracts EPFL-UNIL (2014), S. 382-384.

7 Howard Besser: The Next Stage: Moving from Isolated Digital Collections to Interoperable Digital Libraries, in: First Monday 7,6 (2002), https://doi.org/10.5210/fm.v7i6.958 [zuletzt 14.12.2022].

8 http://www.dublincore.org/specifications/dublin-core/dces/ [zuletzt 4.9.2019].

9 https://tei-c.org/ [zuletzt 4.9.2019].

10 Bernhard Bermeitinger, Simon Doing, Maria Christoforaki, André Freitas und Siegfried Handschuh: Object Classification in Images of Neoclassical Artifacts Using Deep Learning, in: Diane Jakacki u.a. (Hrsg.): Digital Humanities 2017, Montréal, Canada, S. 395-397.

über alle Felder der Metadaten mit unstrukturierten Anfragen ist dabei für eine erste Orientierung im Datenbestand sicher hilfreich. Dies bleibt aber zwangsweise recht grob, denn in welchem Metadatenfeld ein Suchbegriff wie »Goethe« steht, kann kaum vernachlässigt werden. In der Konsequenz ist eine Suche auf einzelnen Feldern des Metadatenschemas wünschenswert. Daneben können einzelne Elemente der Metadatenschemata auch als Filter im Rahmen einer facettierten Suche hilfreich sein. Beispiele wären die Dublin Core Felder »date«, »language« oder »format«. Probleme ergeben sich dabei natürlich aus der Heterogenität der Metadatenschemata. Diese müssen zusammengeführt werden, ohne die Spezifika einzelner Sammlungen zu vernachlässigen. Damit bleibt als Zielvorstellung eine Suche, die sowohl übergreifende als auch spezifische Suchen in den Metadaten unterstützt und dabei die gezielte Nutzung einzelner Elemente für die Suche ebenso erlaubt wie für die facettierte Filterung.

3. Mögliche Architekturen und Konzepte

Für die grundsätzliche Situation einer Suche über verteilte Sammlungsbestände sind unterschiedliche Suchkonzepte möglich. Die Ausgangsbasis stellt sich dabei wie folgt dar: Die einzelnen Sammlungen werden in der Regel in eigenen Systemen (im Weiteren als dezentrale Systeme bezeichnet) verwaltet, die ihrerseits über Such- und Export-Schnittstellen verfügen. Über die Suchschnittstelle kann typischerweise unter Verwendung einer mehr oder weniger mächtigen Anfragesprache in den Daten gesucht werden. In den meisten Fällen existieren verschiedene Schnittstellen, die entweder über eine grafische Benutzerschnittstelle (Graphical User Interface; GUI) direkt von Nutzerinnen und Nutzern bedient werden oder von anderen Systemen über eine technische Schnittstelle (Application Programming Interface; API) angesprochen werden können. Neben Suchschnittstellen bieten die Systeme zur Verwaltung von Sammlungen, oder allgemeiner formuliert von Daten, normalerweise auch eine Exportschnittstelle an. Über diese Exportschnittstelle können die Sammlungsdaten von anderen Systemen, die über entsprechende Berechtigungen verfügen, exportiert werden. Ein typisches Beispiel für eine solche Schnittstelle ist das Protocol for Metadata Harvesting der Open Archives Initiative (OAI-PMH).[11] Zusammenfassend stellt sich damit die typische Basiskonstellation für eine übergreifende Suche so dar, dass die einzelnen Sammlungen in eigenen *dezentralen* Systemen verwaltet werden, die im Idealfall

11 https://www.openarchives.org/pmh/ [zuletzt 4.9.2019].

jeweils über eine Suchschnittstelle und eine Exportschnittstelle verfügen. Auf dieser Ausgangsbasis können nun sehr unterschiedliche Ansätze zur Umsetzung einer Suche verfolgt werden, die sammlungsübergreifende Suchen erfordern. Vier idealtypische Ansätze werden im Folgenden besprochen.

3.1 Direkte Suche

Bei der direkten Suche wird den Nutzerinnen und Nutzern schlicht keine integrierte Suchlösung über mehrere Sammlungen oder dezentrale Systeme angeboten. Vielmehr müssen die Nutzerinnen und Nutzer für eine Suche, die über mehr als eine Sammlung hinausgeht und Daten aus mehreren Sammlungen adressieren soll, selbst alle potenziell relevanten Sammlungen identifizieren und Anfragen über deren jeweilige Suchschnittstellen durchführen. Die mehrfache Abfrage in unterschiedlichen Systemen ist weder nutzerfreundlich noch effektiv. Zusätzlich können sich die jeweiligen Systeme in ihrer Bedienung und Benutzerführung stark unterscheiden und beeinträchtigen somit die Effizienz der Suche. Ein weiteres Problem ergibt sich dadurch, dass eine Integration oder vergleichende Betrachtung der Ergebnisse der verschiedenen Systeme ebenfalls durch die Nutzerin bzw. den Nutzer erfolgen muss.

Trotz dieser Nachteile hat der Verzicht auf eine übergreifende Suchlösung auch Vorteile. So entfallen zunächst die Entwicklungs- und Wartungsaufwände für eine übergreifende Suchlösung. Ferner kann man auch ins Feld führen, dass die direkte Nutzung der jeweiligen Systeme gewährleistet, dass die Daten in ihrer ursprünglichen Form und in ihrem ursprünglichen Kontext wahrgenommen werden können. Übergreifende Suchlösungen müssen hier darauf achten, dass die Daten möglichst unverfälscht bleiben und ihr ursprünglicher (Entstehungs-)Kontext problemlos nachverfolgbar bleibt.

3.2 Metasuche

Bei einer Metasuche werden die Datenbestände mehrerer dezentraler Systeme durch die Anbindung an eine zentrale Suchschnittstelle zugänglich gemacht.[12]

12 Eine Metasuche ist eine Suche, bei der eine Suchanfrage an andere Suchmaschinen weitergeleitet wird. »Meta« bezieht sich hier also auf eine Suche über (oder mithilfe anderer) Suchen. Auf welchen Daten dabei gesucht wird (Daten oder Metadaten), spielt aus Sicht der Suchmaschinenterminologie keine Rolle. Eine Metasuche ist so natürlich auch in Metadaten möglich, aber ebenso in Daten selbst.

Die Daten liegen hier weiterhin nur in den ursprünglichen Systemen. Die zentrale Suchlösung nimmt die übergreifenden Anfragen der Nutzerinnen und Nutzer entgegen und übersetzt diese so, dass sie von den jeweiligen Suchschnittstellen der dezentralen Systeme verstanden werden. Die übersetzten Suchanfragen werden dann über die Suchschnittstellen (APIs) an die jeweiligen Systeme gesendet. Die Systeme bearbeiten die Anfrage auf ihren lokalen Datenbeständen und senden die Ergebnisse zurück. Die zentrale Suchlösung versucht nun in der Ergebnisdarstellung einen Überblick über die Treffer der einzelnen Systeme zu geben und der Benutzerin bzw. dem Benutzer so die übergreifende Recherche zu erleichtern.

Der größte Vorteil dieser Lösung ist, dass Nutzerinnen und Nutzer nur eine Anfrage stellen müssen, die automatisch für die verschiedenen dezentralen Systeme übersetzt und an diese weitergeleitet wird. Da die Suche genau zum Zeitpunkt der Anfragestellung an die dezentralen Systeme weitergeleitet wird, wird sie dort immer auf den aktuellen Daten ausgeführt. Probleme mit einem veralteten Index einer zentralen Suchlösung (siehe Abschnitt 3.3) können so nicht entstehen. Ein anderer Aspekt, der bei dieser Lösung relativ gut zu handhaben ist, liegt in gegebenenfalls komplexen Zugriffsrechten für die einzelnen dezentralen Systeme bzw. Sammlungen. Die Nutzerin bzw. der Nutzer kann sich gegenüber der zentralen Suchlösung identifizieren. Diese Suchlösung kann die Anfrage dann mit den jeweiligen persönlichen Zugriffsrechten an die verteilten Systeme weiterleiten, sodass die Anfrage mit den korrekten Zugriffsrechten auf den dezentralen Systemen ausgeführt wird. Ein weiterer Vorteil einer Metasuchlösung – der eng mit der hohen Aktualität verknüpft ist – liegt in der Tatsache, dass die Suchlösung praktisch keine Daten halten muss. Dadurch kann eine solche Lösung – zumindest was den Speicherplatzbedarf betrifft – problemlos zentral betrieben werden.

Ein Nachteil von Meta-Suchmaschinen ist, dass die Anfrage für jedes dezentrale System in dessen entsprechende Anfragelogik übersetzt werden muss. Dies kann zum einen durchaus aufwendig werden, zum anderen führt es aber auch dazu, dass die zentrale Suchlösung in gewisser Weise nur die Funktionalität der schwächsten dezentralen Suchlösung umsetzen kann. Ein anderes Problem ist das Mischen der Resultate der einzelnen dezentralen Systeme. Da der zentralen Suchlösung weitere Informationen zu den Regeln, die den einzelnen Suchergebnissen zugrunde liegen, fehlen, ist es in der Regel nicht mit vertretbarem Aufwand möglich eine einheitliche integrierende Sicht auf die Suchergebnisse zu erstellen. Deshalb wird häufig auf eine simple Trefferanzeige über die einzelnen dezentralen Systeme mit separaten Ergebnislisten zurückgegriffen. Schließlich setzt die Metasuche natürlich voraus, dass die einzelnen dezentralen Systeme eigene Suchschnittstellen als API anbieten.

3.3 Gathering

Eine Suchlösung auf Basis des Gathering-Konzepts nutzt, im Gegensatz zur Metasuche, nicht die Suchschnittstellen anderer dezentraler Systeme, sondern deren Exportschnittstellen. Über diese Exportschnittstellen werden die Datenbestände der dezentralen Systeme abgerufen und in einem zentralen Index für die Suche aufbereitet und verwaltet. Auf diese Weise kann der zentrale Index eine leistungsfähige, systemübergreifende Suche realisieren, die nicht auf die Schnittmenge aller Suchfeatures der indexierten dezentralen Systeme beschränkt ist.

Damit ist auch bereits der primäre Vorteil einer solchen Lösung genannt. Der Nutzerin bzw. dem Nutzer kann – analog zu Web-Suchmaschinen – eine leistungsfähige Suchfunktionalität zentral bereitgestellt werden, die einerseits eine vereinheitlichte Suchlogik und andererseits insbesondere auch eine integrierte Gesamtsicht auf die Ergebnisse erlaubt. Ferner ist die Suchlösung nicht auf die Leistungsfähigkeit der dezentralen Systeme angewiesen. Selbst wenn diese keine Suchschnittstelle bereitstellen, entstehen keine Probleme, solange auf die Daten zugegriffen werden kann. Dazu reicht allerdings auch eine gegebenenfalls sehr einfache Exportschnittstelle – z.B. über Dateien – aus. Schließlich kann die Suche über den zentralen Index sehr effizient (d.h. mit schnellen Antwortzeiten für die Nutzerinnen und Nutzer) realisiert werden. Dies stellt gegenüber der Metasuche einen deutlichen Vorteil dar, weil bei der Metasuche auf die Antworten der dezentralen Systeme gewartet werden muss bevor eine Antwort präsentiert werden kann.[13]

Diesen Vorteilen stehen allerdings mehrere Nachteile gegenüber. Das zentrale Argument gegen eine Gathering-Lösung ist hierbei der massive Speicherplatzbedarf der Suchlösung, da durch die Verwaltung eines eigenen Suchindex, über alle Daten der indexierten dezentralen Systeme (Redundanz), viel Speicher erforderlich ist, der linear zur Größe des gesamten Datenbestandes anwächst. Ein weiteres Problem bilden Aktualisierungen in den dezentralen Systemen. Diese werden in der Regel nicht an die Suchlösung gemeldet, sodass keine Synchronisierung des Suchindex in Echtzeit erfolgt. Stattdessen wird üblicherweise mit regelmäßigen Aktualisierungen über die Exportschnittstellen gearbeitet. Der Rhythmus kann dabei variieren. Zu berücksichtigen ist, dass eine hohe Aktualität mit häufigem Nutzen der Exportschnittstellen und

13 Aus diesem Grund zeigen Meta-Suchmaschinen oft mit dem Eintreffen der ersten Antworten der eingebundenen Systeme vorläufige Ergebnisse an, die anschließend immer weiter komplettiert werden. Dadurch kann die Nutzerin bzw. der Nutzer relativ früh erste Ergebnisse betrachten, die sukzessive Vervollständigung kann jedoch irritieren.

einem erneuten Aufbau oder einer Aktualisierung des zentralen Suchindex verbunden ist. Daher werden in der Regel wöchentliche oder bei hohem Aktualisierungsbedarf auch tägliche neue Indexierungen infrage kommen. Auch die Problematik der Zugriffsrechte ist für Gathering-Ansätze von Bedeutung. Da der Zugriff auf die Daten über die Exportschnittstelle erfolgt, ist zu diesem Zeitpunkt unklar, welche Nutzerinnen und Nutzer später über die Suche auf die Daten zugreifen werden. Es ist daher üblich, den Export umfassend zu gestalten und dann in der Suche gegebenenfalls bereits auf die Existenz der Daten zu verweisen. Vor dem Zugriff auf die eigentlichen Daten (oder alternativ eben auch bereits vor der Anzeige in der Ergebnisliste) müssen dann aber nochmals die Zugriffsrechte der Nutzerin oder des Nutzers geprüft werden. Auch sind Gathering-Ansätze darauf angewiesen, dass die Exportschnittstellen der dezentralen Systeme vollständige, gut aufbereitete Daten zur Verfügung stellen. Häufig muss dabei für die Suchlösung eine Schemaintegration stattfinden, da der zentrale Suchindex Annahmen zu einem übergreifenden vereinheitlichten Schema machen muss. Gerade zu dieser Problematik wird später im vierten Kapitel ein Ansatz beschrieben, der es erlaubt, sehr flexibel vorzugehen.

3.4 Hybride Lösungen

Sowohl die Metasuche als auch ein Ansatz mithilfe von Gathering weisen Probleme und Einschränkungen auf, die der jeweils andere Ansatz umgeht oder unterschiedlich löst. Es liegt daher nahe, die beiden Ansätze zu kombinieren. Aufgrund der Vorteile der Metasuche im Hinblick auf den geringen Speicherplatzbedarf, die Aktualität der Daten, in denen gesucht wird, und die Reduktion der Redundanz sollte die Metasuche genutzt werden, sofern dies möglich ist, weil die dezentralen Systeme entsprechend leistungsfähige Suchfunktionen zur Verfügung stellen. Daneben wird bei einer hybriden Lösung der Ansatz des Gathering für die dezentralen Systeme eingesetzt, die keine oder eine unzureichende Suchfunktionalität anbieten. Auch Systeme, die von zu häufigen Anfragen über die Metasuche überlastet würden, können per Gathering angebunden werden. Für diese Systeme wird ein ergänzender zentraler Index aufgebaut, der über die Exportschnittstelle der dezentralen Systeme gefüllt und in regelmäßigen Abständen aktualisiert wird.

Eine solche Lösung wird besonders attraktiv, wenn die in der Metasuche angebundenen Systeme eine bereits für die zentrale Suchlösung optimierte Suchschnittstelle anbieten. Bei einer Suchanfrage an die zentrale Suchlösung werden dann die entsprechend angebundenen Bestände im Sinne der Metasuche über deren Suchschnittstellen abgefragt und parallel dazu auch der

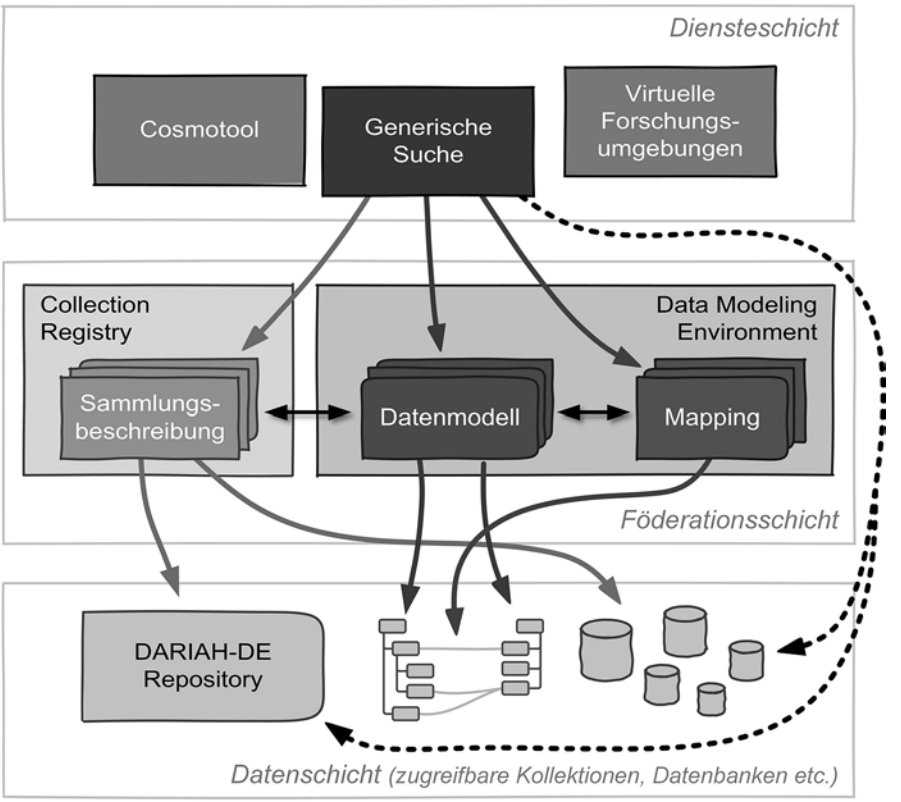

Abb. 1: Kernkomponenten der DARIAH-DE Föderationsarchitektur

im Gathering-Verfahren erstellte zentrale Index für die übrigen dezentralen Systeme angesprochen. Durch einen solchen Ansatz wird die Redundanz im System deutlich reduziert, da für die großen und leistungsfähigen dezentralen Systeme kein zentraler Suchindex erforderlich ist. Auf der anderen Seite wird eine hohe Abdeckung gewährleistet, da weniger leistungsfähige Systeme über den Gathering-Ansatz mit angebunden werden können.

Natürlich verbinden sich durch einen solchen hybriden Ansatz letztlich auch die Nachteile der beiden kombinierten Ansätze. Allerdings kann dies hier gut gesteuert werden, da für die einzelnen Systeme bzw. Sammlungen jeweils entschieden werden kann, ob sie per Gathering oder im Rahmen der Metasuche angebunden werden.

4. Eine übergreifende Suchlösung für föderierte Sammlungen

Um die in Kapitel drei vorgestellten Ansätze nun konkreter fassbar zu machen, soll als Beispiel die in den DARIAH-DE-Projekten entwickelte Suchlösung vorgestellt werden, die auch der Verbundsuche im Forschungsverbund Marbach Weimar Wolfenbüttel (MWW) zugrunde liegt. Dabei handelt es sich aktuell um einen Gathering-Ansatz.[14] Im Folgenden soll nun zunächst das Konzept der Suchlösung dargelegt werden, welches auf einer föderierten Architektur basiert und auf dieser Grundlage eine generische Suchfunktionalität anbietet. Im anschließenden Kapitel fünf wird dann die Nutzung dieses Ansatzes im MWW-Kontext adressiert.

4.1 Die DARIAH-DE Föderationsarchitektur

Sammlungen in den Geisteswissenschaften und auch die Metadaten zu diesen Sammlungen nutzen häufig sehr unterschiedliche Formate. Das Ziel eines föderierten Ansatzes ist es in diesem Kontext, nicht nur Sammlungen in bestimmten standardisierten Formaten zu unterstützen, sondern auch Sammlungen, die in speziellen Formaten vorliegen. Die Heterogenität der Formate ist dabei einerseits durch Altbestände begründet, aber andererseits auch durch die Spezifika der Forschungsprojekte in denen die Sammlungen entstanden sind bzw. entstehen. Die DARIAH-DE Föderationsarchitektur (siehe Abb. 1) wurde dafür konzipiert, heterogene Sammlungen in einer Föderation zu verwalten und für Dienste – wie z.B. übergreifende Suchlösungen – zugänglich zu machen. Dabei werden »sowohl domänenübergreifende als auch forschungsspezifische Sichten«[15] unterstützt.

Um eine Sammlung in die Föderation zu integrieren, muss zunächst ein entsprechender Eintrag in der Collection Registry vorgenommen werden. Hier sind neben administrativen Daten, wie dem Namen der Kollektion, insbesondere auch Angaben zur technischen Zugreifbarkeit erforderlich. Konkret geht es unter anderem um die Registrierung der Exportschnittstelle der entsprechenden Sammlung, um so im Gathering-Ansatz auf die Daten

14 Die Weiterentwicklung zu einer hybriden Lösung ist in Planung. Dabei soll es insbesondere ermöglicht werden, lokale Suchlösungen, die ihrerseits auf der DARIAH-DE Suchlösung basieren, zu einem größeren Verbund zusammenzuschließen, der nach dem Prinzip der Metasuche arbeitet.

15 Tobias Gradl und Andreas Henrich: Die DARIAH-DE-Föderationsarchitektur – Datenintegration im Spannungsfeld forschungsspezifischer und domänenübergreifender Anforderungen, in: Bibliothek, Forschung und Praxis 40,2 (2016), S. 222-228.

zugreifen zu können. Zu diesem Zweck werden verschiedene Schnittstellen vom OAI-PMH Protokoll bis zur Bereitstellung einer XML-Datei auf einem Webserver unterstützt.[16] Die Collection Registry basiert dazu auf dem Dublin Core Collection Application Profile.[17]

Für die Nutzung der Sammlungen durch Dienste in der Diensteschicht (siehe Abb. 1) ist neben der Angabe der technischen Zugriffspunkte die Definition der verwendeten Schemata von hoher Bedeutung. Diese Schemata können im Data Modeling Environment ausgewählt bzw. modelliert werden. Dabei erlauben leistungsfähige Spezifikationsmechanismen die detaillierte Beschreibung der Schemata. Für die Interoperabilität mit anderen Schemata können auch Mappings erzeugt werden, um die Assoziationen zwischen unterschiedlichen Schemata zu modellieren, die für übergreifende Analysen und Werkzeuge notwendig sind. Mithilfe der Beschreibung der Datenmodelle und der Mappings können manuell gepflegte Verarbeitungsregeln verwendet werden, um Elemente beispielsweise zusammenzuführen oder zu zerlegen, etwa – als sehr einfaches Beispiel – die Zerteilung eines Namens in Vor- und Nachnamen.

Für die Neuaufnahme einer Sammlung in ein föderiertes System sind mehrere Voraussetzungen notwendig, die für die Integration einer Sammlung in einem nicht föderierten Ansatz unüblich sind.[18] So ist nicht nur Expertenwissen für die Beschreibung der Daten in den Sammlungen unentbehrlich, sondern auch für den Kontext der Daten, das heißt etwa den organisatorischen Rahmen der Sammlung. Weiterhin ist dieses Fachwissen auch in der Modellierungs- und Transformationsphase im Data Modeling Environment erforderlich, welche die Sammlungen für die weitere Verwendung im jeweiligen Anwendungskontext vorbereiten.

Der föderierte Ansatz in DARIAH-DE ermöglicht Wissenschaftlerinnen und Wissenschaftlern aus den Geisteswissenschaften auch ohne tiefen technischen Hintergrund ihr Expertenwissen über die Sammlungen mittels generischer Werkzeuge zur Verfügung zu stellen. Die Trennung von technischen Abläufen auf der einen und Modellierungs- und Integrationsaufgaben auf der anderen Seite ermöglicht somit den Fokus auf kontextuelle Funktionen und Rahmenbedingungen für die Sammlungen. Die im Folgenden beschriebene Generische Suche stellt einen konkreten Anwendungsfall dar, welcher die Komponenten dieses föderierten Ansatzes nutzt.

16 Gradl, Henrich und Plutte, Heterogene Daten.

17 http://dublincore.org/groups/collections/collection-application-profile/ [zuletzt 4.9.2019].

18 Timo Steyer und Tobias Gradl: A research-oriented and case-based data federation for the Humanities, in: Zenodo (2019), https://zenodo.org/record/2536107 [zuletzt 6.5.2019].

Abb. 2: Einfache Suche

4.2 Generische Suche

Die Generische Suche nutzt die in der Collection Registry und im Data Modeling Environment abgelegten Informationen zur Realisierung einer übergreifenden Suchfunktionalität. Im Gegensatz zum in Abschnitt drei beschriebenen allgemeinen Gathering-Ansatz wird jedoch nicht nur ein zentraler harmonisierter Index aufgebaut, stattdessen werden auf Basis der im Data Modeling Environment hinterlegten Beschreibungen zu den einzelnen Sammlungen einzelne Indexstrukturen für die jeweiligen Sammlungen aufgebaut, die die genauen Informationen dieser Sammlungen und ihre Schemata mit berücksichtigen können. Die Generische Suche hat somit nicht nur Zugriff auf einen homogenisierten – und damit um Details bereinigten – Datenbestand, sondern sie kann die Daten aus den einzelnen Sammlungen in ihrer tiefen Struktur adressieren. Dabei bietet die Generische Suche – wie viele Suchlösungen – einerseits eine einfache Suche und andererseits eine erweiterte Suchmöglichkeit an.

Die einfache Suche (siehe Abb. 2) ist eine vom Mapping zwischen den Sammlungsdaten unabhängige Volltextsuche über alle Attribute und kann auf allen verfügbaren Sammlungen ausgeführt werden. Zur Verfeinerung

Abb. 3: Erweiterte Suche

von Suchanfragen können die üblichen Suchoperatoren (wie AND, OR oder NOT) verfeinert werden.

Anfragen an die Generische Suche führen dabei zu unterschiedlichen Ergebnistypen: den einzelnen Suchtreffern (Ressourcen), verdichteten Ergebnissen je Sammlung (bzw. Kollektion), Subjekten und Termen. Die Darstellung von Ressourcen umfasst, in Abhängigkeit von ihrer Verfügbarkeit, folgende Informationen: die beinhaltende Sammlung, eine Titelangabe, Links zur Anzeige in der Sammlung, assoziierte Subjekte und Inhaltsangaben.

Neben der Präsentation einzelner Suchtreffer werden die Suchergebnisse für jede Sammlung in aggregierter Form dargestellt. Der bzw. dem Suchenden wird dadurch nicht nur ein Überblick über die einzelnen Treffermengen geboten, sondern auch die Relevanzbewertung einzelner Sammlungen für das eigene Informationsbedürfnis erleichtert. Neben den Ergebnissen je

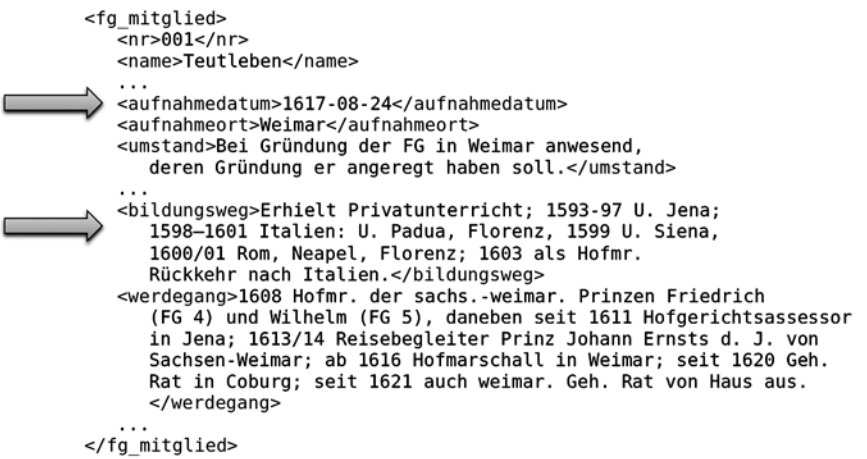

Abb. 4: Ein Beispieldatensatz aus dem Mitgliederverzeichnis der Fruchtbringenden Gesellschaft

Sammlung wird auch eine Aggregation gefundener Subjekte (Personen und Institutionen) sowie häufiger Terme für die Treffermenge dargestellt.[19]

Auch die erweiterte Suche (siehe Abb. 3) ist wie eine Volltextsuche implementiert. Im Gegensatz zur einfachen Suche ist für die Ausführung einer erweiterten Suchanfrage aber die Angabe eines Metadatenschemas erforderlich. Die Suche kann dann aus verschiedenen Teilanfragen zusammengesetzt werden, die sich gezielt auf die einzelnen Elemente des Schemas beziehen können. Eine Besonderheit der Generischen Suche besteht darin, dass zur Überwindung der Heterogenität in den Datenbeständen die Mappings herangezogen werden, die über das Data Modeling Environment verfügbar sind. Die Anfrage wird in alle Schemata übersetzt, zu denen entsprechende Mappings existieren und kann so auf allen Sammlungen mit verbundenen Schemata ausgeführt werden. Neben der Formulierung von Bedingungen zu einzelnen Schemaelementen erlaubt die Generische Suche auch die gezielte Auswahl von Sammlungen oder das Filtern (z.B. nach Ort, Genre oder Datenlieferant).

19 Gradl und Henrich, A novel approach for a reusable federation of research data, S. 382-384.

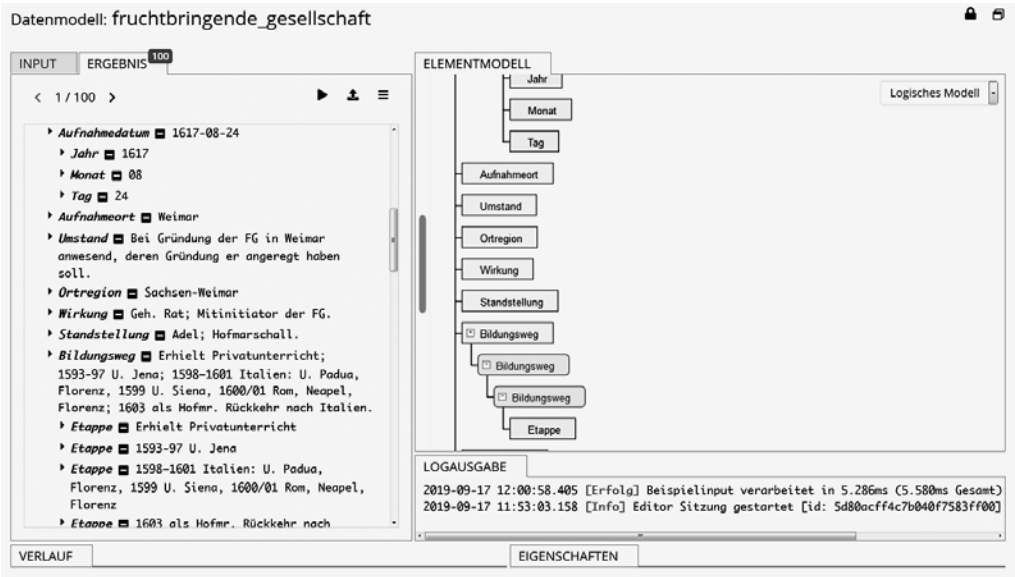

Abb. 5: Verfeinerung des Schemas im Editor für Datenmodelle

5. Nutzung im Forschungsverbund MWW

Der Forschungsverbund Marbach Weimar Wolfenbüttel nutzt die Generische Suche, um eine übergreifende Verbundsuche für Sammlungen aus den drei Häusern umzusetzen. Im Folgenden soll an einem Beispiel verdeutlicht werden, welche Möglichkeiten sich durch die Föderationsarchitektur ergeben und wie diese genutzt werden können. Als konkretes Anwendungsbeispiel wird der Datensatz »Mitgliederverzeichnis der Fruchtbringenden Gesellschaft« betrachtet.[20] Der in Abb. 4 dargestellte Auszug einer XML-Datei beschreibt ein Mitglied der Fruchtbringenden Gesellschaft.

Im ersten Schritt kann das Datenmodell für das Mitgliederverzeichnis der Fruchtbringenden Gesellschaft als XML-Schema im Data Modeling Environment eingelesen werden. Die Analyse der Daten identifiziert dann aber zwei Elemente, für die durch eine Transformation eine Anreicherung des

20 Gabriele Ball, Anne Dickel, Andreas Herz und Timo Steyer: Der gedruckten Edition eine digitale Schwester. Das AEDit-Projekt und die digitale Edition der Fruchtbringenden Gesellschaft, in: Denkströme. Journal der Sächsischen Akademie der Wissenschaften 16 (2016), S. 69-82.

Abb. 6:
Nutzung des Editors zum Verfassen von Grammatiken zur Aufspaltung des Elementes Bildungsweg (linker Teil des Editorfensters oben und rechter Teil unten dargestellt)

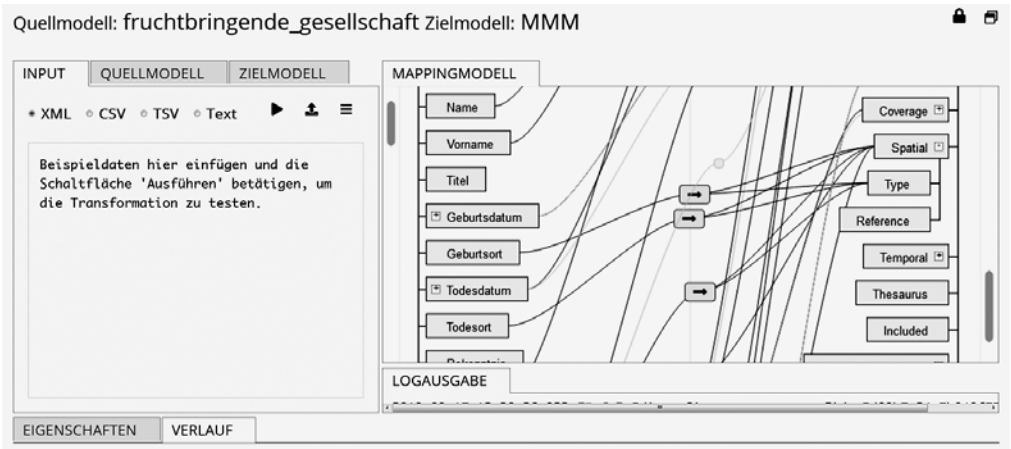

Abb. 7: Mapping zwischen dem Datenmodell für das Mitgliederverzeichnis der Fruchtbringenden Gesellschaft und dem MWW-Metadata-Modell

ursprünglichen Schemas erzielt werden könnte, nämlich die Elemente Aufnahmedatum und Bildungsweg. Das Aufnahmedatum ist im ursprünglichen Schema nur als Ganzes ausgezeichnet. Die einheitlich angewendete Syntax JJJJ-MM-TT erlaubt aber eine genauere Aufteilung, um dann beispielsweise in der Filterung bestimmte zeitliche Intervalle anzugeben. Im Falle des Bildungsweges sind die einzelnen Stationen, durch ein Semikolon getrennt, in einem Textfeld angegeben. Durch eine Aufspaltung in explizit modellierte Etappen ist in Anfragen und durch andere Werkzeuge eine zielgerichtetere Ansprache möglich. Die beiden Elemente werden daher im Data Modeling Environment noch feiner beschrieben, sodass die Daten in einer Form vorliegen, die einen höheren Informationsgehalt verglichen mit dem ursprünglichen Schema aufweist (siehe Abb. 5).

So ist das Aufnahmedatum nun mit drei separaten Elementen für Tag, Monat und Jahr abgebildet. Am Beispiel des Elementes Bildungsweg, das in die unterschiedlichen Etappen der jeweiligen Laufbahn aufgespalten wird, soll die zur Verfeinerung zu nutzende Grammatik angedeutet werden. Diese ist exemplarisch in Abb. 6 dargestellt.

Unten links in Abb. 6 ist die Grammatik zu sehen, die bestimmt, dass ein Weg (oben in der Mitte als Einstiegsregel definiert) aus einer Folge von Etappen besteht, die durch Semikolons getrennt sind. Oben rechts auf der Seite können Testdaten eingegeben werden für die direkt darunter im »Ergebnis

Abb. 8: Startseite der MWW-Verbundsuche

der Transformation« geprüft werden kann, ob die eigene Grammatik wie gewünscht arbeitet. Die Spezifikation der Regeln, die für die Transformation der Elemente angewendet werden, wird somit mit Funktionen einer domänenspezifischen Sprache umgesetzt.[21]

Die so verfeinerten Datenmodelle können dann durch Mappings auch mit anderen Schemata verbunden werden. Im Kontext des MWW-Verbundes hat man sich dabei auf das eigens spezifizierte MWW-Metadata-Modell (MMM) zur Integration auf eine einheitliche Sicht der MWW-Datenmodelle verständigt.[22] Abb. 7 zeigt einen Ausschnitt des Mappings vom Datenmodell für das Mitgliederverzeichnis der Fruchtbringenden Gesellschaft zum MMM. Hier

21 Tobias Gradl und Andreas Henrich: Extending Data Models by Declaratively Specifying Contextual Knowledge, in: DocEng '16: Proceedings of the 2016 ACM Symposium on Document Engineering, S. 123-126.

22 Steyer und Gradl, A research-oriented and case-based data federation for the Humanities.

Abb. 9: Visualisierung von »Migrationsprofilen« im Cosmotool

wird unter anderem der Geburtsort eines Mitglieds der Fruchtbringenden Gesellschaft mit den korrespondierenden Elementen des MWW-Datenmodells (Spatial und Type) verbunden.

Neben dem Geburtsort ist im Beispiel auch der Todesort einer Person an dieses Element geknüpft, da im integrierten MWW-Datenmodell alle Arten von Ortsbezügen über Spatial Elemente modelliert werden, deren genauere Bestimmung dann über das Unterelement Type festgelegt wird. Der Geburtsort wird also auf einen Ort mit dem Typ »Geburtsort« und der Todesort auf einen Ort mit dem Typ »Todesort« abgebildet. Dies verdeutlicht, dass auch Mappings zwischen Schemata möglich sind, die unterschiedliche Modellierungsansätze verfolgen.

Die MWW-Verbundsuche (siehe Abb. 8) ermöglicht derzeit eine Suche über 26 Sammlungen in mehr als 200.000 Dokumenten.[23] Die Suche startet in der einfachen Suche und bietet die Möglichkeit zum Übergang in der er-

23 https://vfr.mww-forschung.de/suche [zuletzt 4.9.2019].

weiterten Suche. Zusätzlich wird unter der Suchleiste eine Schlagwortwolke mit relevanten Termen der Sammlungen angezeigt.

6. Weitere Nutzungsformen und Ausblick

Die im fünften Kapitel beschriebene Nutzung der Generischen Suche für die MWW-Verbundsuche stellt dabei nur einen der möglichen Anwendungsfälle der Generischen Suche aus DARIAH-DE dar. Neben dem Hosten einer eigenen Instanz der Generischen Suche, wie es bei der MWW-Verbundsuche der Fall ist, kann auch eine gehostete Suche verwendet werden, bei der keine Verwaltung eines eigenständigen Index notwendig ist. Stattdessen können die Daten im Kontext des DARIAH-DE-Projekts verwaltet und als eine *gebrandete* Suche angeboten werden. Zusätzlich zu einer Benutzeroberfläche in den eigenen Farben und dem Anzeigen des eigenen Logos kann eine gebrandete Suche auf ausgewählte Sammlungen beschränkt werden. So können Projekte, die ihre Sammlungen in DARIAH-DE registrieren, mit geringem Aufwand eine Suche über ihre Sammlungen realisieren.[24]

Ein weiterer konkreter Anwendungsfall der Generischen Suche, bzw. des in Abb. 1 dargestellten Data Modeling Environments, ist das Cosmotool;[25] es wurde im Rahmen von DARIAH-DE als Prototyp entwickelt.[26] Die Datenquellen des Cosmotools umfassen sowohl biografische Texte aus Wikipedia als auch strukturierte Informationen aus WikiData. Die Daten werden für eine übergreifende Analyse und Visualisierung der Orte, an denen sich historische Persönlichkeiten aufgehalten haben, verwendet, indem unter Zuhilfenahme von NLP-Technologien aus den Quellen Strukturen mit vier Elementen für die beteiligte Person, den Ort, die Zeit sowie das entsprechende Ereignis extrahiert werden. Zusätzlich können diese NLP-Technologien in den Grammatiken, die in Kapitel fünf beschrieben wurden, genutzt werden. Ferner ermöglichen die Grammatiken auch die Analyse von Wikipedia-Seiten

24 Harald Lordick, Tobias Gradl, und Andreas Henrich: Judaica recherchieren – Unterstützung bei der Realisierung forschungsspezifischer Suchlösungen durch die generische Suche von DARIAH-DE, in: DHd 2016: Modellierung, Vernetzung, Visualisierung – die Digital Humanities als fächerübergreifendes Forschungsparadigma, Konferenzabstracts, Universität Leipzig 2016, S. 138-143.

25 Tobias Gradl, Anna Aschauer, Swantje Dogunke, Lisa Klaffki, Stefan Schmunk und Timo Steyer: Daten sammeln, modellieren und durchsuchen mit DARIAH-DE, in: Zenodo (2017), doi: http://doi.org/10.5281/zenodo.582316.

26 https://cosmotool.de.dariah.eu/cosmotool/personsearch/ [zuletzt 4.9.2019].

anhand des Markups. Die extrahierten Strukturen werden anschließend nach Personen geordnet und in einem Profil visualisiert (siehe Abb. 9).[27]

Die verschiedenen Beispiele zeigen, wie vielfältig Suchlösungen über Sammlungen sein können, die auf einer generischen Plattform zur föderierten Suche basieren.

27 Tobias Gradl und Andreas Henrich (2016c): Nutzung und Kombination von Daten aus strukturierten und unstrukturierten Quellen zur Identifikation transnationaler Lebensläufe, in: DHd 2016: Modellierung,Vernetzung, Visualisierung – die Digital Humanities als fächerübergreifendes Forschungsparadigma, Konferenzabstracts, Universität Leipzig, S. 129-132.