Die DARIAH-DE Architektur zur forschungsorientierten Föderation von Kollektionen in den Digital Humanities

Christoph Plutte[1], Tobias Gradl[2] und Andreas Henrich[2]

[1] Berlin-Brandenburgische Akademie der Wissenschaften,
TELOTA und DARIAH-DE, Jägerstr. 22/23, D-10117 Berlin
[2] Universität Bamberg, Lehrstuhl für Medieninformatik,
An der Weberei 5, D-96047 Bamberg

1 Einleitung

Für die kultur- und geisteswissenschaftliche Forschung relevante Ressourcen finden sich zu großen Teilen in den Sammlungen von Museen, Archiven, Bibliotheken, Universitäten und außeruniversitären Forschungseinrichtungen. Mit der Erweiterung des Anwendungsbereiches der Digital Humanities von den Sprachwissenschaften[3] hin zu einer ganzheitlichen Sicht auf die Kultur- und Geisteswissenschaften seit den 1990ern [1] wurden vermehrt Methoden, Anwendungen und Standards für die Digitalisierung, Analyse und Beschreibung von Ressourcen geschaffen. Die Menge der heute durch öffentliche Netzwerke verfügbaren und für die kultur- und geisteswissenschaftliche Forschung relevanten Kollektionen steigt nicht zuletzt aufgrund der Verwendung von Zugriffs- und Beschreibungsstandards stetig an und bietet Forscherinnen und Forschern einen potenziellen Zugang zu einer Vielzahl heterogener Ressourcen.

In diesem Vortrag stellen wir eine neuartige Föderationsarchitektur vor, die auf eine Erfassung und Fall-basierte Zusammenführung von Forschungsdaten nach den individuellen Bedürfnissen von Forschungsprojekten abzielt. Digitale Sammlungen werden zentral verzeichnet, zur Vermeidung von Informationsverlusten jedoch nicht harmonisiert, sondern in Form von Beziehungen auf Schemaebene assoziiert, wodurch die Verwendung einer dynamisch föderierten Datenbasis in breiten und interdisziplinären, wie auch in fachspezifischen Anwendungskontexten ermöglicht werden kann [2]. Ein übergeordnetes Ziel besteht insbesondere in der Nutzbarmachung des durch Experten hinterlegten Wissens zu Kollektionen und Daten sowie deren Beziehungen für einen weiten Anwenderkreis.

2 Anwendungskontext

Traditionelle Integrationsansätze folgen häufig dem Muster eines physisch harmonisierten Datenbestands auf Basis eines zentralen Schemas [3,4]. Verteilte und heterogene, semi-strukturierte Daten werden hierbei in ein gemeinsames Schema

[3] vgl. die Ausführungen zu *Humanities Computing* in [1]

übersetzt und stehen für eine einfache Weiterverarbeitung in integrierter Form zur Verfügung. Eine zentrale Aufgabe dieses Ansatzes besteht in der Umsetzung eines hinsichtlich der notwendigen Granularität geeigneten Integrationsschemas. In Bezug auf die Digital Humanities als ganzheitliche Anwendungsdomäne, die sich in Form spezifischer, interdisziplinärer und auch übergreifender Informationsbedürfnisse äußert, führt die Integration aller Disziplinen und Perspektiven jedoch entweder zu Schemata kaum verwaltbarer Komplexität oder—bei der Verwendung eines einfachen Modells, wie z. B. Dublin Core—zum Verlust großer Anteile disziplinspezifischer Information.

Für die Konzeption der in DARIAH-DE umgesetzten Föderationsarchitektur werden im Folgenden zwei Anwendungsfälle vorgestellt, deren unterschiedliche Anforderungen die Einschränkungen eines solchen zentralistischen Integrationsansatzes verdeutlichen:

Generische Suche Mit der generischen Suche verfolgt DARIAH-DE das Ziel, eine übergreifende Suchmöglichkeit zu schaffen, welche die Eigenschaften der Breiten- und Tiefensuche so vereint, dass eine dynamische Anpassung der Suche—z. B. im Hinblick auf eine mögliche Facettierung—erreicht werden kann [5]. Die übergreifende Suche in eng assoziierten Datenquellen erlaubt—unter Anwendung der in der DARIAH-DE Crosswalk Registry definierten Assoziationen und Transformationsregeln—eine detaillierte Auseinandersetzung mit den betrachteten Daten (Tiefensuche). Mit einer wachsenden Zahl einbezogener Kollektionen wird die Granularität der Betrachtung und Facettierung ggf. mangels vorhandener Verbindungen reduziert und nimmt die Form einer Breitensuche ein. Für die dynamische Funktionalität der generischen Suche ist die ad-hoc-Integration ausgewählter Kollektionen basierend auf den für eine konkrete Anfrage relevanten Kollektionen und den zwischen diesen vorliegenden Assoziationen erforderlich, um die jeweils zur Verfügung stehende Granularität von Daten nutzen zu können.

Datenintegration Im Gegensatz zu der dynamischen, strukturellen Adaption der generischen Suche an die Zusammensetzung der für eine Anfrage ausgewählten Kollektionen zielen Lösungen der Datenintegration oftmals auf eine Konsolidierung einer a-priori definierten Auswahl von Datenquellen ab [3]. Anforderungen an eine kollektionsübergreifende Integration sind wesentlich von der verfolgten Forschungsfrage abhängig und können z. B. im Kontext der Ablösung von Systemen durch Neuentwicklungen, aber auch für die Ausweitung der Datenbasis einer bestehenden Analyse- und Visualisierungslösung, wie beispielsweise dem DARIAH-DE Geobrowser [6], auftreten. Die Anwendung eines zentralen Integrationsschemas bzw. einer zentralen Ontologie führt im Fall der Datenintegration im Gesamtkontext der Digital Humanities zu Problemen, insbesondere wenn eine spezifische Auswahl von Kollektionen für konkrete Forschungsfragen zusammengefasst werden soll. Werden so beispielsweise Kollektionen aus archäologischen und kunsthistorischen Kontexten integriert, so führt die direkte Integration der spezifischen Datenstrukturen zu einem erhöhten Informationsgehalt gegenüber einer globalen Struktur, die den Fachspezifika nicht gerecht werden kann.

3 Föderationsarchitektur

Die in DARIAH-DE gewählte Architektur (vgl. Abbildung 1) besteht aus der *Collection Registry* zur Verzeichnung von Kollektionen, der *Schema Registry* zur Verwaltung von Schemata, und der *Crosswalk Registry* zur Beschreibung von Assoziationen zwischen verschiedenen Schemata. Integrative Dienste wie die *generische Suche* setzen für die Interpretation und Verarbeitung von Daten der verzeichneten Kollektionen auf den durch die Registries angebotenen Webservices auf.

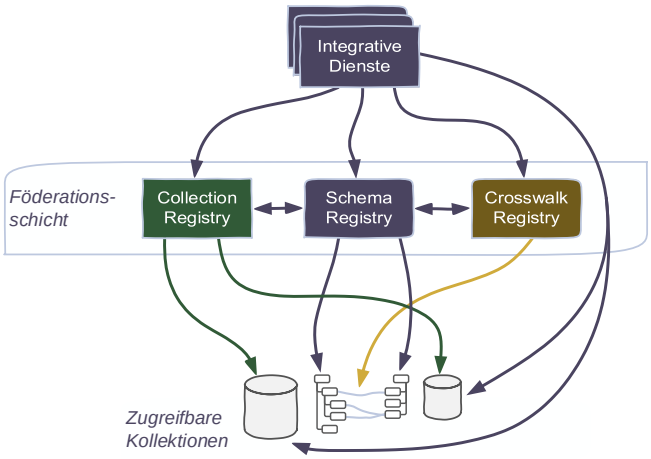

Abb. 1. Komponenten und Zusammenwirken der Föderationsarchitektur

Für eine Forscherin, die eine Sammlung im Rahmen der Förderationsarchitektur registrieren und damit für die Suche, Analyse und den Vergleich mit anderen Sammlungen zur Verfügung stellen möchte, ergibt sich im Zusammenspiel mit der generischen Suche ein Ablauf in vier Schritten (vgl. Abbildung 2):

1. Wenn die entsprechende Kollektion noch nicht in der Collection Registry verzeichnet ist, wird in einem ersten Schritt eine neue Beschreibung der Kollektion und insbesondere ihrer Zugriffsdienste angelegt.
2. Im zweiten Schritt kann die Forscherin das in der Kollektion verwendete Schema (Dublin Core, Lido etc.) beschreiben bzw. die konkrete Verwendung eines allgemeinen Schemas spezifizieren. (Vererbung)
3. Das so erstellte bzw. angepasste Schema kann in Abhängigkeit von konkreten Forschungsfragen im dritten Schritt iterativ mit weiteren Schemata assoziiert werden. (Definition von Crosswalks)
4. Im vierten Schritt indiziert die generische Suche die zugreifbaren Daten der Kollektion anhand der in den Registries hinterlegten Informationen und stellt diese für übergreifende Suchanfragen bereit.

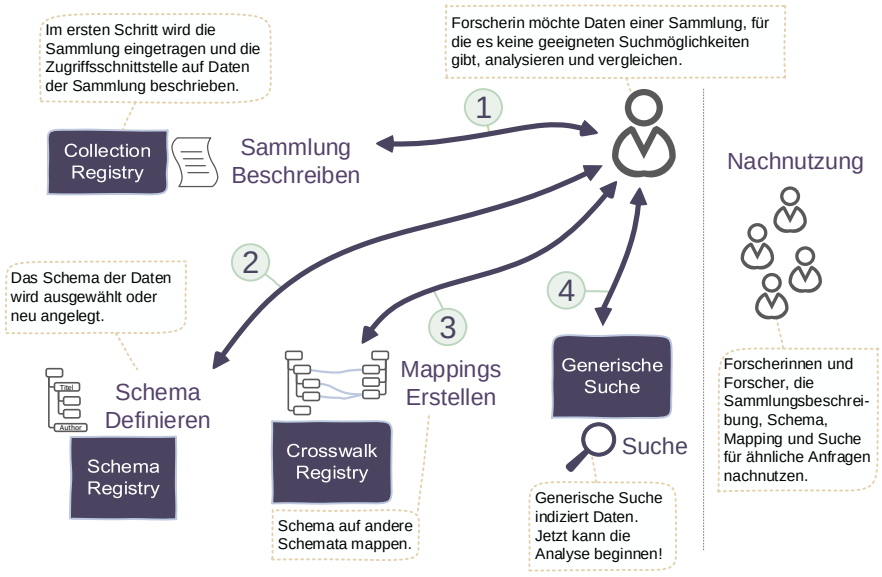

Abb. 2. Schritte der Registrierung von Kollektionen und Schemata

Die sich aus den einzelnen Schritten ergebenden Informationen stehen zur Nachnutzung für verwandte Forschungsinteressen zur Verfügung und können von integrativen Diensten über Webservices abgefragt werden.

3.1 Collection Registry

Die Collection Registry[4] ist ein online zugängliches zentrales Verzeichnis, in dem relevante Sammlungen registriert und durch Fachwissenschaftler beschrieben werden. Das Datenmodell für die Sammlungsbeschreibungen basiert auf dem Dublin Core Collection Application Profile [7], das insbesondere im Hinblick auf die Beschreibung von Zugriffspunkten erweitert wurde. Die Sammlungsbeschreibungen decken neben Verschlagwortung, zeitlichen und geografischen Dimensionen auch Sammlungsformate und Informationen zur Datenpflege ab. Ein Schwerpunkt liegt auf der Beschreibung von Zugriffpunkten wie OAI-PMH-Schnittstellen zur Abfrage der Sammlungselemente für die Weiterverarbeitung durch assoziierte Komponenten. Komponenten können alle erforderlichen Informationen für einen Zugriff auf die Sammlungselemente aus der Collection Registry über Webschnittstellen (REST) beziehen [8].

Neben maschinenlesbaren Schnittstellen für den Zugriff auf die Sammlungsbeschreibungen bietet die Collection Registry ein Benutzerinterface, welches das Anlegen von Sammlungsbeschreibungen und anderen Datenobjekten ebenso unterstützt wie das Suchen, Aktualisieren und Löschen von vorhandenen Beschreibungen. Ausgewählte kontrollierte Vokabulare unterstützen die Eingabe und die

[4] `http://demo2.dariah.eu/colreg/`

Interaktion mit der Schema Registry erlaubt es, eine Sammlungsbeschreibung mit einem bestimmten Schema zu verknüpfen. Für den langfristigen Betrieb wird eine Moderation von DARIAH-DE organisiert, die die Qualität der Daten gewährleisten wird.

3.2 Schema- und Crosswalk Registry

In der Schema- und Crosswalk Registry[5] werden semi-strukturierte Datenmodelle und Korrelationen (siehe Abbildung 3) zwischen diesen aus der primären Zielsetzung heraus beschrieben, expliziertes Expertenwissen zu Kollektionen und den darin verwalteten Daten nachnutzen zu können. Die Spezifikationen von Strukturen z. B. in XML Schema können hierbei in Bezug auf eine Kollektion erweitert und konkretisiert werden, wodurch die Semantik originärer Daten erhalten bleibt und dennoch eine Verfeinerung um zunächst implizites Hintergrundwissen erfolgen kann.

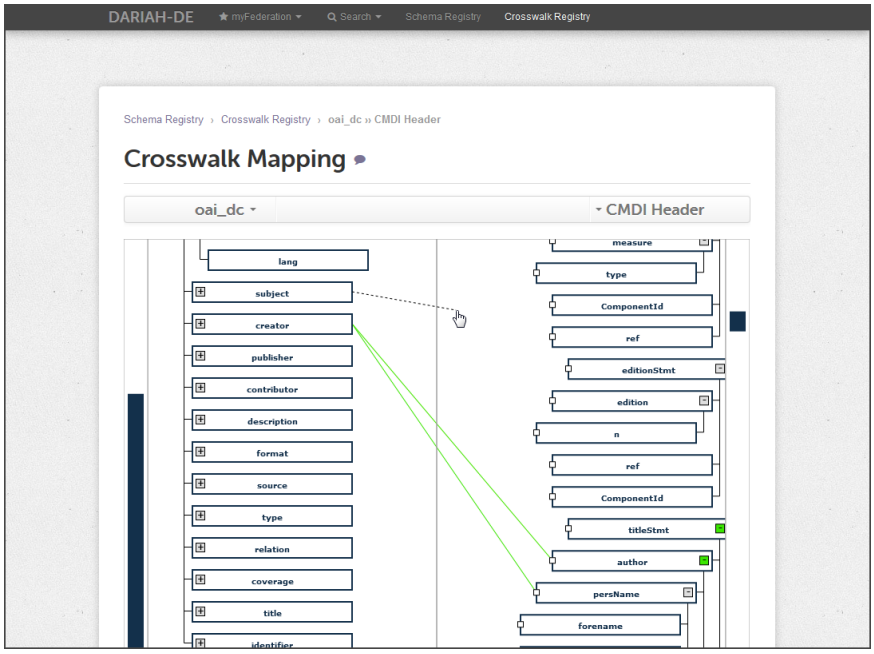

Abb. 3. Assoziation von Schemata in der Crosswalk Registry

Abbildung 4 zeigt beispielhaft Möglichkeiten zur Verfeinerung von Dublin Core basierend auf dem Wissen zu spezifischen Kollektionen. Manuell modellierte Verarbeitungsregeln führen dabei zu einer erweiterten Version eines Datensatzes, welcher für ein Mapping mit komplexeren Strukturen zur Verfügung steht.

[5] `http://dev3.dariah.eu/schereg/`

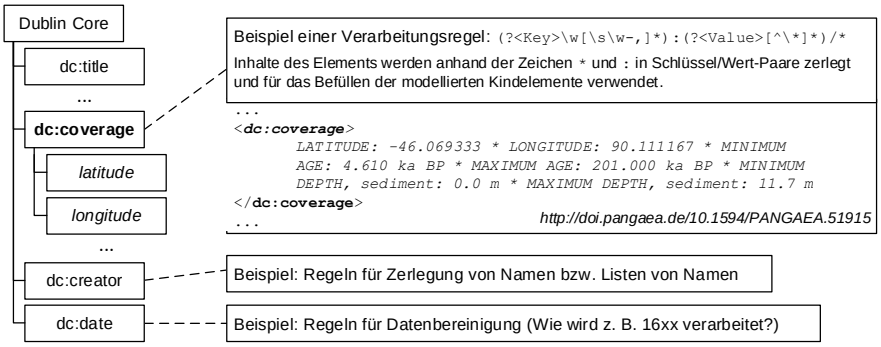

Abb. 4. Beispiele zur kollektionsspezifischen Ergänzung von Dublin Core

Dadurch, dass auch der unveränderte Datensatz weiterhin verwendet werden kann, wird zudem die Kompatibilität zu generischem Dublin Core sichergestellt.

3.3 Generische Suche als durchgeführter Use-Case

Mit der generischen Suche[6] wird im Rahmen von DARIAH-DE ein Anwendungsfall der Datenföderation umgesetzt. Hierbei werden Daten aus den in der Collection Registry verzeichneten Kollektionen nach den in der Schema Registry explizierten Strukturen verarbeitet und indexiert. Die Heterogenität der Ressourcen wird zum Zeitpunkt konkreter Suchanfragen basierend auf der zu durchsuchenden Menge von Kollektionen mit Hilfe der Crosswalk Registry aufgelöst.

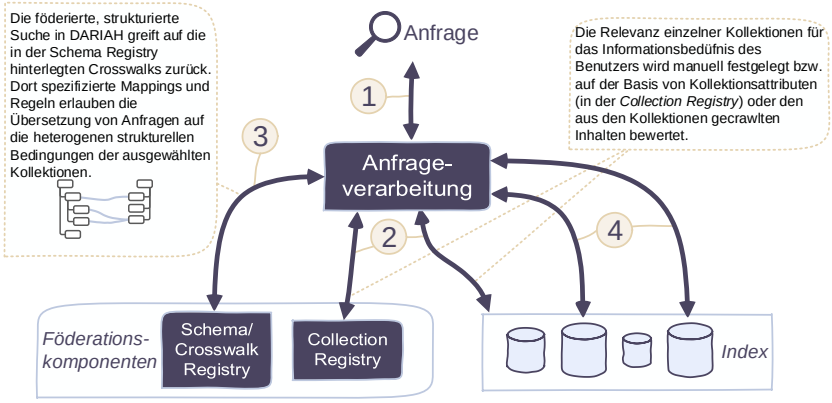

Abb. 5. Anfrageverarbeitung in der generischen Suche

[6] http://dev3.dariah.eu/search/

Abbildung 5 skizziert den Verlauf der Anfrageverarbeitung und die Interaktion mit den Komponenten der Föderationsarchitektur: Am Beginn steht ein Informationsbedürfnis im Rahmen einer Forschungsfrage (1). Zunächst wird nun interaktiv oder automatisch auf Basis der Collection Registry und der von der generischen Suche angebotenen Kollektionssuche die Teilmenge der Kollektionen ermittelt, auf denen die Suche durchgeführt werden soll (2). Je feingranularer die Schemata der gewählten Kollektionen in der Crosswalk Registry miteinander verknüpft sind, umso differenzierter können die Anfragen spezifiziert und ausgeführt werden. Der Nutzer kann die Anfrage dabei in einem Schema seiner Wahl formulieren, das als temporäres Integrationsmodell genutzt wird. Die Anfrage wird auf Basis der relevanten Schemainformationen und Transformationsregeln (3) dann so transformiert, dass sie auf den Indices, die die Daten in ihrem ursprünglichen Schema verwalten, ausgeführt werden kann (4). Ermittelte Ergebnisse werden zusammengefasst und bzgl. ihrer Relevanz für die Anfrage sortiert.

4 Zusammenfassung

Die vorgestellte Förderationsarchitektur folgt dem Prinzip der dezentralen Integration von Daten. Mit der generischen Suche kann gezeigt werden, wie durch die Verwendung der einzelnen Föderationskomponenten ein echter Mehrwert für die Recherche über verschiedene heterogene Datensammlungen entstehen kann und wie eine Alternative zu zentralistischen Ansätzen entwickelt werden kann. Mit einer ad-hoc Föderation kann gegenüber einer domänenweiten Harmonisierung die Möglichkeit der individuellen Integrierbarkeit von Daten geschaffen werden, die auf dem Wissen und der Kollaboration von Spezialisten aus verschiedenen Fachwissenschaften basiert und durch ein breites Publikum in Abhängigkeit von konkreten Forschungsfragen konkret eingesetzt werden kann.

Literatur

1. S. Schreibman, R. G. Siemens, and J. Unsworth, Eds., *A companion to digital humanities*, ser. Blackwell companions to literature and culture. Malden and Mass: Blackwell Pub., 2004, vol. 26.
2. A. Henrich and T. Gradl, "DARIAH(-DE): Digital Research Infrastructure for the Arts and Humanities — Concepts and Perspectives," *International Journal of Humanities and Arts Computing*, vol. 7, no. supplement, pp. 47–58, 2013.
3. M. Lenzerini, "Data Integration: A Theoretical Perspective," in *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, S. Abiteboul, Ed. New York and NY: ACM, 2002, p. 233.
4. S. Peroni, F. Tomasi, and F. Vitali, "Reflecting on the Europeana Data Model," in *Digital Libraries and Archives*, ser. Communications in Computer and Information Science, M. Agosti, F. Esposito, S. Ferilli, and N. Ferro, Eds. Berlin and Heidelberg: Springer Berlin Heidelberg, 2013, vol. 354, pp. 228–240.
5. T. Gradl and A. Henrich, "DARIAH-DE Generische Suche (M 1.4.2.1 - Prototyp): DARIAH-DE Arbeitspapier," 2013. [Online]. Available: https://dev2.dariah.eu/wiki/download/attachments/2295542/Report%20M1.4.2.1.docx

6. M. Romanello, “DARIAH Geo-browser: Exploring Data through Time and Space,” 2013. [Online]. Available: http://de.slideshare.net/56k/dariah-geobrowser-exploring-data-through-time-and-space
7. Dublin Core Metadata Initiative, “Dublin Core Collections Application Profile,” 2007. [Online]. Available: http://dublincore.org/groups/collections/collection-application-profile/
8. C. Plutte and P. Harms, “Collection Registry (M 1.2.2): DARIAH-DE Arbeitspapier,” 2012. [Online]. Available: https://dev2.dariah.eu/wiki/download/attachments/14651583/M1.2.2_Collection_Registry_incl_DCLAP.pdf