Zeitschrift für digitale Geisteswissenschaften

Artikel aus:
Sonderband 1 der ZfdG: Grenzen und Möglichkeiten der Digital Humanities. Hg. von Constanze Baum und Thomas Stäcker. 2015. DOI: 10.17175/sb01

Titel:
Heterogene Daten in den Digital Humanities: Eine Architektur zur forschungsorientierten Föderation von Kollektionen

Autor/in:
Tobias Gradl

Kontakt: tobias.gradl@uni-bamberg.de
Institution: Otto-Friedrich-Universität Bamberg
GND: 1084606585 ORCID: 0000-0002-1392-2464

Autor/in:
Andreas Henrich

Kontakt: andreas.henrich@uni-bamberg.de
Institution: Otto-Friedrich-Universität Bamberg
GND: 111819601 ORCID:

Autor/in:
Christoph Plutte

Kontakt: plutte@bbaw.de
Institution: Berlin-Brandenburgische Akademie der Wissenschaften
GND: 1084027208 ORCID:

DOI des Artikels:
10.17175/sb001_020

Nachweis im OPAC der Herzog August Bibliothek:
830207090

Erstveröffentlichung:
19.02.2015

Letzte Überprüfung aller Verweise:
24.05.2016

GND-Verschlagwortung:
Datenintegration | Digital Humanities | Softwaresystem | Architektur (Informatik) |

Zitierweise:
Tobias Gradl, Andreas Henrich, Christoph Plutte: Heterogene Daten in den Digital Humanities: Eine Architektur zur forschungsorientierten Föderation von Kollektionen. In: Grenzen und Möglichkeiten der Digital Humanities. Hg. von Constanze Baum / Thomas Stäcker. 2015 (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 1). PDF Format ohne Paginierung. Als text/html abrufbar unter DOI: 10.17175/sb001_020.

Tobias Gradl, Andreas Henrich, Christoph Plutte

Heterogene Daten in den Digital Humanities: Eine Architektur zur forschungsorientierten Föderation von Kollektionen

Abstracts

Traditionelle Ansätze der Datenintegration basieren zumeist auf einer Harmonisierung heterogener Daten im Rahmen einer übergeordneten, integrativen Datenstruktur wie einem globalen Schema oder einer globalen Ontologie. Der vorliegende Beitrag verdeutlicht die Limitationen eines derartigen Harmonisierungsansatzes für den speziellen Kontext der Digital Humanities und zeigt, wie im Rahmen von DARIAH-DE eine forschungsorientierte und fallbasierte Föderation von Daten erreicht werden kann. Mit Hilfe von Collection, Schema und Crosswalk Registry können durch die Registrierung und Assoziation von Kollektionen und darin verwendeten Datenschemata sowohl übergreifende als auch disziplinspezifische Sichten auf Forschungsdaten geschaffen werden.

Traditional approaches to data integration are typically based on the harmonization of heterogeneous data with respect to the constraints of a globally integrative data structure, such as a global schema or ontology. This paper illustrates the limitations of such a harmonization-based approach in the specific context of the Digital Humanities and shows how a research-oriented and case-based data federation can be facilitated by the DARIAH-DE federation architecture. Based on the collection, schema, and crosswalk registries, collections and the data schemata they utilize can be registered and flexibly associated – resulting in the ability to create broad and comprehensive as well as discipline-specific views of research data.

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[1] hin zu einer ganzheitlichen Sicht auf die Kultur- und Geisteswissenschaften seit den 1990ern wurden vermehrt Methoden, Anwendungen und Standards für die Digitalisierung, Analyse und Beschreibung von Ressourcen geschaffen.[2] 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 Beitrag stellen wir eine neuartige Föderationsarchitektur vor, die auf eine Erfassung und fallbasierte Zusammenführung von Forschungsdaten nach den individuellen Bedürfnissen von Forschungsprojekten abzielt. Digitale Sammlungen werden zentral

[1] Vgl. die Ausführungen zu Humanities Computing in Schreibman et al. 2004, S. 3ff.
[2] Grundsätzlich Schreibman et al. 2004, insbesondere S. 564ff.

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.[3] 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.[4] Verteilte und heterogene, semi-strukturierte Daten werden hierbei in ein gemeinsames Schema 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 (DCES) – 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.

2.1 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.

[3] Henrich / Gradl 2013, S. 50f.
[4] Lenzerini 2002, S. 234; Peroni et al. 2013, S. 228f.
[5] Gradl / Henrich 2013, S. 8ff.

2.2 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.[6] 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 [7], 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 (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.

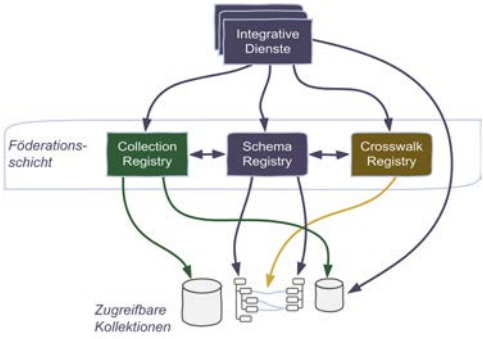

Abb. 1: Komponenten und Zusammenwirken der Föderationsarchitektur [eigene Darstellung].

Für eine Forscherin oder einen Forscher, die oder der eine Sammlung im Rahmen der Föderationsarchitektur registrieren und damit für die Suche, Analyse und den Vergleich mit

[6] Grundsätzlich Lenzerini 2002, besonders S. 234
[7] Überblickend Romanello 2013.

anderen Sammlungen zur Verfügung stellen möchte, ergibt sich im Zusammenspiel mit der generischen Suche ein Ablauf in vier Schritten (Abbildung 2):

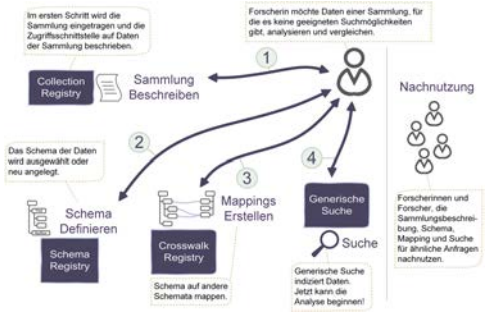

Abb. 2: Schritte der Registrierung von Kollektionen und Schemata [eigene Darstellung].

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 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, das insbesondere im Hinblick auf die Beschreibung von Zugriffspunkten erweitert wurde. Eine beliebig heterogene Sammlung von Ressourcen wird als Collection bezeichnet und beschreibt ein Konstrukt der Anwendungsdomäne, das zur fachlichen Strukturierung von Archiven und Datenquellen eingesetzt werden kann. Collections können selbst direkt Ressourcen oder weitere untergeordnete Teilcollections beinhalten, und sie 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. Für die Auszeichnung mit Schlagworten und Georeferenzierungen werden verschiedene kontrollierte Vokabulare wie LCHS, Dewey Decimal Classification, Geonames u.a. integriert.

Ein Schwerpunkt liegt auf der Beschreibung von Zugriffpunkten wie OAI-PMH-Schnittstellen zur Abfrage der Sammlungselemente für die Weiterverarbeitung durch assoziierte Komponenten. Je Sammlungsbeschreibung können mehrere Zugriffspunkte verzeichnet werden, zu denen neben der URL des Zugriffspunktes auch weitere Angaben wie Zugriffsprotokoll, etwaige Zugriffsbeschränkungen, OAI-PMH subclasses, Dokumentation der Schnittstellen etc. verwaltet werden. Besonders wichtig ist, dass für jeden Zugriffspunkt das von diesem verwendete Schema aus der Schema Registry referenziert wird.

Weiterführende Dienste 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, die 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. Die Collection Registry nimmt eine zentrale Rolle in der Datenföderation ein, dient aber zugleich auch als alleinstehender Dienst und Datenquelle für die Suche und Verwaltung von Metadaten zu Datensammlungen.

3.2 Schema- und Crosswalk Registry

In der Schema- und Crosswalk Registry werden semi-strukturierte Datenmodelle und Korrelationen zwischen diesen (vgl. Abbildung 3) 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-Schemata, 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.

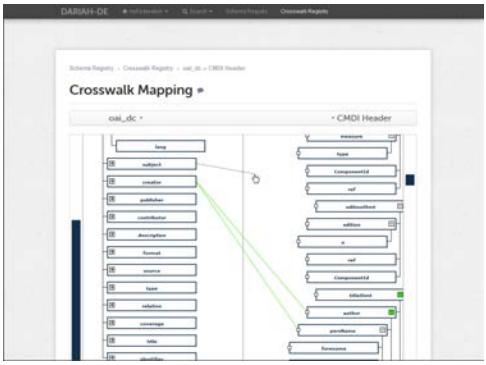

Abb. 3: Assoziation von Schemata in der Crosswalk Registry [eigene Darstellung].

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. Dadurch, dass auch der unveränderte Datensatz

[8] Plutte / Harms 2012, S. 11f.

weiterhin verwendet werden kann, wird zudem die Kompatibilität zu generischem Dublin Core sichergestellt.

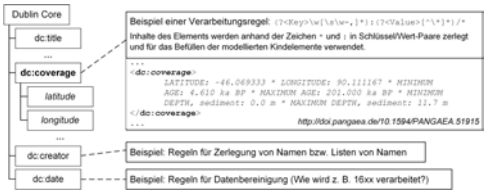

Abb. 4: Beispiele zur kollektionsspezifischen Ergänzung von Dublin Core [eigene Darstellung].

Durch die semantische Assoziation von Kollektionen und darin verwendeten Schemata durch fachwissenschaftliche Experten können integrative Sichten generiert werden, die den Anforderungen konkreter Forschungsfragen entsprechen. Dies wird insbesondere dadurch erreicht, dass nicht technisch motivierte Integrationsziele wie die Vollständigkeit und Korrektheit eines Integrationsschemas im Vordergrund stehen, sondern disziplinspezifische und auch konfliktäre Zusammenhänge modelliert werden können.

Die flexible Anpassbarkeit an übergreifende oder spezifische Fragestellungen wird dabei durch das in Abbildung 5 exemplarisch angedeutete Konzept der forschungsorientierten Föderation digitaler Kollektionen erreicht: Kohärente Bereiche mit eng assoziierten Schemata und Kollektionen werden in der Abbildung durch semantische Cluster widergespiegelt und bilden die Voraussetzung für spezifische Betrachtungen. Für übergreifende Sichten werden wichtige Schemata der einzelnen Cluster (repräsentiert als *S3*, *S5* und *S8*) mit generischen Schemata, wie z. B. Dublin Core – in der Abbildung symbolisiert durch *S10* – assoziiert.

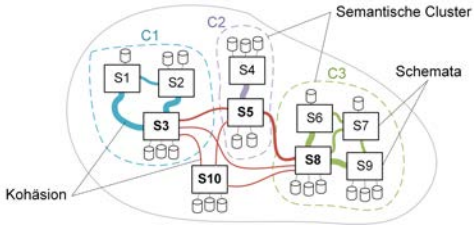

Abb. 5: Prinzip der semantischen Cluster im Beispiel [eigene Darstellung].

3.3 Generische Suche als durchgeführter Use-Case

Mit der generischen Suche 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.

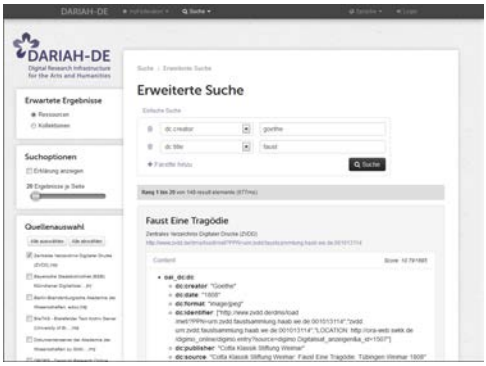

Abb. 6: Anfrageverarbeitung in der generischen Suche [eigene Darstellung].

Abbildung 6 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, in 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 umformuliert, dass sie auf den Indices, die die Daten in ihrem ursprünglichen Schema verwalten, ausgeführt werden kann (4). Ermittelte Ergebnisse werden zusammengefasst und bezüglich ihrer Relevanz für die Anfrage sortiert.

Auf der Basis der eingeführten Föderationsarchitektur und der im Rahmen der generischen Suche implementierten Funktionalität können assoziierte Kollektionen im Rahmen so genannter *Benutzerkollektionen* zusammengestellt und inhaltsspezifisch nachgenutzt werden. Neben der Betrachtung der Daten in externen Tools (wie z. B. dem bereits erwähnten Geobrowser), ermöglicht die generische Suche eine unmittelbare Veröffentlichung der Benutzerkollektion in Form einer so genannten Branded Search, einer eigenen Suchoberfläche, welche sowohl optisch als auch inhaltlich an spezifische Bedürfnisse angepasst ist.

Der Bildschirmausschnitt in Abbildung 7 verdeutlicht insbesondere die optische Abgrenzung von der generischen Suche durch eine konfigurierte Farbgebung und die Verwendung von Such- und Organisationslogos. Neben visuellen Aspekten unterscheidet sich die Branded Search auch in Hinblick auf die zu Grunde liegende Datenbasis: Die in einer Branded Search angebotenen Kollektionen spiegeln bei sämtlichen Such-, Analyse- und Visualisierungsaufgaben die von den Erstellern der Suche getroffene Kollektionsauswahl wider.

Abb. 7: Startseite der generischen Suche in Form einer Branded Search [eigene Darstellung].

4. Zusammenfassung

Die vorgestellte Föderationsarchitektur 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 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 eingesetzt werden kann.

Bibliographische Angaben

A companion to Digital Humanities. Hg. von Susan Schreibman / Ray Siemens / John Unsworth. Oxford 2004. [Nachweis im OPAC]

Tobias Gradl / Andreas Henrich: DARIAH-DE Generische Suche (M 1.4.2.1 - Prototyp): DARIAH-DE Arbeitspapier. 2013. [online, Registrierung erforderlich]

Andreas Henrich / Tobias Gradl: DARIAH(-DE): Digital Research Infrastructure for the Arts and Humanities – Concepts and Perspectives. In: International Journal of Humanities and Arts Computing 7 (2013), S. 47–58. [Nachweis im GBV]

Maurizio Lenzerini: Data Integration: A Theoretical Perspective. In: Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. Hg. von der Association for Computing Machinery. New York 2002, S. 233–246. [Nachweis im GBV]

Silvio Peroni / Francesca Tomasi / Fabio Vitali: Reflecting on the Europeana Data Model. In: Digital Libraries and Archives. Hg. von Maristella Agosti / Floriana Esposito / Stefano Ferilli / Nicola Ferro. Berlin 2013 (= Communications in Computer and Information 354), S. 228–240. [Nachweis im GBV]

Christoph Plutte / Patrick Harms: Collection Registry (M 1.2.2): DARIAH-DE Arbeitspapier. 2012. [online]

Matteo Romanello: DARIAH Geo-browser: Exploring Data through Time and Space. 2013. [online]

Abbildungslegenden und -nachweise

Abb. 1: Komponenten und Zusammenwirken der Föderationsarchitektur [eigene Darstellung].

Abb. 2: Schritte der Registrierung von Kollektionen und Schemata [eigene Darstellung].

Abb. 3: Assoziation von Schemata in der Crosswalk Registry [eigene Darstellung].

Abb. 4: Beispiele zur kollektionsspezifischen Ergänzung von Dublin Core [eigene Darstellung].

Abb. 5: Prinzip der semantischen Cluster im Beispiel [eigene Darstellung].

Abb. 6: Anfrageverarbeitung in der generischen Suche [eigene Darstellung].

Abb. 7: Startseite der generischen Suche in Form einer Branded Search [eigene Darstellung].