

Daten sammeln, modellieren und durchsuchen mit Dariah-DE

Gradl, Tobias

tobias.gradl@uni-bamberg.de
Universität Bamberg

Aschauer, Anna

aschauer@ieg-mainz.de
Leibniz-Institut für Europäische Geschichte (IEG)

Dogunke, Swantje

swantje.dogunke@klassik-stiftung.de
Klassik Stiftung Weimar

Klaffki, Lisa

klaffki@hab.de
Herzog August Bibliothek Wolfenbüttel

Schmunk, Stefan

schmunk@sub.uni-goettingen.de
Niedersächsische Staats- und Universitätsbibliothek
Göttingen

Steyer, Timo

steyer@hab.de
Herzog August Bibliothek Wolfenbüttel

Überblick

Die sammlungsübergreifende Recherche und Nachnutzung geisteswissenschaftlicher Forschungsdaten stehen im Blickpunkt aktueller Forschung in den Digital Humanities. Obwohl das Interesse an einer Zusammenführung digitaler Forschungsdaten bereits kurz nach der Einführung erster digitaler Bibliotheken um die Jahrtausendwende entstand, bleibt die Integration von Forschungsdaten über Sammlungsgrenzen hinweg ein aktuelles Forschungsthema. Bei einer forschungsorientierten Betrachtung von Sammlungen digitaler Daten (also z. B. digitale Texte, Digitalisate, Normdaten, Metadaten) stellt sich die Frage nach den Anforderungen und Erfolgskriterien einer übergreifenden Föderation, Verarbeitung und Visualisierung von Forschungsdaten.

Entgegen der in der Praxis üblichen Orientierung an institutionellen Anforderungen stellen die in Dariah-DE entwickelten Konzepte und Dienste zur Verzeichnung, Korrelation und Zusammenführung von Forschungsdaten

die Bedürfnisse von WissenschaftlerInnen im Kontext ihrer Forschungsfragen in den Mittelpunkt. Dies äußert sich beispielsweise darin, dass Dariah-DE keine strukturellen Bedingungen an Forschungsdaten stellt. Stattdessen können Daten so publiziert, modelliert und integriert werden, dass eine möglichst gute Passung an den jeweiligen geisteswissenschaftlichen Kontext erreicht wird.

Dieser Workshop wird zunächst in Form kurzer Referate Hintergrundwissen zu den Konzepten und Diensten der Dariah-DE Föderationsarchitektur vermitteln. Wichtige Bereiche sind dabei nicht nur die Handhabung der Daten selbst sowie Fragen der Lizenzierung von Forschungsdaten, sondern auch die Nachnutzbarkeit einmal erhobener oder gesammelter Daten für weitere Forschungsfragen oder zur Nutzung durch andere WissenschaftlerInnen. Ein wesentlicher Anteil des Workshops wird dann insbesondere in der Hands-On-Anwendung der Komponenten durch die TeilnehmerInnen selbst bestehen.

Thematische Schwerpunkte

Die wesentlichen Themenschwerpunkte des Workshops können wie folgt zusammengefasst werden:

- Hintergründe und Best Practices zur *Lizenziierung* und *Nachnutzbarkeit* von Forschungsdaten
- Beschreibung und nachhaltige *Referenzierbarkeit* von Sammlungen in der Dariah-DE Collection Registry
- *Modellierung* von Daten in der Dariah-DE Schema Registry zur Beschreibung des Erstellungskontexts von Daten sowie deren Transformation in einen forschungsorientierten Verwendungskontext

Anhand der generischen Suche von Dariah-DE werden die Auswirkungen der Benutzerinteraktion im Rahmen des Workshops sofort erkennbar, d. h. referenzierte Daten werden anhand der entwickelten Datenmodelle verarbeitet und können gemeinsam durchsucht und analysiert werden.

Der gesamte Workshop wird thematisch begleitet von der konkreten, historischen Anforderung (vgl. Szöllösi-Janze, Panter & Paulmann 2015), biographische Daten und Texte aus verschiedenen Datenquellen zu verarbeiten. Die schließlich integrierten biographische Profile (vgl. Gradl & Henrich 2016b) können zur Unterstützung konkreter historischer Forschung herangezogen werden. Das Beispiel ist so gewählt, dass den Teilnehmerinnen und Teilnehmern eine konzeptuelle Übertragung auf ihre eigenen Daten und Forschungsfragen erleichtert wird.

Zielpublikum

Der Workshop richtet sich gleichermaßen an:

- geisteswissenschaftliche WissenschaftlerInnen in den unterschiedlichsten Phasen des akademischen Werdegangs

- VertreterInnen von Institutionen, die Datensammlungen im Rahmen von Dariah-DE auffindbar und zugreifbar machen möchten,
- sowie auch VertreterInnen der Informatik, die ein Interesse an der Implementierung von Dariah-DE Komponenten bzw. den Datenaustausch auf Basis maschinell zugreifbarer Schnittstellen haben.

Wer bereits über digitale Daten verfügt, ist herzlich eingeladen, diese für die Hands-On-Sessions mitzubringen, um an diesen konkreten Beispielen die Dariah-DE-Tools zu erarbeiten. Für TeilnehmerInnen, die keine geeigneten Daten mitbringen können, werden Beispiele zur Verfügung gestellt. Bitte bringen Sie in jedem Fall Ihren eigenen Laptop mit!

Der Workshop ist Teil zweier konzeptionell eigenständiger Einreichungen zu den Angeboten der digitalen Forschungsinfrastrukturen TextGrid und Dariah-DE. Der erste Workshop hat den Titel "Annotieren und Publizieren mit Dariah-DE und TextGrid". Der Besuch beider Workshops ermöglicht eine grundlegende und umfassende Einführung in und Anwendung von Architektur, Tools, Diensten und Workflows zum Annotieren, Sammeln, Modellieren, Recherchieren und Publizieren geisteswissenschaftlicher Forschungsdaten.

Inhalte und Ablauf des Workshops

I - Impulsreferate "Sammeln"

- Lizenzierung, Referenzierung und Nachnutzbarkeit von Forschungsdaten (*Lisa Klaffki*)
- Transnationale Biographien als Beispiel einer historischen Motivation für die forschungsorientierte Föderation von Dariah-DE (*Anna Aschauer*)

II - Impulsreferat "Modellieren"

- Forschungsorientierte Modellierung und Korrelation von Daten in der Föderationsarchitektur von Dariah-DE (*Stefan Schmunk, Tobias Gradl*)

III - Impulsreferat "Durchsuchen"

- Integriertes Suche über heterogene Datenbestände – Anforderungen und Lösungsansätze im Bereich des kulturellen Erbes (*Timo Steyer, Swantje Dogunke*)

IV - Hands-on Session "Sammeln, Modellieren & Durchsuchen"

- Anwendung der Föderationsarchitektur und generischen Suche von Dariah-DE (*Tobias Gradl*)
 - Modellierung von Daten und Vorbereitung einer Nachnutzung
 - Assoziation heterogener wissenschaftlicher Sammlungen
 - Verfeinerung der benutzerdefinierten Suchmöglichkeiten in der generischen Suche (Suchbild, Ranking etc.)
 - Anpassung der generischen Suche und Bereitstellung benutzerdefinierter Suchen

Komponenten des Workshops

Abbildung 1 zeigt die Zusammenhänge zwischen den für die Dariah-DE Infrastruktur zugänglichen Kollektionen, den Registries und der generischen Suche. In der Übersicht dargestellte Komponenten und Verbindungen werden im Rahmen des Workshops live durch die TeilnehmerInnen beeinflusst, weshalb wir in diesem Abschnitt eine vorbereitende Einführung anbieten möchten. Für weitere Informationen erlauben wir uns einen Verweis auf die weiterführenden Publikationen am Ende des Dokuments.

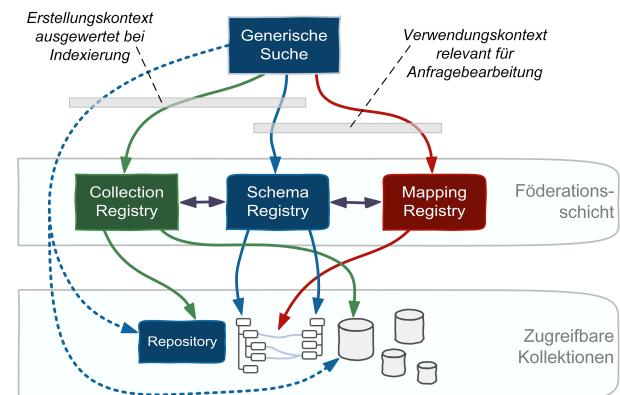


Abbildung 1: Dariah-DE Föderationsarchitektur

"Sammeln": Collection Registry

Die Collection Registry (vgl. Abbildung 2) 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.

Vereinheitlicht	Titel	KollektionsTyp	Zugang	Aktuelle
Bürokratische Foto Marburg (Bürokratische Foto Marburg)	Fotosammlung	Beacon	14. Februar 2016	✓
Deutsch-jüdische Geschichte (user-dgl)	Presentationen, Poster, Forschungsdaten	OAI-PMH	14. Februar 2016	✓
Deutsch-jüdische Geschichte (digital) (dgd)	Blog	OAI-PMH;	14. Februar 2016	✓
Deutsch-jüdische Publizistik (DJP)	Textsammlung	Beacon	14. Februar 2016	✓
Deutsches Textarchiv (DTA)	Textsammlung	OAI-PMH;	14. Februar 2016	✓
Digitales Personenregister der Germania Sacra (Germania Sacra)	Datenbank, Textsammlung	Beacon	14. Februar 2016	✓
Epos (Epos)	Sammlung	Beacon	14. Februar 2016	✓
Katloope (Katloope)	Nachlässe	Beacon	14. Februar 2016	✓
Kalonymos. Beiträge zur deutsch/jüdischen Geschichte aus dem Ravennischen Judentum	Journal (PDF)	Beacon	15. Februar 2016	✓

Abbildung 2: Bildschirmausschnitt „Sammlungen“ der Dariah-DE Collection Registry

“Modellieren”: Schema Registry / Mapping Registry

In der Schema und Mapping Registry werden Datenmodelle und Korrelationen zwischen diesen beschrieben. Die grundlegende Zielsetzung besteht in der Definition und nachnutzbaren Modellierung der Erstellungs- und Verwendungskontexte von Daten:

- Erstellungskontext:** Ausgehend beispielsweise von einem XML-Schema wird ein Datenmodell angelegt, verfeinert und um Hintergrundwissen z. B. zur Sammlung, Institution erweitert (vgl. Abbildung 3). Hierdurch wird insbesondere eine Nachnutzung von Daten außerhalb des originären Sammlungskontexts ermöglicht.
- Verwendungskontext:** Durch die Definition eines fallspezifischen Integrationsmodells können Datenmodelle miteinander assoziiert werden. Durch eine Formulierung von Transformationsregeln werden Daten so umgewandelt und integriert, wie sie für eine weiterführende Untersuchung benötigt werden (vgl. Abbildung 4).

The Schema-Editor interface shows a 'Parser Grammatik' (Parser Grammar) and an 'Elementmodell' (Element Model) side-by-side. The 'Parser Grammatik' contains code snippets for parsing names and first names. The 'Elementmodell' shows a hierarchical structure of elements like 'Name', 'First', 'Last', and 'Subject'. Arrows point from the grammar to the model, and from the model back to the grammar, illustrating the bidirectional relationship between them.

Abbildung 3: Bildschirmausschnitt des Schema Editors

The Mapping Editor interface shows a 'Beispieltransformation' (Example Transformation) and 'Transformationsanweisungen' (Transformation Instructions). The transformation example shows a document with coordinates being converted into a placemark. The transformation instructions show a script using XSLT to reverse latitude and longitude values.

Abbildung 4: Bildschirmausschnitt des Mapping Editors

“Durchsuchen”: 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 in der Schema Registry 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 Mapping Registry aufgelöst.

Über die Möglichkeit der einfachen Suche über die Daten verzeichneter Kollektionen hinaus, können auf Basis der Funktionalität der generischen Suche weiterführende, fachspezifische Suchmaschinen implementiert werden (s. Abbildung 5).

The Verbundsuche search interface features a search bar and a results page. The results page includes a word cloud visualization where words related to historical biographies are highlighted in purple, such as 'Leipzig', 'historische', 'biographien', 'rezezieren', and 'Vasilia'. Logos for various partner institutions like MARBACH, WEIMAR, WOLFSBURG, and the Deutsches Literaturarchiv Marbach are visible on the right.

Abbildung 5: Fachwissenschaftliche Spezialsuche im Rahmen der generischen Suche

“Durchsuchen”: Historischer Biographien Use-Case

Der Use-case *Biographien* verbildlicht wie man eine historische Fragestellung anhand digitaler Werkzeuge bearbeiten kann.

Prosopographische historische Forschung orientiert sich immer noch stark an nationaler Geschichtsschreibung: religiöse, berufliche, gesellschaftliche Gruppen werden oft innerhalb der nationalen Grenzen, die selbst ein Konstrukt der Moderne sind, untersucht. Das Zusammenführen der Daten aus unterschiedlichen biographischen Datenbanken kann helfen dieses Problem zu lösen und biographische Recherchen über die nationalen Grenzen hinweg zu gestalten.

Zu diesem Zweck implementiert Dariah-DE derzeit das CosmoTool (vgl. Gradl & Henrich 2016b), welches auf die Unterstützung historischer Forschung an biographischen Daten abzielt. Das 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 in der Dariah-DE Schema / Mapping Registry
- 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

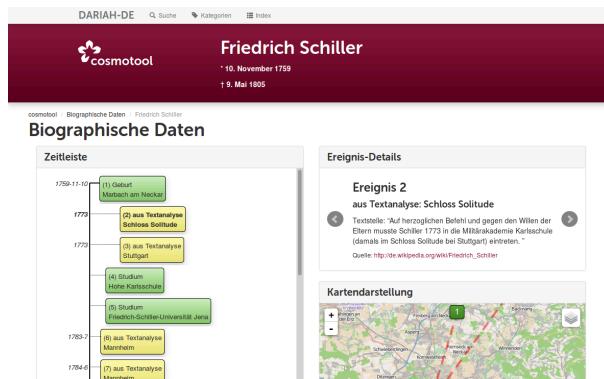


Abbildung 6: Bildschirmausschnitt des CosmoTools

Zusammenfassung

Insgesamt werden den TeilnehmerInnen im Rahmen dieses Workshops verschiedene Kenntnisse im Kontext der Sammlung, Modellierung und Suche geisteswissenschaftlicher Forschungsdaten vermittelt. Durch die Anwendung der entsprechenden Komponenten von Dariah-DE werden die in vorausgegangenen Referaten vorgestellten Ideen vertieft.

Die Begleitung des Workshops durch Forschungsfragen und Daten im Kontext biographischer Daten soll

den TeilnehmerInnen die praktische Anwendung der Komponenten deutlich machen. Idealerweise wird dadurch die Übertragbarkeit auf andere Daten und Fragen vermittelt, wodurch eine nachhaltige Zugänglichkeit wissenschaftlicher Forschungsdaten erreicht werden kann.

Kontaktdaten aller Beitragenden

Anna Aschauer, Leibniz-Institut für Europäische Geschichte (IEG), Querschnittsbereich, Alte Universitätstraße 19, 55116 Mainz

Forschungsinteressen: Pietismusforschung, Geschichte Russlands, Migration der religiösen Minderheiten in der Frühen Neuzeit, Digital Humanities.

Swantje Dogunke, Forschungsverbund Marbach Weimar Wolfenbüttel / Klassik Stiftung Weimar, Direktion Verwaltung, Abteilung Informationstechnik, Burgplatz 4, 99423 Weimar

Forschungsinteressen: Dokumentation im Museum, Museumsmanagement, digital curation, digitale Langzeitarchivierung, Digital Humanities

Tobias Gradl, Otto-Friedrich-Universität Bamberg, Lehrstuhl für Medieninformatik, An der Weberei 5, 96052 Bamberg

Forschungsinteressen: Forschungsdaten und Forschungsdatenmanagement, Digital Humanities, Datenintegration, Information Retrieval

Lisa Klaffki, Herzog August Bibliothek Wolfenbüttel, Abteilung 1, Lessingplatz 1, 38304 Wolfenbüttel

Forschungsinteressen: Archäologie der germanischen Provinzen, Bestattungssitten der römischen Kaiserzeit, Digital Humanities

Stefan Schmunk, Niedersächsische Staats- und Universitätsbibliothek Göttingen, Abt. Forschung und Entwicklung, Papendiek 14, 37073 Göttingen,
Forschungsinteressen: Forschungsdaten und
Forschungsdatenmanagement, Digitale
Geschichtswissenschaft, Virtuelle
Forschungsumgebungen, Digitale
Forschungsinfrastrukturen

Timo Steyer, Forschungsverbund Marbach Weimar Wolfenbüttel / Herzog August Bibliothek Wolfenbüttel, Abteilung 1, Lessingplatz 1, 38304 Wolfenbüttel

Forschungsinteressen: Digitale Editionen, Datenmodellierung und Metadaten, Digital Humanities

Zahl der möglichen Teilnehmerinnen und Teilnehmer

Die Zahl der möglichen Teilnehmer ist aus unserer Sicht nicht eingeschränkt. Einer sehr großen Zahl müsste ggf. durch mehrere Helfer in der Hands-On-Session entgegnet werden

Angaben zu einer etwa benötigten technischen Ausstattung

Es wird keine zusätzliche Ausstattung neben der üblichen Präsentationstechnik benötigt. Von den TeilnehmerInnen wird das Mitbringen eines eigenen Laptops für die aktive Teilnahme an der Hands-On-Session erwartet.

Fußnoten

1. Repository, Collection Registry, Schema / Mapping Registry und Generische Suche von Dariah-DE (vgl. Gradl & Henrich 2016a, Schmunk & Funk 2016)

Bibliographie

Gradl, Tobias / Henrich, Andreas (2016a): „Die Dariah-DE Föderationsarchitektur - Datenintegration im Spannungsfeld forschungsspezifischer und domänenübergreifender Anforderungen“, in: *Bibliothek - Forschung und Praxis* 2016 40 (2): 222–228 10.1515/bfp-2016-0027.

Gradl, Tobias / Henrich, Andreas (2016b): „Nutzung und Kombination von Daten aus strukturierten und unstrukturierten Quellen zur Identifikation transnationaler Lebensläufe“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* 129–132.

Gradl, Tobias / Lordick, Harald / Henrich, Andreas (2016): „Judaica recherchieren: Unterstützung bei der Realisierung forschungsspezifischer Suchlösungen durch die generische Suche“, in: *DHd 2016: Modellierung - Vernetzung - Visualisierung* 132–136.

Schmunk, Stefan / Funk Stefan (2016): „Das Dariah-DE- und das TextGrid-Repositorium: Geistes- und kulturwissenschaftliche Forschungsdaten persistent und referenzierbar langzeitspeichern“, in: *Bibliothek - Forschung und Praxis* 2016 40 (2): 213–221 10.1515/bfp-2016-0020.

Szöllösi-Janze, Margit / Panter, Sarah / Paulmann, Johannes (2015): „Mobility and Biography. Methodological Challenges and Perspectives“, in: *Jahrbuch für Europäische Geschichte / European History Yearbook* 16: 1–14 10.1515/9783110415162-001.