Entwicklung einer Wissensbasis für Lehr- und Lernmaterialien zu FAIRem Forschungsdatenmanagement und Data Science im Projekt DALIA (Data Literacy Alliance)

Canan Hastik[1], Frank Lange[2], Jan-Michael Haugwitz[3] und Peter Pelz[4]

Abstract: DALIA ist ein Projekt zur Entwicklung einer Wissensbasis für Data Literacy, die über einen semantischen Knowledge-Graph zugänglich ist. Die Plattform basiert auf etablierten Technologien des Semantic Web und wird aktuelle Methoden der künstlichen Intelligenz (KI) implementieren, um die effiziente Bereitstellung von Lehr- und Lernmaterialien in personalisierten Lernpfaden sicherzustellen. Suchergebnisse und Empfehlungen können generisch oder disziplinspezifisch sein und sich auf alle Kompetenz- und Erfahrungsstufen beziehen. Ferner ist die Integration von modernen und nutzungsfreundlichen Technologien, wie Frage-Antwort-Interaktion, vorgesehen. Knowledge-Graphen verbessern nicht nur die Interoperabilität und Sichtbarkeit der Ressourcen gemäß den FAIR-Prinzipien, sondern ermöglichen auch eine Vernetzung mit anderen fachspezifischen Modellen innerhalb der NFDI. Die Entwicklung erfolgt in enger Zusammenarbeit mit Expertengruppen und zukünftigen Nutzenden - Lernende und Lehrpersonen aller Disziplinen.

Keywords: Datenkompetenz, Wissensbasis, Open Educational Resources, Semantic Web, KI

1 Einleitung

In diesem Beitrag wird die DALIA Wissensbasis des vom BMBF bis 2025 geförderten Projektes „Knowledge-Graph der Data Literacy Alliance (Dalia) für FAIRe Datennutzung und -bereitstellung auf der Basis von Semantic-Web-Technologie" [Da23] hinsichtlich nachhaltiger Vernetzung von Anbietenden- und Nutzenden-Communitys von Lehr- und Lernmaterialien vorgestellt. Sie dient der disziplinübergreifenden Förderung der Datenkompetenz (Data Literacy) mittels Nutzung etablierter Technologien des Semantic Web wie Resource Description Framework (RDF), Ontologien und fachspezifischen Vokabularen, sowie Techniken des maschinellen Lernens, wodurch Datenaustausch durch die Anreicherung mit Metadaten maschinenlesbar und

[1] Technische Universität Darmstadt, Institut für Fluidsystemtechnik, Karolinenplatz 5, 64289 Darmstadt, canan.hastik@fst.tu-darmstadt.de, https://orcid.org/0000-0003-1729-4642
[2] RWTH Aachen University, IT Center, Seffenter Weg 23, 52074 Aachen & Institut für Anorganische Chemie, Landoltweg 1A, 52072 Aachen, f.lange@itc.rwth-aachen.de, https://orcid.org/0000-0002-9346-6031
[3] RWTH Aachen University, IT Center, Seffenter Weg 23, 52074 Aachen, haugwitz@itc.rwth-aachen.de, https://orcid.org/0009-0007-3576-3947
[4] Technische Universität Darmstadt, Institut für Fluidsystemtechnik, Karolinenplatz 5, 64289 Darmstadt, peter.pelz@fst.tu-darmstadt.de, https://orcid.org/0000-0002-0195-627X

kontextsensitiv gestaltet wird. Das Projekt arbeitet eng mit den Communitys der NFDI-Sektion Training & Education zusammen und soll nach Abschluss an den NFDI-Verein übergeben werden.

2 Was ist DALIA?

Datenkompetenz von Anfang an. Mit diesem Paradigma betonen Prof. Dr. Peter Pelz und Prof. Dr. rer. nat. Sonja Herres-Pawlis die Relevanz der Datenkompetenzförderung und damit verbunden der nachhaltigen Nutzung und Bereitstellung von Forschungsdaten [Pe21]. Das von beiden initiierte Projekt hat zum Ziel, einen nennenswerten Beitrag bei der Umsetzung dieses Paradigmas durch die Entwicklung einer Plattform für Lehr- und Lernmaterialien als Wissensbasis zu leisten. Durch die enge Verbindung zum NFDI e.V. und den Sektionen werden starke Synergien bei der Entwicklung von DALIA ermöglicht. Das Projekt leistet einen wichtigen Beitrag zur Steigerung der Datenkompetenz und der Etablierung einer Datenkultur in Forschung und Lehre. Darüber hinaus fördert DALIA den Kulturwandel, indem von den neuen Technologien und Trends Gebrauch gemacht wird. Dadurch werden Microservices für unterschiedliche Community-Bedarfe adressierbar und möglich.

Die verfügbaren Inhalte und Plattformen für Bildungsmaterialien zur FAIRer Datennutzung und Bereitstellung sind durch große Heterogenität gekennzeichnet. Materialien sind oft generisch und es mangelt an inhaltlich-fachlich-didaktischen Qualitätskriterien sowie formalen Kriterien, beispielsweise hinsichtlich der Nachnutzbarkeit dieser. Zudem sind verwendete Metadaten in Beschreibungsumfang und -tiefe, aber auch Qualität nicht ausreichend, um für Suchende für deren individuelle und ebenso heterogenen Bedarfe passende Materialien zu finden. DALIA soll als Single Point of Entry den einheitlichen Zugang zu einer Vielzahl von Repositorien ermöglichen und durch Kurationsprozesse zu einer Anreicherung von Metadaten und damit zu Erfolgen auch bei hoch spezifischen Suchanfragen führen. In DALIA kuratierte Metadaten können im Sinne eines Metadaten-Ökosystems wieder in die Repositorien zurückfließen und die Qualität für lokale Suchen verbessern. DALIA wird als Plattform konzipiert, die Lehr- und Lernmaterialien rund um den Themenkomplex Data Literacy in Form einer Knowledge-Base sichtbar, verfügbar und zugänglich macht. Diese wird, entsprechend zu aktuellen Entwicklungen in der Nationalen Forschungsdateninfrastruktur (NFDI) und dem FAIR-Konzept [Wi16] passend, als semantischer verknüpfter Knowledge-Graph umgesetzt. Zielsetzung ist es, bestehende edukative Materialien der NFDI-Konsortien, Datenkompetenzzentren, der FDM-Landesinitiativen und internationaler Stakeholder wie EOSC, OpenAIRE und RDA zu vernetzen, besser sichtbar und zugänglich zu machen. Die Verweise und Vorschläge der Knowledge-Base sind fachspezifisch und abhängig von Interesse und Bedarf der Lehrenden und Lernenden am relevanten Thema. Dabei kommen verschiedene Methoden der künstlichen Intelligenz, wie Semantic Reasoning, Machine Learning und Natural Language Processing (NLP) zum Einsatz. Diese leiten die

Interaktion der Nutzenden an und verbessern Suchergebnisse durch Analyse von Kriterien wie fachspezifische Data-Literacy-Kompetenzprofile [Le20; Pe23], Disziplin- und Zielgruppenspezifik.

2.1 Fragen wir ChatGPT…

Durch Klartextfragen kann aus der Menge von vorgegebenen Daten durchaus Neues entstehen. Fragen wir ChatGPT [Op23] (Wissensstand 2021), „Was ist Dalia?", bekommen wir derzeit folgende Antworten:

- „DALIA steht für "Data Analysis and Logical Inference Assistant". Es handelt sich dabei um eine computergestützte Analyse- und Inferenzplattform, die speziell für den Umgang mit großen Datenmengen entwickelt wurde. DALIA kombiniert verschiedene Techniken der Datenanalyse, des maschinellen Lernens und der logischen Inferenz, um Muster, Zusammenhänge und Erkenntnisse in den Daten zu identifizieren."

und

- „DALIA basiert auf fortgeschrittenen Algorithmen und Modellen des maschinellen Lernens, einschließlich Deep Learning und künstlicher Intelligenz. Die Plattform ermöglicht es den Benutzern, komplexe Abfragen und Analysen durchzuführen, um tiefere Einblicke in ihre Daten zu gewinnen. Sie bietet auch Funktionen zur Visualisierung der Ergebnisse und zur interaktiven Exploration der Daten."

ChatGPT und ähnliche Angebote liefern auf Trainingsdaten basierende logisch konstruierte Aussagen. Die Fragestellung bedingt dabei die Musterantworten, die aus statistischen Modellen und Algorithmen generiert werden. Die beiden oben zitierten spekulativen Aussagen zeigen eine Überbetonung von klangvollen Technologiebegriffen. Inhaltlich sind die generierten Antworten recht weit entfernt von DALIA, auch wenn die Antworten sehr überzeugend aussehen - zumindest, wenn Kenntnisse fehlen, Information und Daten kritisch zu hinterfragen und zu bewerten. DALIA ist keine spezialisierte Datenanalyseplattform für große Datenmengen, wie ChatGPT behauptet, stattdessen nutzt DALIA Ressourcen aus verschiedenen Quellen, um Nutzenden bei der individuell optimierten Suche passende Ergebnisse zu Data-Literacy-Themen, wie Forschungsdatenmanagement, Datenethik oder Datenpräsentation zu liefern.

3 Was hat DALIA vor

DALIA wird als Single Point of Entry, also als zentraler Einstiegspunkt konzipiert. Zielsetzung von DALIA ist es, bestehende oder in Entwicklung befindliche verteilte Plattformen und heterogene Materialsammlungen, wie das Training und Education Repositories der NFDI-Konsortien, z.B. das NFDI4Ing Education Repository [Ed23],

oder kuratierte Materialien auf Plattformen wie Zenodo [Mo15], über standardisierte Schnittstellen mit der Wissensbasis zu verbinden. Die Materialien aus verschiedenen Quellen umfassen Standards und Best Practices, Selbstlern- und Lehrmaterialien unterschiedlicher Medienformate, die durch Lehrende nachgenutzt werden können. Damit leistet DALIA einen wichtigen Beitrag zur Vernetzung und Zugänglichmachung heterogener Quellen für fachspezifische Lehr- und Lernmaterialien in den jeweiligen Community-eigenen Plattformen. Darüber hinaus sollen Mehrwerte für die Anbietenden von Lehr- und Lernmaterialien geschaffen werden, u.a. indem Inhalte verteilter Quellen durch hochwertige Metadaten und Vokabulare angereichert und Werkzeuge zur Qualitätsbewertung von Materialien entwickelt werden.

3.1 Quellen und Materialien

Als Quelle für Open Educational Resources (OER) dienen verschiedene Repositorien und vielfältige Quellen unterschiedlicher Anbieter, wie z. B. Audio- und Video-Portale von Hochschulen, welche grundlegende Qualitätsstandards erfüllen. Qualitätsmerkmale sind u. A. eine Mindestverfügbarkeit und die Möglichkeit der Nachnutzbarkeit. Außerdem wird der Diskurs zur Anwendung der FAIR-Prinzipien auf OER verfolgt und berücksichtigt. In DALIA zugänglichen Materialien umfassen vielfältige Formate, darunter sind Standards und Best Practices, Selbstlern- und Lehrmaterialien unterschiedlicher Medientypen (Text, Bild, Audio, Video) und vielfältigen Formaten wie u. A. Slides, Poster, Podcasts, Lehrfilme oder Tutorial-Videos.

3.2 Graphentechnologie und Wissensbasis

Für die Implementierung der Knowledge-Base wird auf etablierte Technologien des Semantic Web zurückgegriffen, insbesondere wird der Knowledge-Graph mittels RDF und Vokabularen/Ontologien modelliert. SPARQL dient als standardisierte Schnittstelle sowohl für das DALIA-Portal zur Kuratierung und zum Lernen, als auch für einen öffentlichen Endpunkt zur Datenabfrage des Knowledge-Graphen. Die Umsetzung als Knowledge-Graph dient nicht nur als Datengrundlage für Empfehlungs-Algorithmen der DALIA-Plattform, sondern bringt auch Vorteile für Ersteller und Kuratoren der Lehr- und Lernmaterialien. So werden Ressourcen im Knowledge-Graphen mit maschinenlesbaren Metadaten angereichert, was insbesondere deren Interoperabilität verbessert. Durch Zurverfügungstellung über den öffentlichen SPARQL-Endpunkt werden die Ressourcen Teil der Linked Open Data Cloud und erhalten somit eine bessere Sichtbarkeit und Nachnutzbarkeit im Sinne der FAIR-Prinzipien.

Um das edukative Material in den verteilten Repositorien zu beschreiben und auffindbar zu machen, wird ein ontologiebasiertes Informationsmodell (DALIA Core Model) mit einem Maximum an semantischer Expressivität und disziplinübergreifenden Interoperabilität entwickelt. Die Verwendung von RDF-basierten Technologien und des graphenbasierten Modells bieten die Repräsentation des Wissens über die Domäne und

der Beziehungen zwischen diesen, was wiederum den Such- und Empfehlungsdienst unterstützt.

3.3 Nutzerbedarfe stehen im Vordergrund

Als Community-zentrierter Ansatz legt DALIA Wert darauf, aktuelle Trends wie Chatbots zur Interaktion von Nutzenden bei der Entwicklung angemessen einzusetzen und den Angebotscharakter (Affordance) zu berücksichtigen. Mit dieser Technologie ist es möglich, Fragen und Aufgabenstellungen in Prosa zu formulieren aber auch Programmieren, Testen und Konvertieren von Quellcode in der Softwareentwicklung zu unterstützen, aus Prosa Quelltext zu generieren und diesen in eine Vielzahl von Programmiersprachen zu konvertieren. Die KI ist darüber hinaus beim Generieren von API-Aufrufen, Erzeugen von Datenbankabfragen, der Fehlersuche und dem Testen recht fortgeschritten.

Das DALIA Core Model wird sich sukzessive mit weiteren fach- und disziplinspezifischen Modellen innerhalb der NFDI vernetzen. Perspektivisch können NutzerInnen über eine Vorauswahl der zur Verfügung gestellten Lernangebote in Form von Kursen und Lernpfaden einen individuellen Zugang zur Plattform und somit DALIA zu einem maßgeschneiderten personalisierten Angebot konfigurieren. Darüber hinaus sind noch weitere Anwendungsmöglichkeiten vorgesehen, wie beispielsweise die Integration von Abfrage-Wizards oder rudimentäres Natural Language Question Answering, sodass Forschende, Lehrende und Lernende individuelle Anfragen gemäß ihrem jeweiligen Bedarf formulieren können, wie

- „Ich möchte meine Messdaten FAIR ablegen, bitte liefere mir eine kommentierte Best-Practice-Anleitung"
- „Ich möchte Daten anderer nutzen, bitte liefere mir Metriken und Werkzeuge zur Beurteilung der formalen und inhaltlichen Datenqualität"
- „Ich möchte meinen Lernenden einen Lernpfad in DALIA vorschlagen, den Lernfortschritt tracken und für einzelne Lerneinheiten Leistungsnachweise vergeben"
- „Ich möchte für meinen Beruf relevante Kompetenz entwickeln, bitte schlage mir einen Lernpfad vor"
- „Bitte schlage mir weitere Themen vor, die für mich relevant sind."

4 Nächste Schritte

Nach der allgemeinen Erhebung und Analyse der bestehenden Angebote im Bereich Data Literacy wird der Ansatz verfolgt, zielgruppenorientiert die Nutzendenbedarfe zu identifizieren. Um eine bestmögliche Nutzerorientierung zu erreichen, werden in

aufeinanderfolgenden Fokusgruppen Workshops zusammen mit der NFDI-Sektion Training & Education kontinuierlich weitere Anforderungen und Nutzendenperspektiven systematisch erhoben und zielgerichtet in die Entwicklung von DALIA integriert. Ein Vorteil dieser Herangehensweise ist, dass die Community aktiv teilnehmen und teilhaben kann und selbst Einfluss hat, das „Endprodukt" DALIA mitzugestalten. Daher ist das Community-Building und -management ein wesentlicher Aspekt hinsichtlich des Monitorings von Zielgruppen, Bedarfen und Beständen. Außerdem werden aus den NFDI-Communitys heraus generische und fachspezifische modulare skalierbare Vermittlungs- und Schulungskonzepte, aber auch Zertifikatskurse entwickelt, die in DALIA Anwendung finden sollen.

5 DALIA Kommunikationskanäle

Aktuelles über das DALIA Projekt, zu technologischen Entwicklungen und Community-Themen aber auch Terminankündigungen werden im DALIA Projekt Newsletter [Ne23], in dem NFDI-DALIA-Rocket Chat Kanal und in Social Media veröffentlicht und verbreitet.

6 Danksagung

Dieses Projekt mit dem Förderkennzeichen 16DWWQP07A wurde gefördert vom Bundesministerium für Bildung und Forschung (BMBF) und der Fördermaßnahme aus der Aufbau- und Resilienzfazilität der EU.

7 Bibliographie

[Da23] DALIA: Knowledge-Base für „FAIR data usage and supply" als Knowledge-Graph, https://www.fst.tu-darmstadt.de/forschung_fst/zusammenarbeit_in_der_forschung/dalia/dalia_ueberblick.de.jsp, Stand: 29.05.2023.

[Ed23] Education Repository, NFDI4Ing, https://git.rwth-aachen.de/nfdi4ing/education, Stand: 29.05.2023.

[Le20] Lemaire, M. et al. (2020), Das DIAMANT-Modell 2.0: Modellierung des FDM-Referenzprozesses und Empfehlungen für die Implementierung einer institutionellen FDM-Servicelandschaft, https://doi.org/10.25353/ubtr-xxxx-f5d2-fffb.

[Mo15] Molloy, L. (2015), Research data management (RDM) open training materials, https://zenodo.org/communities/dcc-rdm-training-materials/?page=1&size=20, Stand: 29.05.2023.

[Ne23] DALIA Newsletter, https://dalia.pages.rwth-aachen.de/newsletter.

[Op23] OpenAI Introducing ChatGPT, https://openai.com/blog/chatgpt, Stand: 29.05.2023.

[Pe21] Pelz, P. et al. (2021), Sektionskonzept Training & Education zur Einrichtung einer Sektion im Verein Nationale Forschungsdateninfrastruktur (NFDI) e.V., https://zenodo.org/record/5599770.

[Pe23] Petersen, B. et al. (2023), Lernzielmatrix zum Themenbereich Forschungsdatenmanagement (FDM) für die Zielgruppen Studierende, PhDs und Data Stewards, https://doi.org/10.5281/zenodo.8010617.

[Wi16] Wilkinson, M.D. et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018, 2016.