Vernetzung von Kulturdaten in Sachsen: Auf dem Weg zum DIKUSA-Forschungsdatenregister als Schlüssel zur Datenintegration

Goldhahn, Dirk
goldhahn@saw-leipzig.de
Sächsische Akademie der Wissenschaften, Deutschland
ORCID: 0000-0003-1681-567X

Naether, Franziska
naether@saw-leipzig.de
Sächsische Akademie der Wissenschaften, Deutschland
ORCID: 0000-0003-4652-6836

Mühleder, Peter
muehleder@saw-leipzig.de
Sächsische Akademie der Wissenschaften, Deutschland
ORCID: 0000-0001-6593-5673

DIKUSA (Sächsische Akademie der Wissenschaften zu Leipzig 2024a, Goldhahn et. al. 2023) ist ein Verbundprojekt aus den Bereichen Geschichte und Digital Humanities der sechs außeruniversitären geisteswissenschaftlichen Forschungseinrichtungen in Sachsen, koordiniert durch das KompetenzwerkD (Sächsische Akademie der Wissenschaften zu Leipzig 2021), mit dem Ziel, eine digitale Infrastruktur zu schaffen, welche es ermöglicht, die in individuellen Forschungsprojekten entstehenden digitalen Daten nachhaltig zugänglich und recherchierbar zu machen. Damit wird die Kompetenz der Häuser für die digitale Erfassung von Archivmaterial und Objektdaten bzw. Metadaten, deren Integration bzw. Verlinkung, Visualisierungen sowie der Abgleich mit Normdatensätzen gestärkt.

Fokus dieses Posters ist dabei der Weg der Forschungsdaten in Einzelschritten – dargelegt anhand von Beispieldaten eines Projektpartners – von der Arbeit im Projekt über eine Konvertierung und Integration hin zu einem zentralen Forschungsdatenregister mit Referenzierungs- und Reconciliation-Diensten als Datenhub für Kulturdaten sächsischer DH-Projekte.

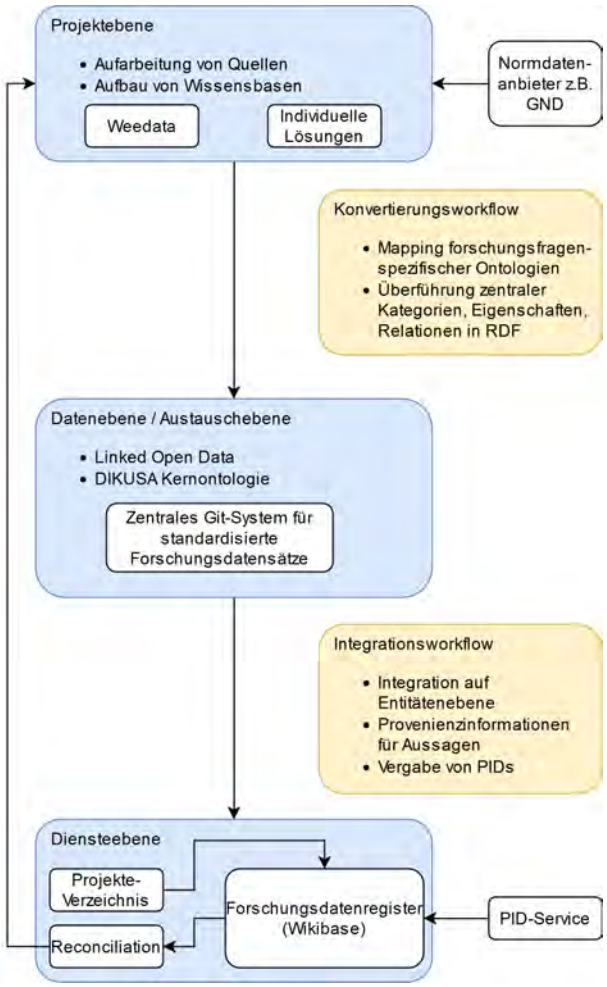

Abb. 1: Aufbau der DIKUSA-Dateninfrastruktur

1. Projektebene:

Ausgangspunkt der vorgestellten Integrationsbemühungen sind Forschungsprojekte der sächsischen Forschungsinstitute im Bereich der digitalen Geschichtswissenschaften, die im Rahmen des DIKUSA-Verbundvorhabens durchgeführt werden.

In den Projekten erfolgt typischerweise eine Aufarbeitung von Quellenmaterialien zur Erstellung von Wissensbasen . Die dafür eingesetzten Technologien sind abhängig von den jeweiligen Anforderungen der Forscher:innen und daher stark heterogen (bspw. Weedata (Sächsische Akademie der Wissenschaften zu Leipzig 2024a), Plone (Plone Foundation 2024), und individuelle Datenbanklösungen). Darüber hinaus erfolgt der Aufbau dieser strukturierten Datensammlungen unter Verwendung forschungsfragenspezifischer Datenmodelle. Ein wichtiger Schritt in diesem Prozess ist die Vergabe von Normdatenidentifiern in den jeweiligen Wissensbasen zu Personen, Orten usw. Hierfür haben wir im Rahmen des DIKUSA-Projekts eine Schnittstelle (Dikúdex) entwickelt, die den parallelen Abgleich mit mehreren Normdatenanbietern vereinfacht und in die Datenbanktools des Projekts integriert werden konnte (Sächsische Akademie der Wissenschaften zu Leipzig 2024c).

Die Forscher:innen in den Projekten werden bei der Datenerfassung angehalten, GND, Wikidata, Geonames und HOV-Identifikatoren (Institut für Sächsische Geschichte und Volkskunde (ISGV) 2024) zu vergeben.

1. Datenebene, Austauschebene (LOD):

Um einen Austausch von Daten und eine sich anschließende automatisierte Datenintegration zu ermöglichen bzw. zu erleichtern, bedarf es einer gemeinsamen Repräsentation der erfassten Daten. Aus diesem Grund wurde die DIKUSA Kernontologie (Goldhahn et. al. 2024) entwickelt, die eine einheitliche Darstellung der zunächst heterogenen Daten ermöglicht.

Die Entwicklung orientierte sich dabei an:

- typischen Forschungsfragen der Partnerinstitutionen und den vorkommenden zentralen Kategorien, Aussagen und Beziehungen (Sächsische Akademie der Wissenschaften zu Leipzig 2024d),
- allgemeinen Abhandlungen zum Thema Datenmodellierung (Flanders et. al. 2015; Tomasi 2018),
- Lessons Learned und Best Practices aus anderen Projekten (exemplarisch Hyvönen 2021) und
- an bestehenden Ontologien wie SARI (Universität Zürich 2020)
- an generellen Prinzipien wie FAIR-Data.

Die entstandene Ontologie basiert auf den Linked-Open-Data-Technologien RDF (W3C 2024a) und OWL (W3C 2024b). Darüber hinaus wird RDF-Star (W3C 2023) eingesetzt, um Aussagen über Tripel zu ermöglichen und somit flexibel jeder beliebigen Aussage Provenienzinformationen anfügen zu können. Es wurde ein SHACL-basiertes Tool entwickelt (W3C 2017), um die aus den Projekten entstehenden RDF-Datensätze zu verifizieren (KompetenzwerkD 2022). Die Bereitstellung und Versionierung der konvertierten Datensätze erfolgt in einem zentralen Git-basierten System.

Im nächsten Schritt findet eine Datenintegration auf Entitätenebene der bisher isolierten Datensätze statt. Diese erfolgt auf Basis der ausgezeichneten Normdaten. Alle Einträge des Registers werden mit einer eigenen PID versehen, wobei aktuell eine Nutzung der Handle.Net-Registry (DONA Foundation 2023) umgesetzt wird. Aussagen werden stets mit Provenienzinformationen zu den bereitstellenden Projekten versehen; eine tiefergehende Integration, Kuratierung oder das Auflösen von widersprüchlichen Aussagen sind derzeit nicht geplant. Update-Prozesse werden so einfach wie möglich gestaltet. Der Datensatz eines Projekts wird stets als Ganzes importiert bzw. geupdatet. Entitäten bleiben jederzeit erhalten, Aussagen werden vollständig ersetzt.

Der beschriebene Datenintegrationsprozess wird derzeit implementiert. Er wird im Rahmen der Posterpräsentation anhand von Beispieldaten eines DIKUSA-Teilprojekts dargelegt.

1. Diensteebene:

Um den nachhaltigen Zugang und die Recherchierbarkeit der Daten zu gewährleisten, entwickeln wir im Rahmen des KompetenzwerkD eine Reihe von Diensten, die langfristig betrieben werden sollen. Zentral ist hier das Forschungsdatenregister , das den öffentlichen Zugang zu den Forschungsdaten bereitstellt. Die Umsetzung erfolgt auf Basis von Wikibase (Wikimedia Deutschland 2024). Das Wikibase-System erlaubt eine übersichtliche Darstellung strukturierter und verknüpfter Daten und ermöglicht deren Recherchierbarkeit – zentrale Anliegen des Registers. Das Forschungsdatenregister beinhaltet für alle Aussagen Provenienzangaben, darunter Informationen zu den Forschungsprojekten, welche die Daten bereitgestellt haben. Diese Informationen werden in einem dedizierten Verzeichnis von Forschungsprojekten zusammengetragen. Dessen Datenmodell ist an die NFDI core ontology (Bruns et. al. 2024) angelehnt, wodurch ein zukünftiger Datenaustausch ermöglicht wird. Zusätzlich wird das Forschungsdatenregister als Basis für einen Reconciliationdienst dienen und in die bestehende Dikúdex-Infrastruktur integriert werden. Dies soll eine Referenzierbarkeit sicherstellen und eine Integration zukünftiger Forschungsdaten ermöglichen.

Durch diese Schritte entsteht eine Infrastruktur, die es den geisteswissenschaftlichen Forschungseinrichtungen in Sachsen erlaubt, digitale Forschungsprojekte durchzuführen, die Daten miteinander zu verknüpfen, und an größere Forschungsdateninitiativen wie die NFDI-Konsortien anschließbar zu machen, ohne individuelle Lösungen dafür konzipieren zu müssen. Durch die Sonderstellung des KompetenzwerkD im sächsischen Kontext kommt dem Unterfangen dabei Pioniercharakter zu.

Weiterhin soll diese Infrastruktur zukünftig technische Basis für zusätzliche Dienste sein. Insbesondere ein Angebot zum Einpflegen von Entitäten aus dem Kontext der DH-Forschung in Sachsen in die GND - im Rahmen einer eigenen GND-Agentur, über Dienste der NFDI-Konsortien oder weiterer Bestrebungen (Balzer et. al. 2019) - sei hier genannt. Die so entstehende Infrastruktur und ihre Dienste rund um Datenerfassung und Datenintegration sollen dadurch für zukünftige Forschungsprojekte attraktiv werden, um eine langfristige Nutzung und Weiterentwicklung sicherzustellen. Basis dafür ist auch die dauerhafte Förderung des KompetenzwerkD durch den Freistaat Sachsen, die einen Weiterbetrieb und eine Weiterentwicklung auch über die DIKUSA-Projektförderung hinaus erst ermöglicht.

Bibliographie

Balzer, D., Fischer, B.K., Kett ‚J., Laux, S., Lill, J.M., Lindenthal, J., Manecke, M.,Rosenkötter, M.,Vitzthum, A. (2019). „Das Projekt ‚GND für Kulturdaten' (GND4C)". *o-bib. Das offene Bibliotheksjournal* / Herausgeber VDB 6 (4): 59–97. https://doi.org/10.5282/o-bib/2019H4S59-97.

Bruns, O., Tietz, T., Posthumus, E., Waitelonis, J., Sack ‚H. (2024). „NFDIcore Ontology. Revision: v2.1.0".

https://ise-fizkarlsruhe.github.io/nfdicore/ (zugegriffen: 26.11.2024).

DONA Foundation (2023). „Handle.Net Registry“. https://handle.net (zugegriffen: 26.11.2024).

Flanders, J., Jannidis, M. (2015), „Knowledge Organization and Data Modeling in the Humanities“. https://nbn-resolving.org/html/urn:nbn:de:bvb:20-opus-111270 .

Goldhahn, D., Mühleder, P., Naether, F. (2023). „There is no ‚I‘ in ‚Infrastructure‘: Creating a shared data-centric DH Infrastructure for Cultural Heritage Research in Saxony/Germany“. *DH*, DOI: 10.5281/zenodo.8107515.

Goldhahn, D., Naether, F., Mühleder, P. (2024). „DIKUSA core ontology v1.0“. RADAR4Culture, DOI: 10.22000/xxDiXtLrXbLCedbS (Metadaten und Datenpaket).

Hyvönen, E. (2021). „How to Create a National Cross-domain Ontology and Linked Data Infrastructure and Use It on the Semantic Web“. In *Keynote presentation at DCMI 2021 conference (Dublin Core Matadata Initiative)*. Paper: https://seco.cs.aalto.fi/publications/2022/hyvonen-infra-2022.pdf.

Institut für Sächsische Geschichte und Volkskunde (ISGV) (2024). „Historisches Ortsverzeichnis von Sachsen“. https://hov.isgv.de/ (zugegriffen: 26.11.2024).

KompetenzwerkD (2022). „DIKUSA RDF Validator“. https://github.com/KompetenzwerkD/dikusa_rdf_validator (zugegriffen: 26.11.2024).

Plone Foundation (2024). „Plone CMS: Open Source Content Management“. https://plone.org/ (zugegriffen: 26.11.2024).

Sächsische Akademie der Wissenschaften zu Leipzig (2021). „KompetenzwerkD: Sächsisches Forschungszentrum und Kompetenznetzwerk für Digitale Geisteswissenschaften und Kulturelles Erbe“. https://www.saw-leipzig.de/de/akademie-digital/akademie-digital/kompetenzwerkd-saechsisches-forschungszentrum-und-kompetenznetzwerk-fuer-digitale-geisteswissenschaften-und-kulturelles-erbe (zugegriffen: 26.11.2024).

Sächsische Akademie der Wissenschaften zu Leipzig (2024a). „DIKUSA - Vernetzung Digitaler Kulturdaten in Sachsen“. https://dikusa.saw-leipzig.de (zugegriffen: 26.11.2024).

Sächsische Akademie der Wissenschaften zu Leipzig (2024b). „DIKUSA - Vernetzung Digitaler Kulturdaten in Sachsen: Weedata“. https://dikusa.saw-leipzig.de/toolbox/weedata/ (zugegriffen: 26.11.2024).

Sächsische Akademie der Wissenschaften zu Leipzig (2024c). „DIKUSA - Vernetzung Digitaler Kulturdaten in Sachsen: Reconciliation Service". https://dikusa.saw-leipzig.de/toolbox/dikudex/ (zugegriffen: 26.11.2024).

Sächsische Akademie der Wissenschaften zu Leipzig (2024d). "DIKUSA - Vernetzung Digitaler Kulturdaten in Sachsen: Projekte“. https://dikusa.saw-leipzig.de/projekte/ (zugegriffen: 26.11.2024).

Tomasi, F. (2018). „Modelling in the Digital Humanities: Conceptual Data Models and Knowledge Organization in the Cultural Heritage Domain“. *Historical Social Research*, Supplement, 31, 170-179, https://doi.org/10.12759/hsr.suppl.31.2018.170-179.

Universität Zürich (2020). „Swiss Art Research Infrastructure project - SARI: Documentation“. https://docs.swissartresearch.net/ (zugegriffen: 26.11.2024).

W3C (2017). „Shapes Constraint Language (SHACL)“. https://www.w3.org/TR/shacl/ (zugegriffen: 26.11.2024).

W3C (2023). „RDF-star W3C Community Group Draft Report“. https://w3c.github.io/rdf-star/cg-spec/editors_draft.html (zugegriffen: 26.11.2024).

W3C (2024a). „Resource Description Framework (RDF)“. https://www.w3.org/RDF/ (zugegriffen: 26.11.2024).

W3C (2024b). „Web Ontology Language (OWL)“. https://www.w3.org/OWL/ (zugegriffen: 26.11.2024).

Wikimedia Deutschland (2024). „Wikibase“. https://www.wikimedia.de/projekte/wikibase/ (zugegriffen: 26.11.2024).