DIKUSA

Datenmanagementplan

Erstellt durch: KompetenzwerkD

Datum: 09.06.2023

Version: 1.2

Administrative Angaben

Projekttitel: Vernetzung digitaler Kulturdaten in Sachsen – Aufbau einer technischen Infrastruktur für die Forschung zu Mobilität, Migration und Transformation von Orten, Personen und Artefakten (in zeitlicher und räumlicher Perspektive) – DIKUSA
Antragsnummer (SAB): 100611724
Kontonummer (SAB): 3000954790; Bewilligung vom 24.02.2022
Kundennummer (SAB): 2002094846
Clusternummer (SAB): 4192
Fördergeber: Landesmittel des Freistaats Sachsen, Titelgruppe 70. Die Zuwendung stammt aus Steuermitteln. Diese Steuermittel werden auf der Grundlage des vom Sächsischen Landtag beschlossenen Haushaltes zur Verfügung gestellt. Details siehe Projektskizze, Vollanträge der 6 Partner und 6 Zuwendungsbescheide der SAB.
Zeitraum: 01.02.2022-31.12.2025 (Teilprojekte können beim Beginn leicht abweichen)
Mitarbeitende/Datenurheber: siehe DMPs der Teilprojekte; aus DI, GWZO, HAIT, ISGV, SAW, SI, nachfolgenden auch als die „Partner“ bezeichnet
Koordinierung: KompetenzwerkD an der SAW - Dr. Dirk Goldhahn, Mag. Peter Mühleder, PD Dr. Franziska Naether
Antragsteller: Sächsische Akademie der Wissenschaften zu Leipzig, vertreten durch den Präsidenten Prof. Dr. Hans Wiesmeth und den Generalsekretär Dr. Christian Winter
Kontakt: kompetenzwerkd@saw-leipzig.de

Kurzbeschreibung/Abstract Gesamtvorhaben:

Im Forschungsalltag stellt die niedrigschwellige digitale Wissenserfassung und -integration die außeruniversitären geisteswissenschaftlichen Forschungseinrichtungen Sachsens vor unterschiedlich große Herausforderungen. Ziel dieses Verbundvorhabens ist es, eine entsprechende technische Infrastruktur zu entwickeln und die Kompetenz der Häuser für die digitale Aufnahme von Archivmaterial und Objektdaten, deren Verlinkung, Visualisierungen sowie den Abgleich mit Normdatensätzen zu ermöglichen. Dafür wird zusammen mit dem KompetenzwerkD und der SLUB eine technische Basis bestehend aus mehreren Komponenten erstellt: eine zentrale Wissensbasis mit Referenzierungs- und Reconciliation-Diensten als Datenhub, eine „Toolbox“ mit nachnutzbaren Softwarekomponenten für die Datenarbeit und Dienste für die Visualisierung auf historischem Kartenmaterial. Dies ermöglicht zugleich eine attraktive Außendarstellung der Teilprojekte. Die technische Infrastruktur wird im Rahmen von sechs Teilprojekten aus dem Bereich der Geisteswissenschaften und der Digital Humanities entwickelt, die Kulturdaten über Orte, Personen und Artefakte in der zeitlichen Perspektive seit dem Mittelalter bis zur Gegenwart in spezifischen Fragestellungen in den Blick nehmen.

Erster Pressebericht: https://www.saw-leipzig.de/de/aktuelles/neues-verbundprojekt

Beschreibung der Daten und Metadaten

Das Gesamtvorhaben besteht neben der Forschungstätigkeit in den Teilprojekten aus einem übergeordneten Vorhaben, das die in den Wissensbasen der Partner enthaltenen Daten zusammenführt und darauf aufbauend Dienste bereitstellt („KOWED Toolbox“). Diese Dienste sollen auch nach dem Ende von „DIKUSA“ allen Partnern zur Verfügung stehen.

In den Teilprojekten werden die Daten in Wissensbasen entsprechend von Datenmodellen, die forschungsfragen-spezifisch ausgearbeitet wurden, erfasst. Details behandelt der spezifische DMP.

Zur Vereinfachung der Integration werden die Daten in ein gemeinsames Datenmodell überführt. Dieses Modell ist orientiert an bestehenden Ontologien wie den Swiss Art Research Infrastructures Ontologies (SARI). Es umfasst in erster Linie Kategorien wie Personen, Orte, Objekte/Artefakte, Werke, Gruppen und Ereignisse, die einen stark vernetzten Wissensgraphen bilden. Um die nötige Interoperabilität sicherzustellen, werden Klassen und Properties auf Elemente aus CIDOC-CRM zurückgeführt (Work in Progress). Das Schema samt Dokumentation (in Entwicklung) wird auf GitHub bereitgestellt: https://github.com/KompetenzwerkD/dikusa-core-ontology.

Die Bereitstellung der im Projekt erstellten Datensätze erfolgt durch die Partner im RDF-star-Format unter Verwendung einer der üblichen, verbreiteten Serialisierungen wie Turtle oder JSON-LD. Abgelegt werden diese Daten in einem gemeinsamen vom Kompetenzwerk betriebenen Git-basierten System. Derzeit erfolgt die Planung dieses Systems als Teil der zentralen Infrastruktur und des Datenimports in diese. Derzeit gehen wir von einer Umsetzung mittels Gitea aus. Die Bereitstellung soll über einen Server der SAW erfolgen, welcher gerade im Rahmen des Projekts angeschafft wird.

Das Schema des gemeinsamen Datenmodells wird als RDF-Schema beschrieben. OWL-Syntax findet Verwendung zur Auszeichnung inverser Relationen. RDF-star wird eingesetzt, um Provenienzangaben an beliebige Aussagen im Wissensgraph annotieren zu können. Dies ermöglicht die Nachverfolgung des Ursprungs von Aussagen und erhöht somit die Nachnutzbarkeit der von den Partnern bereitgestellten Daten.

Ergänzt wird das Schema durch SHACL-Constraints, die eine Validierung der von den Partnern bereitgestellten Instanzdaten gegen das Schema ermöglichen.
Hierzu wird seitens des KompetenzwerkD ein Webservice bereitgestellt werden, welcher im Rahmen des Datenimports in das Git-System Verwendung findet. Dieser Service wird auf dem bereits jetzt bereitgestellten DIKUSA-Validierungstools basieren:
https://github.com/KompetenzwerkD/dikusa_rdf_validator

Die Entitäten in den Wissensbasen werden, wann immer möglich, mit Normdaten aus Dienstes wie GND, Wikidata, HOV oder Geonames verbunden und stets mit diesen Informationen (sie sind Teil des zentralen Datenmodells) veröffentlicht.
Zur Unterstützung dieser Auszeichnung wird vom KompetenzwerkD ein Reconciliation-Dienst angeboten, der im Rahmen der lokalen Dateneingabe bzw. Datenerfassung durch die Projektpartner genutzt werden kann, die Entitäten entsprechend auszuzeichnen.
Dieser Dienst wird eine im Vergleich zur Reconciliation-API erweiterte Funktionalität bereitstellen, um Einträge mehrerer Normdaten-Dienste übersichtlich zu präsentieren.

In diesem Zusammenhang wird ein projektspezifischer Normdatendienst bzw. Index aufgesetzt, der allen in den Teilprojekten vorkommenden Entitäten eine eindeutige gemeinsame ID zuweist,

unabhängig davon, ob ein Eintrag in externen Normdatenbeständen vorhanden ist. Dieser Datensatz wird ebenfalls in den Reconciliation-Dienst integriert und ermöglicht so die Verknüpfung mit den Daten der anderen Teilprojekte. Somit wird eine eindeutige Referenzierbarkeit auch bei bisher nicht in Normdatensätzen bekannten Entitäten gewährleistet. Eine Bereitstellung des beschriebenen eigenen Normdatensystems über die Projektgrenzen hinaus wird geprüft.

Die Teilprojekte werden angehalten, im Rahmen der Modellierung ihrer Datenmodelle die Verwendung bestehender Vokabulare zu prüfen. Eine Abstimmung zu den verwendeten Vokabularen erfolgt im Projekt mit dem Ziel der Vereinheitlichung und der Übernahme ins zentrale Datenmodell.
Sämtliche im Rahmen des Projektes verwendeten Vokabulare werden als Teil der Dokumentation der Ontologie mit betrachtet und alle im Projekt entwickelten Vokabulare werden zusammen mit der Ontologie veröffentlicht.

Ethische und rechtliche Aspekte

Seitens der Teilprojekte werden nur vollständig freie und rechtlich unproblematische Daten bereitgestellt, da die Datensätze frei verfügbar gemacht werden.

Speicherung, Archivierung und Sicherung der Daten

(Wo werden Daten gespeichert? Wie wird mit Daten gearbeitet? Wo und wie erfolgen Backups? Welche Daten werden wo und wie lange verfügbar sein?)

Speicherung, Archivierung und Sicherung der Daten der Teilprojekte werden im jeweiligen spezifischen Datenmanagementplan erörtert.
Daten aus den in den Teilprojekten erstellten Wissensbasen werden seitens der Teilprojekte für eine zentrale Verwendung bereitgestellt. Hierzu werden die in forschungsspezifischen Ontologien vorliegenden Daten in ein gemeinsames Datenmodell (https://github.com/KompetenzwerkD/dikusa-core-ontology) überführt. Die Bereitstellung erfolgt im RDF-Format unter Verwendung einer der üblichen, verbreiteten Serialisierungen. Die Daten werden über ein vom KompetenzwerkD betriebenes Git-basiertes System bereitgestellt, in das die Partner ihre Daten einspeisen.
Derzeit erfolgt die Planung dieses Systems als Teil der zentralen Infrastruktur und des Datenimports in diese. Wir gehen von einer Umsetzung mittels Gitea aus. Die Bereitstellung soll über einen Server der SAW erfolgen, welcher gerade im Rahmen des Projekts angeschafft wird.
Für jede projektbezogene Wissensbasis wird dabei ein eigenes Repositorium angelegt, zusammengefasst unter einem gemeinsamen Organisations-Account. Im Rahmen des Uploads erfolgt eine Validierung der Daten gegen das gemeinsame Schema, um die syntaktische Korrektheit der Daten zu gewährleisten und die Projektpartner gegebenenfalls über nötige Schritte zur Behebung von Problemen mittels Mailversand aus dem System heraus zu informieren. Das hierfür benötigte Werkzeug befindet sich in Entwicklung, ist aber bereits abrufbar:
https://github.com/KompetenzwerkD/dikusa_rdf_validator
Es wird die Basis für einen Webservice darstellen, der in den Prozess integriert wird. Dieser wird ebenfalls auf einem SAW-internen Server laufen.

Die erste Bereitstellung von Daten durch die Partner erfolgt - sobald die digitale Erfassung im Projekt eingesetzt werden kann und erste Daten erfasst wurden - zeitnah und danach fortlaufend bei entscheidenden Änderungen der Datenbasis wie umfangreichen Erweiterungen oder Korrekturen, jedoch mindestens einmal pro Halbjahr. Die Nachvollziehbarkeit von Änderungen wird durch Verwendung von Git-Features gewährleistet.

Die so bereitgestellten Datensätze werden in eine zentrale Wissensbasis eingefügt, welche vom KompetenzwerkD betrieben wird und in erster Linie als Datenbasis für den Reconciliation-Dienst

dient. Dieses System befindet sich in Entwicklung und wird voraussichtlich auf Neo4J aufbauen. Betrieben werden die Wissensbasis und die dazugehörigen Dienste auf einem dediziertem Server der SAW Leipzig, auf welchem Docker Compose zur Orchestrierung aller Bestandteile eingesetzt wird.
Alle nötigen Tools (eigener Code und externe Software) sowie Dokumentation zum Aufsetzen des Servers werden unter freier Lizenz mittels Git, GitHub oder docker hub bereitgestellt bzw. sind dort bereits vorhanden. Auch die Datengrundlage kann und wird aus Git eingespielt werden.

Die zentrale Wissensbasis wird die einzelnen Datensätze der Teilprojekte beinhalten, wobei nur die grundlegenden Kategorien und Properties im Fokus stehen, die seitens des Reconciliation Service angeboten werden (insbesondere Personen und Orte). Das Mapping der im RDF-Format vorliegenden Daten auf das in Neo4J verwendete Datenmodell erfolgt beim Import in die Wissensbasis nach der Validierung durch das bereitgestellte Tool:
https://github.com/KompetenzwerkD/dikusa_rdf_validator

Mittels eines persistenten internen Index, welcher die Basis für einen eigenen Normdatensatz darstellt, werden Konzepte der Teildatensätze miteinander verbunden. Eine Identität zweier Konzepte ist dabei initial gegeben, wenn eine Übereinstimmung bezüglich mindestens eines Normdatenregisters besteht. Um dies zu ermöglichen, werden die Einträge des Index eine der vom Reconciliation-Dienst bereitgestellten Datenquellen sein.
Die Wissensbasis wird jeweils nur den aktuellsten Release der Wissensbasen der Teilprojekte enthalten. Ein Update erfolgt jeweils, wenn eine Version einer Wissensbasen durch den jeweiligen Projekt- Partner als neuen Release im Git-System getaggt wird.
Da der zentrale interne Index nicht nur vom aktuellen Zustand der Wissensbasen abhängt und nicht ohne weiteres rekonstruiert werden kann, ist das Vorhandensein einer funktionierenden Backup-Strategie von großer Bedeutung. Nach dem Export (in ein noch festzulegendes Datenformat) wird dieser Vorgang mindestens einmal wöchentlich in der von der SAW Leipzig bereitgestellten Backup-Architektur durchgeführt, wobei ergänzend die einzelnen Wissensbasen aus dem Git-System als RDF-Datei gesichert werden.
Außer des Index können alle Komponenten der zentralen Infrastruktur (Tools und Daten) aus den entsprechenden gut dokumentierten Git-Repositorien wiederhergestellt werden.
Details zur Umsetzung der Backup-Strategie werden im Rahmen der Umsetzung und Bereitstellung der Tools und Dienste festgelegt und dokumentiert.

Aufbauend auf der zentralen Wissensbasis und dem dazugehörigen Index wird ein Reconciliation-Service angeboten, welcher auf derselben Server-Infrastruktur betrieben wird. Dieser Dienst wird es Projektpartnern erlauben, eigene Datensätze (unter Verwendung eigener Tools oder eines vom Kompetenzwerk bereitgestellten Datenerfassungstools) mit den Datensätzen in der zentralen Wissensbasis und in externen Normdatensätzen wie GND oder HOV abzugleichen und zu verknüpfen. Externe Normdatensätze werden dabei über die standardisierte Reconciliation-API angebunden. Der bereitgestellte Dienst hingegen wird einer eigenen erweiterten API-Spezifikation folgen, welche mit dem Quellcode auf GitHub veröffentlicht wird.

Datenverfügbarkeit und Dokumentation

(Auffindbarkeit, Zugriffsbeschränkungen)
TODO gemeinsam:
- gegebenenfalls gemeinsame Strategie zur Bereitstellung der vollständigen Datensätze
- Lizenzen festlegen, Einheitlichkeit diskutieren, z. B. hier orientieren: https://creativecommons.org/licenses/?lang=de

]

Die von den Partnern für die Datenintegration bereitgestellten Wissensbasen werden in Git-basierten Repositorien öffentlich und unter einer freien Lizenz frühzeitig (nach Erstellung der nötigen Infrastruktur) zugänglich gemacht. Außerdem werden die Datenveröffentlichungen auf den gemeinsamen und individuellen projektbezogenen Webauftritten der Partner verlinkt. Darüber hinaus ist die Veröffentlichung aller Wissensbasen spätestens zu Projektende in einschlägigen Forschungsdatenrepositorien vorgesehen. Geplant ist eine Veröffentlichung über das fachspezifische Repositorium RADAR4Culture, ein Angebot von NFDI4Culture und eventuell zusätzlich über Zenodo, ein universelles Forschungsdatenrepositorium.
Die zusätzliche Eintragung in eventuelle Such- oder Aggregationsangebote von Institutionen wie den NFDI-Konsortien (z. B. im Rahmen von NFDI4Memory) oder anderer Forschungsdateninfrastrukturen wie OstData wird geprüft.
Da nur frei zugängliche Daten durch die Partner bereitgestellt werden, können alle Daten unter einer freien Lizenz veröffentlicht werden. Eine Festlegung auf eine spezifische Lizenz folgt.

Die Partner werden angehalten ergänzend zu den Datenexporten auch weitere im Teilprojekt erzeugte Daten über die Repositorien zu veröffentlichen. Details werden im spezifischen DMP festgelegt.

Die Bereitstellung von im Projekt entwickeltem Quellcode erfolgt offen und unter freier Lizenz über den GitHub-Account des KompetenzwerkD. Durch die ergänzende Bereitstellung der nötigen Dokumentation ist somit die Nachnutzbarkeit der eingesetzten Softwarelösungen gegeben.

Die technische Dokumentation des zentralen Datenmodells, der Vokabulare, des veröffentlichten Quellcodes und der Wissensbasen erfolgt öffentlich im Rahmen der Veröffentlichung auf GitHub. Ergänzende wissenschaftliche Veröffentlichungen werden gegebenenfalls beigefügt oder verlinkt. Zusätzliche formale Metadaten werden entsprechend der Vorgaben genutzter Repositorien oder Meta-Suchangebote angegeben.

Sämtliche im Rahmen der Kernontologie verwendeten Vokabulare werden als Teil der Dokumentation der Ontologie mit betrachtet und alle im Projekt entwickelten Vokabulare werden zusammen mit der Ontologie veröffentlicht.

Die im Projekt entwickelte Reconciliation-Schnittstelle wird vom KompetenzwerkD im Rahmen ihrer dauerfinanzierten Tätigkeit auch über das Projektende hinaus betrieben und offen zur Verfügung gestellt.

Datenmanagementpläne (DMPs)

Neben diesem Dokument, das den Datenmanagementplan zu den Aspekten des FDM darstellt, die gemeinsam im Verbundprojekt geplant und umgesetzt werden und die alle oder die meisten Partner gleichermaßen betreffen, werden ergänzend von jedem Partner teilprojektspezifische Pläne erstellt. Das Datenmanagement aller dieser Pläne wird im Folgenden behandelt.

Die Ablage der Datenmanagementpläne erfolgt in einer von der SAW gehosteten Nextcloud-Instanz. Die Verwaltung wird vom KompetenzwerkD übernommen. Jedes Teilprojekt hat dabei schreibenden Zugriff auf den eigenen Plan und lesenden Zugriff auf alle weiteren.
Die Erstellung der Pläne der Teilprojekte erfolgt in Verantwortlichkeit der Teilprojekte, Details sind im jeweiligen Dokument geregelt. Das KompetenzwerkD begleitet den Bearbeitungsprozess und ist bestrebt, durch Bereitstellung von Feedback die Korrektheit und Aktualität der Pläne sicherzustellen.

Es erfolgen mindestens einmal jährlich Treffen der Teilprojekte und des KompetenzwerkD, um die Fortschritte beim Datenmanagement zu besprechen und offene Punkte zu identifizieren und voranzubringen bzw. zu klären.

Ab Q2/2023 ist es geplant, die DMPs öffentlich bereitzustellen. Die Pläne sollen auf den öffentlichen Webseiten der Teilprojekte und des Gesamtprojekts verlinkt bzw. zum Download bereitgestellt werden.

Nach Projektende sind die Datenmanagementpläne als Teil der Dokumentation zu betrachten und wie in diesem Dokument beschrieben entsprechend mit zu veröffentlichen und zu archivieren.

Verantwortlichkeiten

Gesamt-DMP: KompetenzwerkD, Verantwortliche Person: Dirk Goldhahn, Kontakt siehe oben
DMPs der Teilprojekte: siehe ebenda

Kostenfragen

Das Projekt „DIKUSA“ ist zu 100% gefördert durch Landesmittel des Freistaats Sachsen, siehe den o. g. Zuwendungsbescheid, und wird unterstützt durch die Mitarbeitenden des KompetenzwerkD. Darüber hinaus gehende Kosten fallen nach aktuellem Wissensstand nicht an.