

Kurzbericht KI-Allianz Datenplattform

Teilprojekt Datenraum Medizin

Das Projekt „KI-Datenplattform“ ist Teil der KI-Allianz Baden-Württemberg. Das Teilvorhaben wurde vom 20.07.2023 bis 30.06.2026 durch das Ministerium für Wirtschaft, Handwerk und Tourismus Baden-Württemberg finanziell gefördert.

1. Projektziele

Das Projekt Datenplattform der KI-Allianz Baden-Württemberg hatte zum Ziel, Daten und KI-Modelle unterschiedlicher Akteure über standardisierte Metadaten auffindbar zu machen und damit die Nutzung in Forschung und Wirtschaft zu erleichtern. Die Plattform sollte dabei domänenübergreifend einsetzbar sein und unterschiedliche Datenräume technisch miteinander verbinden.

Das UKF verantwortete im Gesamtprojekt den domänenspezifischen Datenraum Medizin. Ziel war es, medizinische Daten trotz hoher Anforderungen an Datenschutz, IT-Sicherheit und Governance über standardisierte Metadaten besser auffindbar und für Forschung und KI-Entwicklung nutzbar zu machen. Dazu wurden Ansätze für synthetische und anonymisierte Daten, die geschützte Ausführung externer KI-Algorithmen im Hoheitsbereich des UKF sowie ethische, rechtliche und vertragliche Rahmenbedingungen für Kooperationen untersucht.

2. Beteiligte Partner

Das Teilvorhaben „KI-Datenplattform“ wurde von neun Einrichtungen umgesetzt: Fraunhofer IOSB als Projektleitung, KIT, FZI Forschungszentrum Informatik, Universität Stuttgart/HLRS, Fraunhofer IPA, Universitätsklinikum Freiburg, Stadt Freiburg, Hochschule Aalen und Universität Tübingen.

Das UKF verantwortete den Datenraum Medizin und war zudem an Core-Modulen zu technischer Architektur, Datensuche, Ökosystem und Skalierbarkeit sowie am Service-Modul „KI-Ethik-Assessment“ beteiligt. Innerhalb des UKF wirkten insbesondere das Institut für Digitalisierung in der Medizin, das Datenintegrationszentrum und das Zentrum für Klinische Studien mit. Nutzungsszenarien wurden unter anderem mit der Averbis GmbH und dem HLRS bearbeitet.

3. Vorhabenverlauf

Nach Projektstart im Juli 2023 standen zunächst der Aufbau der lokalen technischen Infrastruktur, die Personalgewinnung und die Abstimmung mit den Projektpartnern im

Vordergrund. Nach anfänglichen Verzögerungen konnten die operativen Arbeiten 2024 deutlich ausgeweitet werden. Am UKF wurde ein lokaler piveau-Hub aufgebaut und mit dem zentralen Datenkatalog verbunden. Parallel wurden Verfahren zur Integration medizinischer Metadaten, zur Generierung synthetischer Daten und zur geschützten lokalen Ausführung externer Algorithmen erprobt.

2025 und 2026 lag der Schwerpunkt auf Konsolidierung und praktischer Erprobung. Gemeinsam mit dem Zentrum für Klinische Studien wurde ein NFDI4Health Local Data Hub aufgebaut und dessen Metadaten an HealthDCAT-AP angebunden. Zusätzlich wurde eine Confidential-Kubernetes-Infrastruktur innerhalb des Hoheitsbereichs des UKF umgesetzt und mit einem Industriepartner erprobt. Die Projektlaufzeit wurde kostenneutral bis zum 30.06.2026 verlängert

4. Erzielte Ergebnisse

Im Gesamtprojekt wurde eine domänenübergreifende Open-Source-Datenplattform auf Basis standardisierter Metadaten aufgebaut. Sie umfasst Werkzeuge zur Beschreibung und Integration heterogener Datenquellen, einen zentralen Metadatenkatalog mit Suchfunktion sowie Erweiterungen für unterschiedliche Anwendungsdomänen. Die technischen Komponenten wurden veröffentlicht und im Testbetrieb erprobt.

Das UKF übertrug diese Ansätze auf die medizinische Domäne und entwickelte hierfür spezifische Lösungen für die Beschreibung und Auffindbarkeit medizinischer Daten, die Integration bestehender Forschungsdateninfrastrukturen sowie die geschützte Nutzung sensibler Daten für KI-Anwendungen.

Medizinische Metadaten und zentrale Suche: Medizinische Datenbestände wurden auf Metadatenebene nach DCAT-AP beziehungsweise HealthDCAT-AP beschrieben und in einem lokalen piveau-Hub bereitgestellt. Dieser wurde mit dem zentralen Hub synchronisiert; die medizinischen Metadaten wurden an die zentrale Such- und Abfrageschnittstelle angebunden und im Testbetrieb erfolgreich erprobt.

NFDI4Health und HealthDCAT-AP: Gemeinsam mit dem Zentrum für Klinische Studien wurde ein NFDI4Health Local Data Hub aufgebaut. Metadaten zu Forschungsprojekten und -datensätzen wurden automatisiert in HealthDCAT-AP transformiert und damit an die KI-Datenplattform anschlussfähig gemacht.

Das hierfür entwickelte Transformationswerkzeug wurde als Open-Source-Software veröffentlicht und dokumentiert. Es schafft eine technische Brücke zwischen einer medizinischen Forschungsdateninfrastruktur und domänenübergreifenden Datenkatalogen und stellt zugleich eine Vorarbeit für mögliche spätere Anbindungen an nationale und europäische Katalogstrukturen dar.

Synthetische Daten und Anonymisierung: Für synthetische medizinische Daten wurden unterschiedliche Ansätze erprobt. Mit Synthea wurden Datensätze erzeugt, MIMIC wurde für Entwicklung und Erprobung genutzt und für komplexere UKF-Routinedaten wurde HALO eingesetzt. Mit zunehmender medizinischer Detailtiefe stiegen Speicher- und Validierungsaufwand jedoch erheblich. Ein ursprünglich vorgesehener Server mit hinreichend realistischen synthetischen UKF-Daten konnte deshalb nicht fertiggestellt werden.

Ergänzend wurde ein Konzept zur risikoadaptierten Generalisierung von ICD-Diagnosen entwickelt, mit dem der Informationsgehalt medizinischer Datensätze möglichst weit erhalten und zugleich die für den Nutzungskontext erforderliche Anonymisierung beziehungsweise Generalisierung unterstützt werden soll.

„Bring the algorithms to the data“: Für den Ansatz „Bring the algorithms to the data“ wurde die am UKF aufgebaute Confidential-Kubernetes-Infrastruktur auf Basis einer Distribution von Edgeless Systems evaluiert sowie die on premise verfügbare Kubeflow-Infrastruktur. Die Infrastruktur unterstützte GPU-fähige Container und ermöglichte die geschützte Ausführung externer Algorithmen innerhalb des Hoheitsbereichs des UKF.

Der Ansatz wurde unter anderem anhand der Software Health Discovery der Averbis GmbH zur Erstellung de-identifizierten Daten erprobt. Der Algorithmus konnte de-identifizierte Daten innerhalb des Verantwortungsbereichs des UKF erzeugen, die dadurch für Modelltraining nutzbar werden.

Im Rahmen der vorgesehenen Machbarkeitsstudie wurde geprüft, ob der Confidential-Kubernetes-Ansatz auf Hochleistungsrechenressourcen des HLRS übertragen werden kann. Die praktische Übertragung war aufgrund der Lizenzbedingungen beziehungsweise der fehlenden Verfügbarkeit der erforderlichen NVIDIA-Lizenzen für den vorgesehenen GPU-basierten Betrieb nicht möglich. Eine technische Umsetzung am HLRS erfolgte daher nicht.

Zusätzlich stellte Edgeless Systems die verwendete Confidential-Kubernetes-Distribution im Projektverlauf ein. Die am UKF aufgebaute Lösung war damit in der realisierten Form nicht mehr als dauerhaft gepflegte Produktbasis verfügbar. Für einen künftigen Regelbetrieb wäre eine Migration auf eine weiterhin unterstützte technische Lösung sowie eine erneute Sicherheits- und Betriebsbewertung erforderlich.

Ethik, Recht und Governance: Die technische Entwicklung wurde durch eine Taskforce Ethik & Recht begleitet. Ergebnisse eines interdisziplinären Workshops sowie Impulse aus dem Austausch mit dem Landesbeauftragten für den Datenschutz und die Informationsfreiheit Baden-Württemberg flossen in ein im Februar 2026 veröffentlichtes Whitepaper ein. Dieses enthält unter anderem Empfehlungen zu Rollen- und Rechtenkonzepten, risikobasierten Prüfungen, Datenqualität, Audits, Monitoring, Dokumentation sowie möglichen Governance- und Compliance-Strukturen.

5. Mögliche Anwendungsfelder und Nutzen für die Wirtschaft

Die entwickelten Lösungen adressieren zentrale Einstiegshürden für Unternehmen, die medizinische Daten für Forschung und KI-Entwicklung nutzen möchten.

Standardisierte Metadaten und eine zentrale Suchfunktion verbessern die Auffindbarkeit vorhandener Datenbestände, Forschungsaktivitäten und Kooperationsmöglichkeiten. Für besonders schützenswerte Daten bietet der Ansatz „Bring the algorithms to the data“ einen alternativen Kooperationsweg: Externe Algorithmen können unter kontrollierten Bedingungen innerhalb der Krankenhausinfrastruktur ausgeführt werden, ohne medizinische Daten an Unternehmen zu übertragen.

Open-Source-Werkzeuge wie die Transformation von NFDI4Health-Metadaten nach HealthDCAT-AP erleichtern zudem die Nachnutzung der Projektergebnisse. Die begleitenden Arbeiten zu Datenschutz, Governance und Vertragsgestaltung schaffen Grundlagen für zukünftige Kooperationen zwischen Krankenhäusern, Forschungseinrichtungen und Unternehmen.

Zum Ende der Projektlaufzeit bestand noch kein standardisierter medizinischer Regelbetrieb mit wiederkehrenden Datenzugängen für externe Unternehmen. Die Pivau-Plattform verbindet aber auf Ebene der Metadaten die Domäne Medizin mit anderen Datenräumen und ist grundsätzlich kompatibel mit Konzepten des kommenden EHDS. Die Projektergebnisse sind daher überwiegend als technisch erprobte Pilotkomponenten, interoperable Metadatenlösungen und konzeptionelle Grundlagen für die weitere Umsetzung einzuordnen.