TP1 – Datenanalyse und -austausch
- ZLG Plattformkonzept -
Version 2.3

ZUKUNFTSLABOR Gesundheit

ZDIN
Zentrum für digitale Innovationen Niedersachsen

Eingereicht von:
Teilprojekt 1 – Zukunftslabor Gesundheit

Zukunftslabor Gesundheit
Stand: 13.09.2022

Version	Datum	Bearbeiter	Kommentar	Prüfer
0.1	08/20 – 01/21	J. Richter	Initialer Aufschlag	/
0.4	02/21	J. Richter	Re-Strukturierung, Aufschlag UC Datenanalyse und Roadmap	/
0.5	02/21	J. Richter	Formulierung Relevante Technologien und Update Technischer Ansatz sowie Beispieldaten, Formulierung Plattform-Komponenten + neue Grafik + Architektur-Skizze mit Erläuterung	/
0.6	02/21	J. Richter	Formulierung Plattform-Komponenten und Architektur-Skizze, Darstellung von Diskussionspunkten zur Roadmap	/
	08.02.2021	J. Richter	Vorstellung aktueller Stand + Diskussion	TP1
0.7	02/21	J. Richter	Formulierung Stand der Technik, Dokument als Basis für Erstimplementierungen	/
2.0	04/21 – 05/21	J. Richter	Straffen des Dokuments, Fokus auf verteilte Infrastruktur mit DI-Pipeline und Datensätzen bei den Partnern bzw. im Verantwortungsbereich der Partner	/
2.1	10.05.2021	TP1	Vorstellung aktueller Stand + Diskussion	TP1
2.2	07/21	J. Richter	Überarbeitung Datenanalyse-Part, Darstellung des Gesamtvorhabens und der Komponenten	/
2.3	05/22 – 09/22	J. Richter, M.Katzensteiner	Finalisierung	

Inhaltsverzeichnis

Abkürzungsverzeichnis

Akronym	Langform
AP	Arbeitspaket
CLI	Command Line Interface
EHR	Electronic Health Record
FHIR	Fast Healthcare Interoperability Resources
MII	Medizininformatik Initiative
TA	Task
TP	Teilprojekt
ZDIN	Zentrum für digitale Innovation Niedersachsen
ZL-G	Zukunftslabor-Gesundheit
MeDIC	Medizinisches Datenintegrationszentrum
TMF	Technologie- und Methodenplattform für die vernetzte medizinische Forschung
TTP	Trusted Third Party
UMG	Universitätsmedizin Göttingen
MHH	Medizinische Hochschule Hannover
DSGVO	Datenschutzgrundverordnung
BDSG	Bundesdatenschutzgesetz
LDSG	Landesdatenschutzgesetz
HsH	Hochschule Hannover
PSN	Pseudonym
IDAT	Identifizierende Daten
MDAT	Medizinische Daten

Abbildungsverzeichnis

Tabellenverzeichnis

1 Zusammenfassung

Die am Teilprojekt 1 beteiligten Wissenschaftler des Zukunftslabor Gesundheit erproben verschiedene Werkzeuge mit Blick auf Austausch, Interoperabilität, Nachnutzbarkeit und Analyse von Daten. Die daraus entstehende Plattform ist eine Sammlung von Werkzeugen bzw. Ein Werkzeugkasten für die Verwendung, Nutzung und Steuerung von interoperabler Datenhaltung, Datentransformation und Datenmanagement unter Verwendung von openEHR als Modellierungsansatz. Die notwendigen Schritte zur Inbetriebnahme und Nutzung der Werkzeuge werden als Dokumentation zu den Werkzeugen aufbereitet und das Gesamtpaket wird als „ZLG-Plattform“ der Öffentlichkeit zugänglich gemacht werden.

Im Kern soll Anwendern ermöglicht werden, Daten in ein ausgewähltes Austauschformat zu transformieren, die Daten in sicherer Umgebung in eigener Infrastruktur vorzuhalten sowie die Daten für berechtigte Forscher durchsuchbar und letztendlich nutzbar zu machen.

Der standortübergreifende Austausch von medizinischen Versorgungs- und Forschungsdaten wird dabei durch die Etablierung des openEHR-Frameworks unterstützt. Adressiert wird der Import von Daten bzw. deren Transformation vom Quellformat hin zum standardisierten openEHR-Format. Technische Aspekte werden in einer Anleitung zusammengefasst, organisatorische Aspekte werden als Governance-Framework beschrieben.

2 Einleitung

2.1 Motivation

Das vorliegende Dokument beinhaltet das Konzept einer standardbasierten und damit interoperablen ZLG-Datenplattform und nimmt Bezug auf Ziele, Nutzungsprozesse, Anforderungen, Architektur sowie das Datenmanagements unter Beachtung der FAIR-Prinzipien[1] sowie die Plattform-Governance. Das Plattformkonzept wird die im Projektstrukturplan aufgezeigte Vision einer standardbasierten Forschungsplattform und dazugehöriger Komponenten konkretisiert. Die exemplarische Umsetzung einzelner Nutzungsprozesse wird in Form einer Roadmap strukturiert.

[1] https://www.forschungsdaten.org/index.php/FAIR_data_principles (abgerufen: 20.05.2021)

Mit der Konzeption einer standardbasierten Forschungsplattform wird auf eines der Kernziele des Zukunftslabor Gesundheit bzw. des Zentrums für digitale Innovationen Niedersachsen (ZDIN) hingearbeitet. Für die Zukunftslabore des organisatorisch übergeordneten ZDIN steht die Vernetzung von Akteuren, Ideen sowie Kompetenzen aus Forschung und Praxis im Mittelpunkt. Der Standort Niedersachsen soll durch Vernetzung und Digitalisierung von niedersächsischen Spitzenreitern aus Wissenschaft und Wirtschaft gestärkt werden. Die in den Zukunftslaboren beteiligten Partner initiieren interdisziplinäre und standortübergreifende zukunftsweisende Forschungsprojekte und setzen diese gemeinsam um.

Im Rahmen des ZL-G finden Arbeiten statt, welche sowohl die strategischen Ziele des ZL-G und des ZDIN unterstützen, als auch Projekte mit vergleichbarer Zielsetzung. Ähnliche Projekte, wie z.B. das HiGHmed-Konsortium, zeigten bereits die Komplexität und den finanziellen Aufwand für sowohl Personal, als auch Hard- und Software bei der Etablierung von Forschungsdateninfrastrukturen auf. Im HiGHmed-Projekt werden sogenannte „Medizinische Datenintegrationszentren“ (MeDICs) an Universitätsklinika aufgebaut. Ebenso setzt das COFONI-Projekt, welches eine Plattform für die Verwaltung und Nutzung von COVID-Forschungsdaten aufbaut, auf im ZLG ausgewählte Standards. Durch Kooperationen sowie Nutzung festgelegter Standards fügen sich die Arbeiten des ZL-G in die aktuellen Digitalisierungsbestrebungen sowohl inhaltlich als auch organisatorisch sinnvoll ein.

Die strategischen Ziele des ZDIN bzw. Zukunftslabors werden auf operativer Ebene in Forschungsvorhaben in einrichtungsübergreifender Kollaboration durch die ZL-G Partner umgesetzt. Im ZL-G finden diese Arbeiten in den drei Teilprojekten statt, welche mit den Themen Datenaustausch und –Analyse (Teilprojekt 1), Sensorik (Teilprojekt 2) und Aus-, Fort- und Weiterbildung (Teilprojekt 3) befasst sind. Die Konzeption und Bereitstellung einer interoperablen Forschungsplattform, welche im Fokus dieses Dokuments steht, ist als Arbeitspaket in Teilprojekt 1 angelegt. Korrespondierende Arbeitspakete mit Inhalten zu Datenanalyse und Datenverwaltung bzw. Governance sind ebenfalls im Teilprojekt 1 verortet. Zentrale Grundlage der standardbasierten Datenmodellierung und Bereitstellung ist der internationale openEHR-Standard für Systeme zur Bereitstellung elektronischer Gesundheitsakten.

Auf Arbeitsebene wird in der Projektanfangsphase (Jahr 1 und 2) auf die Erlangung wissenschaftlicher Erkenntnisse in den Bereichen interoperabler Datenaustausch und Analyse

von standardbasiert modellierten Daten abgezielt. In einem auf Literaturrecherche aufbauenden Evaluationsbericht zu bestehenden Plattformen wurden wichtige Aspekte identifiziert und für die Konzeption berücksichtigt. Eine Forschungsplattform setzt sich aus verschiedenen Einzel-Komponenten zusammen, welche verschiedene Prozesse für Implementation, Wartung, Betrieb und Nutzung der Plattform beinhalten bzw. voraussetzen. Die Erforschung derartiger Plattformen kann daher über den Aufbau, die Erprobung und die Dokumentation dieser Einzel-Prozesse realisiert werden. Konkret wird dies anhand eines exemplarischen ausgewählten Datensatzes erprobt. Berücksichtigt werden dabei insbesondere Governance-Prozesse, die Erprobung der technischen Umsetzung und die anschließende wissenschaftliche Aufbereitung der Ergebnisse. Die Translation der gemachten Erfahrungen hin zu bestehenden und zukünftigen openEHR-basierten Datenmanagementplattformen wird in Form von wissenschaftlichen Veröffentlichungen, in Form von Lehrinhalten oder Demonstratoren für die ZL-G Lehre und anderen öffentlichkeitswirksamen Darstellungen der erfolgten Arbeiten geschehen.

In den Folgejahren (Jahr 3 - 5) steht die Validierung bzw. Evaluation der Erst-Implementation der ZL-G Infrastruktur mit anschließender Weiterentwicklung an.

2.2 Stand Forschungsdatenmanagement

Der Stand des Forschungsdatenmanagements ist im Dokument Evaluationsbericht[2], welches im Rahmen der Arbeiten im AP1.4 „Umsetzung der Datenhaltung und Aufbau der Plattform“ erarbeitet wurde, beschrieben. Wichtige Ergebnisse aus dem Evaluationsbericht werden im Folgenden herausgestellt.

Der Evaluationsbericht zu Forschungsdatenplattformen hat das Vorhandensein verschiedener Kategorien von Forschungsplattformen aufgezeigt. Besonders signifikant sind hierbei die sich ergebenden unterschiedlichen Anforderungen bei der Haltung kurierter Daten bzw. Forschungsergebnissen gegenüber der Haltung von Daten zu Analysezwecken (wobei Letzteres im Zukunftslabor Gesundheit im Fokus steht). Kurierte Daten / Forschungsergebnisse weisen bereits einen hohen Grad an Strukturiertheit auf und unterscheiden sich damit gegenüber Rohdaten für Analysezwecke im Hinblick auf die Komplexität bzgl. Datenmodellen, in der Art

[2] Siehe TA1.4.1_1.1 Evaluationsbericht Forschungsplattformen.pdf

der Umsetzung der FAIR-Prinzipien, der Art des Zugriffs und der Weiterverwendung der Daten oder beim Grad an Strukturiertheit bzgl. der Metadaten / Standards.

Für den inter-institutionellen Austausch auf Datenebene muss für die Sicherstellung von Interoperabilität auf geeignete Datenmodelle und Technologien (wie beispielsweise openEHR[3] und FHIR[4]) zurückgegriffen werden. Mögliche Umsetzungsvarianten openEHR-basierter Plattformarchitekturen werden im Folgenden skizziert. Die Festlegung auf die Verwendung des openEHR-Technologieframeworks im Zukunftslabor Gesundheit ist im Projektstrukturplan festgelegt und wurde bereits im Evaluationsbericht adressiert.

Die Fokussierung auf die openEHR-Technologie, die bereits im ZL-G Plattform-Evaluationsbericht begründet wurde, wird im Plattformkonzept fortgesetzt. Eine Erläuterung der Grundlagen und Rahmenbedingungen zum openEHR-Standard und openEHR-Werkzeugen wurde im Rahmen des Evaluationsberichts erarbeitet. Erläuterungen bezüglich openEHR finden sich ebenfalls in Kapitel 3.2.2 „openEHR-Server“, in dem der openEHR-Server EHRBase näher beschrieben wird und in Kapitel 3.2.3 zur “DI-Pipeline“, in welchem auf die Erstellung von openEHR-Ressourcen, dafür notwendige ETL-Schritte sowie die zu verwendende REST-API eingegangen wird.

Existierende Forschungsdatenbanken

Es existiert bereits eine Vielzahl verschiedenster Forschungsdatenbanken und Forschungsdatenplattformen.

Unterscheiden lässt sich hier zwischen Lösungen, welche sich als „Publikations-Verzeichnisse“ und denen die sich als „Forschungsdaten- / Datenintegrationsplattformen“ bezeichnen lassen. Erstere dienen der Verwaltung von für die Forschung bereits verwendeten Datensätze im Sinne der FAIR-Kriterien z.B. aus Gründen der Reproduzierbarkeit. Letztgenannte dagegen fokussieren auf die Verwaltung von Daten aus der klinischen Versorgung („Rohdaten“) oder anderen Quellen, um sie qualitätsgesichert und kodiert zu integrieren und danach für zugriffsberechtigte Wissenschaftler für zukünftige Forschungsarbeiten zur Verfügung stellen und sie zu diesem Zweck durchsuchbar, vor-auswertbar und analysierbar vorhalten. Eine Forschungsdatenplattform für die Nutzung und Verwaltung von „Rohdaten“ muss damit andere Anforderungen erfüllen, als eine Plattform für „Publikationsdaten“.

[3] https://www.openehr.org/about/what_is_openehr (abgerufen: 02.05.2021)

[4] https://www.hl7.org/fhir/overview.html (abgerufen: 02.05.2021)

Moderne Forschungsplattformen sind meist modular aufgebaut und bieten dadurch eine erhöhte Flexibilität und Erweiterbarkeit, auch gegenüber neuen Anforderungen. Als Beispiel kann hier insbesondere die GeRDI-Plattform[5] oder die Medizinischen Datenintegrationszentren im Rahmen des HiGHmed-Projekts[6] benannt werden. Ersteres verdeutlicht den modularen Aufbau aus Self-Contained-Systems bzw. Mikroservices, letzteres verdeutlicht besonders die prozessinterne Verknüpfung verschiedener Datenformate und Austauschformate.

Herausforderungen und Erkenntnisse

Anhand der in der Literaturrecherche erfassten Plattformen konnten verschiedene Herausforderungen identifiziert werden. Aufgrund des begrenzten Umfangs, der Komplexität sowie Diversität der Lösungen zum Thema Forschungsdatenmanagement ist die folgende Auflistung nicht erschöpfend, beinhaltet allerdings wichtige Aspekte.

Identifizierte Aspekte:

- Iterative Implementierung und Erweiterung (Wie wird die Plattform (weiter-) entwickelt)
- Skalierbarkeit (Datenmenge, bei neuen Datensätzen, bei neuen Partnern)
- Workflows / Anwendungsszenarien (Was unterstützt die Plattform)
- Auswahl der Softwarewerkzeuge (Funktionen und Erweiterbarkeit der Lösungen)
- Datennutzungsbedingungen (Patienteneinwilligung, Data Governance)
- Pseudonymisierung und Anonymisierung (Infrastruktur)
- Konfigurations- und Entwicklungsaufwand

Verfügbare Tools

Kritisch für die Umsetzung der Plattform ist die Wahl der technischen Lösungen bzw. Lösungsansätze bzgl. ETL und Datenhaltung. Die Überführung der Daten von den Quellsystemen in openEHR-basierte Ressourcen ist eine zentrale Anforderung zur Nutzung der Daten. Die Tools und Sprachen für die Auswertung/Analyse der modellierten Daten sind ebenso kritisch für die Plattformnutzung. Die auszuwählenden Libraries und Tools bestimmten maßgeblich die Komplexität in der Umsetzung, die Performance, Skalierbarkeit, Wartbarkeit und Erweiterbarkeit der Plattform.

[5] https://www.gerdi-project.eu (abgerufen: 19.02.2021)

[6] https://www.medizininformatik-initiative.de/de/konsortien/highmed (abgerufen: 19.02.2021)

Vielversprechend ist hier die „openEHR_SDK“[7] Java-Library der EHRBase-Entwickler. Die Kombination dieser Library mit flexiblen und erweiterbaren Lösungen für Frontend und Backend ist maßgeblich für die Umsetzung. Dies muss in der Detailplanung des Systems (z.B. Web-Applikation, automatisierte ETL-Strecke, Anbindung der Datenrepositorien) berücksichtigt werden. Durch die im openEHR-Standard bzw. der zugehörigen Spezifikation definierte REST-Schnittstelle ist ein Zugriff auf Daten jederzeit auf aus verschiedensten Tools und Sprachen möglich, solange mit diesen REST-Anfragen möglich sind.

Der openEHR-Server ist aufgrund der in der openEHR-Spezifikation definierten REST-Schnittstelle eine einfach austauschbare Komponente. Mit der EHRBase[8] existiert ein geeigneter Server, der Open-Source verfügbar ist, die openEHR-Spezifikation bereits zu großen Teilen umsetzt, aktiv weiterentwickelt und von verschiedenen Akteuren genutzt und unterstützt wird. An der Entwicklung der EHRBase wirken unter anderem zentrale Mitglieder der openEHR-Community mit, die aktiv an der Weiterentwicklung der Spezifikation beteiligt sind. Neue Funktionalitäten werden teilweise initial in der EHRBase implementiert, bevor sie in die Spezifikation übernommen werden.

Die Modellierung sowie die Modellierungstools können getrennt von anderen Komponenten betrachtet werden. Hier ist aufgrund der inhaltlichen und organisatorischen Trennung der Rollen und Zuständigkeiten eine Integration mit anderen Softwarelösungen für die Datenhaltung nicht zwingend. Die Modellierung wird im Governance Framework des ZL-G detailliert erläutert. Von Modellierern erstellte Artefakte (Archetypen, Templates, Value-Sets) basieren auf einheitlichen Standards / Spezifikation und lassen sich einfach übertragen und in der Plattform für z.B. ETL, Formularerstellung, etc. nutzen.

2.3 Problemstellung und Zielsetzung

Datensammlungen in isolierten „Daten-Silos“, proprietäre Formate und fehlende Metadaten sind ein großes Hindernis für die Nachnutzung und Wiederverwendung von Forschungsdaten [1]. Die Interoperabilität von Systemen, welche (medizinische) Daten verarbeiten bzw. bereitstellen, ist die Grundlage für Austausch und Nutzung der Daten. Durch Verwendung von Standards bei Datenmodellen, Austauschformaten sowie Kodierung bei der Erfassung, Speicherung und Auswertung von (medizinischen) (Forschungs-) Daten, können diese

[7] https://github.com/ehrbase/openEHR_SDK (abgerufen: 08.05.2021)
[8] https://github.com/ehrbase/ehrbase (abgerufen: 08.05.2021)

zukunftsweisende Technologien und Methoden wie bspw. Künstliche Intelligenz, Big Data oder mobile Anwendungen ermöglichen. [2,3]

Ein Standard für die Modellierung von medizinischen Daten ist openEHR. Das openEHR-Technologieframework ist eine Lösung für die „community-unterstützte“ Umsetzung eines einheitlichen Datenmodells mit standardisiertem Zugriff (z.B. via REST-API). Der openEHR-Ansatz verbindet dabei die Beiträge und das Wissen von Gesundheitsdienstleistern, Forschern, IT-Experten und Datenwissenschaftlern. Vorteile und Erläuterungen des openEHR-Ansatzes, bei welchem die Modellierung getrennt von der Implementierung ist und durch Domänenexperten realisiert wird finden sich zum einen in der openEHR-Spezifikation[9] selbst, als auch in zugehörigen Veröffentlichungen zum Standard und inzwischen immer öfter in Erfahrungs- und Erfolgsberichten zu Forschungs- und Praxisanwendungen mit openEHR.

Durch die interoperable Modellierung von Daten für Datenverwaltung und Datennutzung kann die damit einhergehende Standardisierung die Daten im Hinblick auf die Qualität steigern und damit besser nachnutzbar machen. Innerhalb des Zukunftslabors wurde der openEHR-Standard und dessen Nutzung, angelehnt an das Projekt HiGHmed, als Forschungsobjekt ausgewählt:

> Die Partner im Zukunftslabor Gesundheit sollen zum einrichtungs- und institutionsübergreifenden Datenaustausch (standardbasiert mittels openEHR) befähigt werden.

Eine Eingrenzung des Betrachtungsgegenstandes bei der Erforschung ist im Hinblick auf die Umsetzbarkeit zwingend notwendig. Aus diesem Grund wird die Erforschung der Plattform heruntergebrochen auf einzelne Aspekte, welche für zum Setting des Zukunftslabors passende Anwendungsfälle relevant sind.

2.4 Leit- und Forschungsfragen

Die Konkretisierung des Konzepts im Arbeitspaket AP1.4 „Umsetzung der Datenhaltung und Aufbau der Plattform“ und des Plattform-Use Case „Datenanalyse“ wird durch die folgenden Fragestellungen geleitet.

[9] https://specifications.openehr.org (abgerufen: 10.05.2021)

1. Welche Anforderungen haben die am ZL-G beteiligten Partner an eine interoperable Datenplattform für den Austausch und die Analyse von Daten mit medizinischem Kontext?
2. Wie können die verfügbaren Werkzeuge verwendet werden bzw. welche Anforderungen können mit bestehenden Werkzeugen nicht erfüllt werden?
3. Wie können die im ZL-G erprobten Werkzeuge und Prozesse weiteren Nutzern definiert und dokumentiert und damit niedrigschwellig nutzbar gemacht werden?

Die Leit- und Forschungsfragen werden teils in den verschiedenen Arbeitspaketen des Teilprojekts 1 des ZL-G adressiert. So steht zum Beispiel die Art und Weise der Modellierung von Gesundheitsdaten im Fokus von AP1.2 / AP1.3 oder die exemplarische Umsetzung in AP1.1. Hinsichtlich dieses Konzeptpapiers steht die technische Implementation der Werkzeuge im Fokus.

Inhaltlich wird bei den geplanten Arbeiten die Beantwortung der im Plattform-Use Case „Datenanalyse im Kontext kardiologischer Erkrankungen“ festgelegten Fragestellungen verfolgt. Die im Folgenden vorgenommene weitere Konkretisierung der Fragestellung bezüglich der Prozesse ist daher sinnvoll für die Strukturierung der Arbeiten.

1. **Forschungsfrage Modellierungsprozess (AP1.2/AP1.3):**
   1.1 Wie können (innerhalb des Zukunftslabors Gesundheit) Prozesse für die standardbasierte Modellierung der Daten mit openEHR gestaltet werden, um Beispiel- bzw. reale Daten von Partnern in einem gemeinsamen Datenmodell abzubilden?
2. **Forschungsfrage Datenaufbereitung und -Import bzw. ETL-Prozess (AP1.4):**
   2.1 Wie können Prozessschritte für das Extrahieren, Transformieren und Laden von Beispiel- bzw. Partnerdaten gestaltet werden, um unter Verwendung von (idealerweise open-source) Werkzeugen Daten von Partnern in eine openEHR-basierte Datenhaltung zu überführen?

   2.2 Wie kann der Datenintegrationsprozess so gestaltet werden, dass Anwender, ohne umfassende Vorerfahrung bezüglich openEHR, diesen ohne größeren Aufwand durchführen können?

2.3 Welche (open-source-) Werkzeuge können für das Importieren von Daten eingesetzt werden und wie sind diese zu konfigurieren?

3. **Forschungsfrage Datenexploration und -export (AP1.3/AP1.4/AP1.5):**

3.1 Wie können Prozessschritte für den Export der Daten zu Analysezwecken oder die Anbindung von ausgewählten Analysewerkzeugen an die Datenhaltung per Schnittstelle ausgestaltet werden?

3.2 Welche Nutzungsprozesse sind möglich und wie sieht die Verwendung von openEHR für End-Anwender aus?

3.3 Welche (open-source-) Werkzeuge können für das Explorieren und Exportieren von Daten eingesetzt werden und wie sind diese zu konfigurieren?

Thematisch lässt sich das Forschungsdatenmanagement im Kontext der Plattform allgemein in die folgenden Bereiche unterteilen:

1. Datenintegration (Import von Daten und Speicherung von (Meta-) Daten),
2. Datenexploration, Datenabfrage und Datenexport
3. (Privatheitsbewahrende) Datenanalyse
4. Datenmanagement / Governance der Plattform.

2.5 Ausschluss

Für die operative Nutzung einer Forschungsdatenplattform für einrichtungsübergreifende Vernetzung und Austausch konnten im Rahmen des Evaluationsberichts[10] verschiedene Aspekte bzw. Teilthemen identifiziert, die im Rahmen des Plattformkonzepts zwar relevant sind, allerdings außerhalb des Scopes liegen. Unter anderem lassen sich hier Prozesse rund um Verwaltung und Organisation der Plattform, die Umsetzung der Governance bezüglich Datensätzen und Datennutzungsbedingungen oder Anforderungen an Patienteneinwilligungen als Basis einer Datennutzung, nennen. Diese Aspekte stehen in diesem Konzept, das die Anforderungen, Architektur und technische Ansätze beschreibt, nicht im Fokus.

[10] Siehe TA1.4.1_1.1 Evaluationsbericht Forschungsplattformen.pdf

In erster Instanz findet die Nutzung von Beispieldaten für die Umsetzung des Plattform-Use Case „Datenanalyse“ im Kontext kardiologischer Erkrankungen statt. Die modellierten und abgelegten Daten werden innerhalb des Use Cases und in der frühen Projektphase nur durch die im TP1 beteiligten Wissenschaftler genutzt.

3 Plattformkonzept

Grundlage für die Detail-Planung einzelner Plattformkomponenten ist die Erfassung und Auswertung existierender Anforderungen an eine interoperable Datenbasis und die dafür notwendige (Plattform-)Infrastruktur. Aufgrund des Anwendungsbereichs Gesundheitssystem sind besondere Anforderungen aufgrund der Sensibilität der behandelten Daten und Informationen zu berücksichtigen.

Eckpunkte einer Datenplattform im Gesundheitsbereich:

- Sensible Patientendaten (Verarbeitung besonderer Kategorien personenbezogener Daten – Art. 9 DSGVO[11])
- Nicht-anonyme Daten nur mit Einwilligung (Consent) und Ethikvotum nutzbar (NDSG[12]).
- Austausch von Daten mit Unternehmen aufgrund von Nutzungseinschränkungen oft erschwert.
- Austausch von Ergebnissen daher meist leichter als Austausch von Rohdaten.
- Fokus bei Datenintegration von Rohdaten auf kleineren Nutzerkreis (Kliniker, Forscher).
- Translation von Ergebnissen erfolgt typischerweise zu anderen Forschern oder zu Bürger*n.

Orientiert an den verfügbaren Ressourcen und dem Fokus auf die Vorgehensschritte des Use Case Datenanalyse wurden Anforderungen sowohl von beteiligten ZLG-Partnern erfasst, aber auch sich aus dem technischen Kontext ergebende Anforderungen gesammelt. Über User-Stories wurden durch die Partner relevante Nutzungsszenarien skizziert. Ergänzt wurden diese durch Implementierungstests und erste konzeptionelle Skizzen.

[11] https://dsgvo-gesetz.de/art-9-dsgvo/ (abgerufen: 19.02.2021)

[12] http://www.nds-voris.de/jportal/?quelle=jlink&query=DSG+ND+%C2%A7+17&psml=bsvorisprod.psml&max=true (abgerufen: 19.02.2021)

3.1 Kern-Anforderungen

Für eine Forschungsdatenplattform ist das Vorhandensein einer entsprechenden (Web-) Applikation, welche an den Bedürfnissen der Nutzer ausgerichtet, die Funktionalitäten der Plattform nutzbar und zugreifbar macht, zentral. Organisatorisch einfach wäre hierfür eine zentral verwaltete und zentral entwickelte Plattform wobei die Datenhaltung sowohl zentral als auch dezentral betrieben werden könnte. Der Aufwand für die Modellierung von (neuen) Datensätzen sowie Import, Weiterentwicklung läge damit zentral beim Plattform-Betreiber. Vergleichbare Plattformen (z.B. HiGHmed MeDICs) haben gezeigt, dass der Aufwand und Komplexität hierbei immens sein kann und ein großes und erfahrenes Team benötigt wird. Die dafür notwendigen Voraussetzungen bezüglich verfügbarer Ressourcen sind im Zukunftslabor nicht gegeben. Das Ziel des Zukunftslabors ist es daher nicht, eine umfassende zentralisierte Plattform als Service zu betreiben, sondern eher den openEHR-Ansatz und Daten in der Breite für eine Vielzahl von Partnern nutzbar zu machen. Hierfür sind eine effiziente Ressourcennutzung, schlanke Verwaltungsstrukturen sowie das Automatisieren von technischen Prozessen notwendig. Das Vorhandensein von openEHR-Expertise kann bei den beteiligten Partnern nicht vorausgesetzt werden.

Während in der initialen Entwicklungsphase die Umsetzung der technischen Basis im Fokus steht, soll die Weiterentwickelbarkeit und Erweiterbarkeit der Plattform durch entsprechendes Design sichergestellt werden. Die Kern-Funktionalitäten der Plattform ergeben sich aus den grundlegenden Zielen und aus unter den ZL-G Partnern diskutierten Nutzungsszenarien.

3.2 Architektur und Komponenten

Die konzeptionelle Herausforderung liegt in der Planung einer technischen Infrastruktur und der Definition organisatorischer Prozesse, welche den formulierten Kern-Anforderungen gerecht werden. Ausgehend von einer simplen Architektur (Kernkomponente Datenhaltung mit Webserver) soll diese anhand der Anforderungen so angepasst werden, dass die Anforderungen erfüllt und die Ziele des Zukunftslabors realisiert werden können.

3.2.1 Mögliche Plattformarchitektur

In Abbildung 1 ist eine mögliche Plattformarchitektur aufgezeigt. Die Datenhaltung (EHRBase, FAIR-Daten-Repository, File-Store) ist hierbei an einen Web-Server / Web-Applikation angebunden, über welche die Plattformfunktionalitäten den Nutzern bereitgestellt werden könnten. Diese zentrale Komponente müsste hierfür eigens entwickelt werden. Weitere

Komponenten für Betrieb und Nutzung (Identity-Service für Nutzerverwaltung, Zertifikatsservice, Terminologie-Server) wären hierfür ebenfalls anzubinden. Die Komponenten für openEHR-Modellierung und Clinical Knowledge Management (CKM) sind eigenständig und unabhängig von den restlichen Komponenten der Plattform des Zukunftslabors.

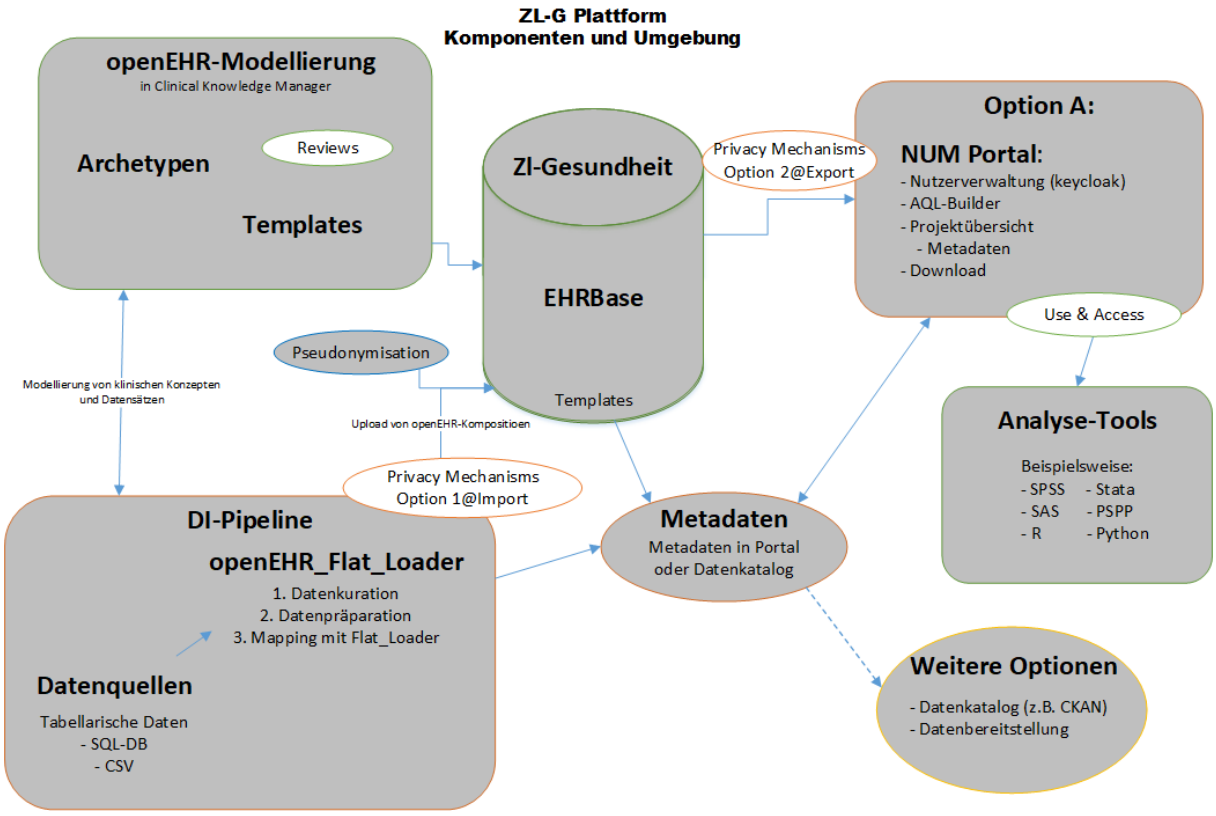

Abbildung 1: Mögliche ZL-G Plattform Komponenten

Der Übergang zu einer schlankeren Organisation und Verwaltung einer solchen Plattform soll im Zukunftslabor Gesundheit durch die Aufteilung der Zuständigkeiten über einzelnen Komponenten mit einhergehender Automatisierung von Prozessen geschehen. In Abbildung 2 ist eine Plattformarchitektur abgebildet, die diese Anforderungen erfüllen kann. Die Datenhaltungskomponente (EHRBase, Filestore) wird hier als CLI-Container zusammen mit einer halb-automatisierten ETL- / Datenintegrations-Pipeline (DI-Pipeline) umgesetzt. Damit ist sowohl eine zentrale (z.B. für gemeinsame Datensätze) als auch dezentrale Datenhaltung (z.B. bei einem ZLG-Partner) möglich oder auch ein Schicken der generierten Ressourcen an einen externen Data Store z.B. einen openEHR-Server in einem MeDIC.

Terminologie-Server, ein Pseudonymisierungsdienst (z.B. Mainzelliste[13]) und Modellierungstools können als zentrale Komponenten betrieben werden. Insbesondere hinsichtlich Record Linkage und Einheitlichkeit bei der Modellierung ist dies sinnvoll.

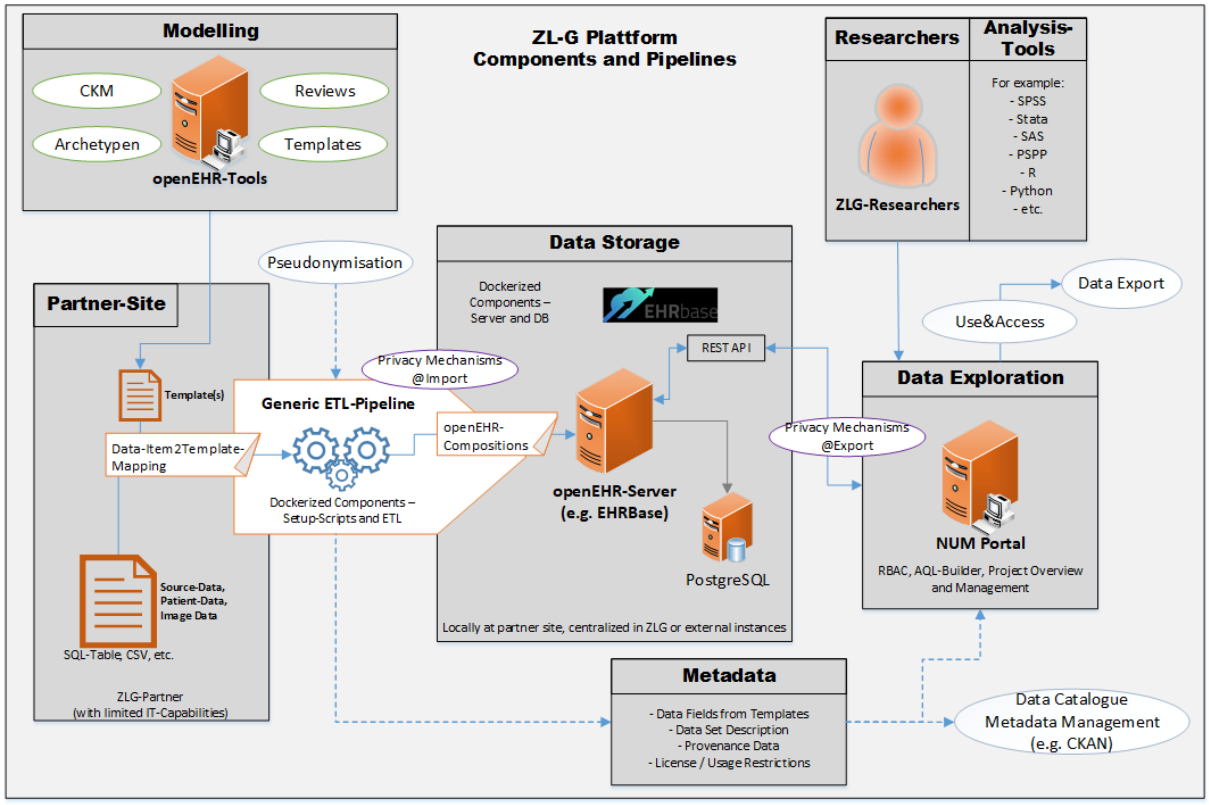

Abbildung 2: ZLG-Plattform Komponenten und Prozesse

Moderne Forschungsdatenplattformen sind oft modular und bestehen aus verschiedenen Komponenten. Dieser Design-Ansatz eignet sich auch für die Plattform des Zukunftslabors. Ein modularer Aufbau erlaubt, im Vergleich zu einem monolithischen Aufbau ein einfacheres Austauschen, Anpassen oder Erweitern der Plattform. Die eigenständige Entwicklung von Plattformkomponenten erlaubt ein flexibles Reagieren auf Anforderungen aus den Teilprojekten, sowie das Erfüllen der allgemeinen Anforderungen bezüglich Organisation und Verwaltung der Plattform. Dies wäre bei der Mitnutzung einer bestehenden Plattform nicht möglich. Der modulare Ansatz und die Verwendung gängiger Standards garantieren hierbei die spätere Erweiterbarkeit und Anpassbarkeit der entwickelten Lösung. Es wird versucht, den Austausch zwischen ähnlichen Plattformprojekten und das Ausnutzen von Synergieeffekten

[13] https://bitbucket.org/medicalinformatics/mainzelliste/src/master/ (abgerufen: 08.05.2021)

zwischen solchen Projekten (z.B. HiGHmed MeDICs) bei allen thematischen, infrastrukturellen und organisatorischen Überschneidungspunkten zu nutzen.

Backend

Zentrale Komponenten sind die Datenhaltungslösungen im sogenannten Backend der Plattform. Als openEHR-Server ist hier die EHRBase abgebildet, welche es erlaubt, openEHR-Ressourcen über die openEHR-REST-API entgegenzunehmen, unter Nutzung einer PostgresSQL-Datenbank abzuspeichern und über die openEHR-REST-API in der Archetype-Query-Language (AQL) wieder abfragbar zu machen. Im openEHR-Server sollen klinische Forschungsdaten in der logischen Struktur von Patientenakten / Electronic Health Records (EHRs) abgespeichert werden.

Frontend

Das Frontend der Plattform stellt die Benutzerschnittstelle zur indirekten Nutzung der Datenhaltungs-Komponenten im Backend dar. Das Frontend kann beispielsweise über einen Web-Server / Web-Applikation mit GUI oder durch Skripte für den Datenabruf umgesetzt werden. Voraussichtlich wird das NUM Portal[14] als Frontend verwendet. Dieses erlaubt die Anbindung eines openEHR-Repositoriums als Datenhaltung und darüber hinaus die Verwaltung von Projekten und Forschenden. Über eine Kohortensuche lassen sich Abfragen auf dem Datensatz im openEHR-Repositorium stellen, welche nicht die die Daten selbst zurückgeben, sondern Informationen über die Größe der verfügbaren Kohorte zur ausgeführten Suche.

Proxy

Über einen Proxy (z.B. NGINX[15]) kann die dahinterliegenden Plattform-Komponenten außerhalb des Host-Netzwerks kontrolliert erreichbar gemacht und abgesichert. Über die Anbindung z.B. der Let's-Encrypt-Services[16] können Zertifikate automatisiert erneuert und die Kommunikation SSL-verschlüsselt werden.

[14] https://github.com/num-forschungsdatenplattform (abgerufen 03.08.2022)
[15] https://www.nginx.com (abgerufen 07.02.2021)
[16] https://letsencrypt.org/de/ (abgerufen 07.02.2021)

openEHR-Governance

Die openEHR-Governance umfasst die Verwaltung, Erstellung und Anpassung von Modellierungsressourcen für die Verwendung von openEHR. Dies sind z.B. Software-Tools für die Erstellung von Archetypen und Templates oder ein Clinical Knowledge Management-Server, welcher erstellte Archetypen und Templates verwaltet und zugreifbar macht.

Die Anbindung eines Ontologie-Servers für die Bereitstellung von Ontologien in Form von Value Sets und Localizations ist möglich.

Identity Server

Die Anbindung von Authentifizierungs-Lösungen ist z.B. über OAuth2[17] möglich. Angebunden werden sollte ein Identitätsmanagement an eine entsprechende Schnittstelle des Frontends / Web-Server.

Plattform Management

Die Plattformverwaltung muss Prozesse für die Datenverwaltung, Metadatenverwaltung und den Datenzugriff etablieren. Hierbei handelt es sich nicht notwendigerweise um eine technische Lösung, diese Komponente kann auch aus organisatorischen Vorgaben wie SOPs und Leitlinien bestehen.

ETL-Komponente

Die Plattform soll nicht nur über die GUI des Frontend / Web-Applikation verfügbar gemacht werden, sondern auch die (teil-) automatisierte Anbindung weiterer Datenquellen ermöglichen. Daten können entweder über die openEHR-REST-Schnittstelle hochgeladen oder perspektivisch über die FHIR-Bridge[18] integriert werden. Beide Wege setzen voraus, dass die abzuspeichernden Daten in openEHR modelliert wurden und im EHR-Server ein entsprechendes Template für den Datensatz abgespeichert wurde bzw. dies vor dem Laden der Daten hochgeladen wird.

Trusted Third Party

Für die Haltung personenbezogener Patientendaten werden nach DSGVO bzw. den Bundes- und Landesdatenschutzgesetzen besondere Maßnahmen zur Gewährleistung der Rechte betroffener Personen benötigt. Das generische Datenschutzkonzept der TMF sieht die Pseudonymisierung von Datensätzen als notwendige Maßnahme für die Wahrung der Betroffenenrechte vor. Ein hierfür benötigter Pseudonymisierungsdienst, welcher die

[17] https://oauth.net/2/ (abgerufen: 08.05.2021)

[18] https://github.com/ehrbase/fhir-bridge (abgerufen: 08.05.2021)

Bereitstellung von Pseudonymen während des ETL-Prozesses bietet, muss durch eine organisatorisch von den anderen beteiligten Parteien getrennte Instanz (vertrauenswürdige dritte Partei) bereitgestellt werden (siehe Kapitel 6.3).

3.2.2 openEHR-Server

Durch die Konformität der openEHR-Server-Lösungen mit der openEHR-Spezifikation ist diese Plattformkomponente modular und damit austauschbar. Als openEHR-Server wird für die exemplarische Umsetzung der Open-Source-Server „EHRBase“ der vitagroup AG verwendet.

In Abbildung 3 ist der EHRBase Server als zentrale Komponente einer openEHR-basierten Plattform dargestellt. An den Server sind in dieser Abbildung ein „Identity Server“ für die Verwaltung und Authentifizierung von Benutzern angeschlossen, eine „Ontology Server“ für die Abfrage von Ontologien (wie z.B. SNOMED-CT, o.ä.), eine Clinical Modeling Komponente für die Erstellung von Archetypen und Templates (z.B. Better Archetype Editor[19], Ocean Editor/Designer[20]) oder ein Clinical Knowledge Manager (CKM) für die Verwaltung von Archetypen und Templates sowie eine Datenbank (in diesem Fall PostgreSQL) angebunden. Die Kommunikation zwischen den Komponenten erfolgt REST-basiert, entweder durch den Austausch von FHIR-Ressourcen oder openEHR-Ressourcen.

[19] https://tools.openehr.org/designer/#/ (abgerufen 08.02.2021)

[20] https://oceanhealthsystems.com/applications/template-designer/ (abgerufen 08.02.2021)

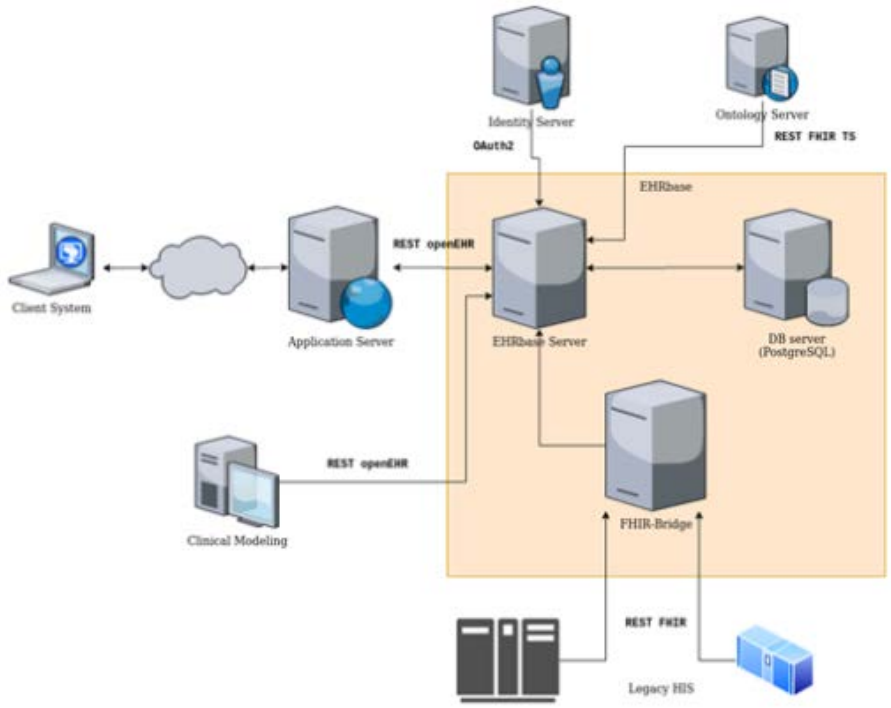

Abbildung 3: EHRBase Platform Context[21]

Durch die abgebildete „FHIR-Bridge“-Komponente können im FHIR-Format übermittelte Daten im openEHR-Format in der EHRBase abgespeichert werden. Dies geschieht innerhalb der FHIR-Bridge über ein hinterlegtes Mapping von FHIR-Profilen auf die in der EHRBase verwendeten openEHR-Templates.

[21] https://www.slideshare.net/openEHR-Japan/ehrbase-open-source-openehr-cdr (abgerufen am 07.01.2021)

In Abbildung 4 sind die Komponenten der EHRBase Plattformarchitektur abgebildet. Die Client-Library inzwischen als „openEHR_SDK“ [22] benannt ist in der Programmiersprache Java geschrieben. Mithilfe der Client-Library können openEHR-Ressourcen erzeugt, validiert und an einen Server übermittelt werden.

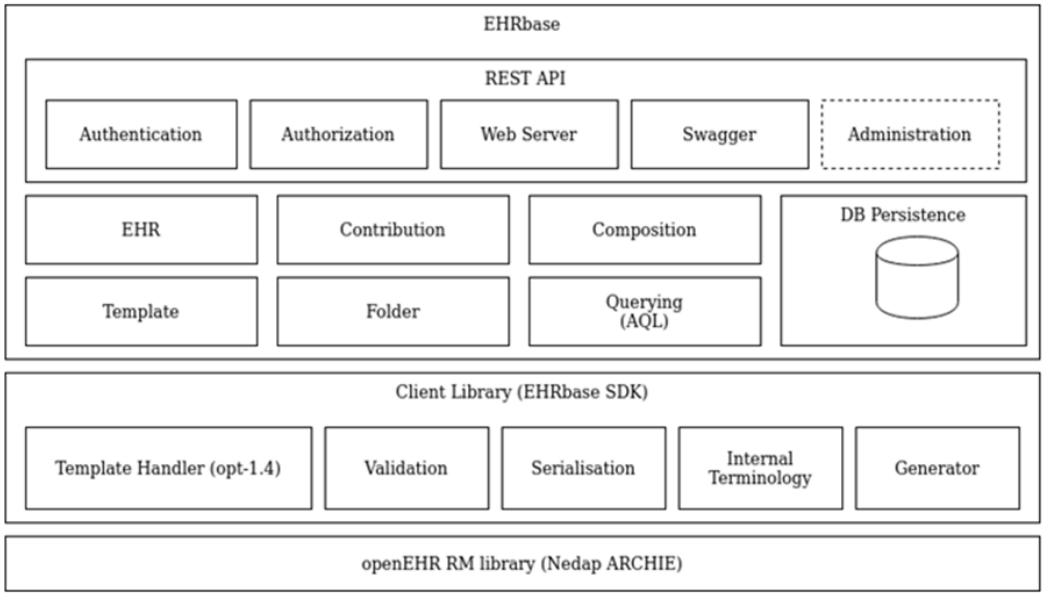

Abbildung 4: Plattformarchitektur der EHRBase

3.2.3 Datenintegrationspipeline

Die Datenintegrationspipeline stellt eine zentrale Komponente der Plattform dar, wie in Kapitel 3.2 erläutert und in Abbildung 2 abgebildet. Über die Transformation von Quelldaten in ein interoperables Datenformat und die nachfolgende Bereitstellung durch die Pipeline werden die Quelldatensätze weiternutzbar gemacht. Voraussetzung für die Weiterverwendung von entsprechenden Datensätzen ist eine entsprechende Zweckbestimmung und Einwilligungserklärung bei der Erfassung des Datensatzes. Als Beispiel für eine Einwilligung mit breiter Zweckbestimmung ist der Broad Consent[23] der MI-Initiative geeignet.

Für fest definierte openEHR-Templates lassen sich in verschiedenen Implementationssprachen ETL-Jobs bauen. Das hierfür benötigte (openEHR-Fachwissen) und die benötigte Zeit sind für eine spätere Plattform, in die zu beliebigen Templates Daten gespeichert werden sollen, nicht vernachlässigbar. Im Sinne der Automatisierung und Optimierung von organisatorischen und

[22] https://github.com/ehrbase/openEHR_SDK (abgerufen 08.02.2021)
[23] https://www.medizininformatik-initiative.de/de/mustertext-zur-patienteneinwilligung (abgerufen: 18.05.2021)

technischen Prozessen wird daher versucht, eine generische ETL-Pipeline zu implementieren, die den Import zu beliebigen Templates ermöglicht.

Der für die DI-Pipeline notwendige ETL-Prozess soll eine direkte Überführung von Quelldaten in das openEHR-Format erlauben. Durch umfassende Dokumentation soll die Pipeline auch durch Personen mit wenig ETL- oder openEHR-Erfahrung verwendbar sein.

Grundlage für die Auslagerung der Datenhaltungskomponente „EHR-Server“ und der zugehörigen ETL-Strecke bzw. Datenintegrations-Pipeline ist die einfache Verwendbarkeit der Lösungen. Insbesondere für Partner bzw. Mitarbeiter ohne umfassendes Wissen zu openEHR ist dies wichtig. Sichergestellt werden kann die Verwendbarkeit über gut dokumentierte Docker CLI-Container.

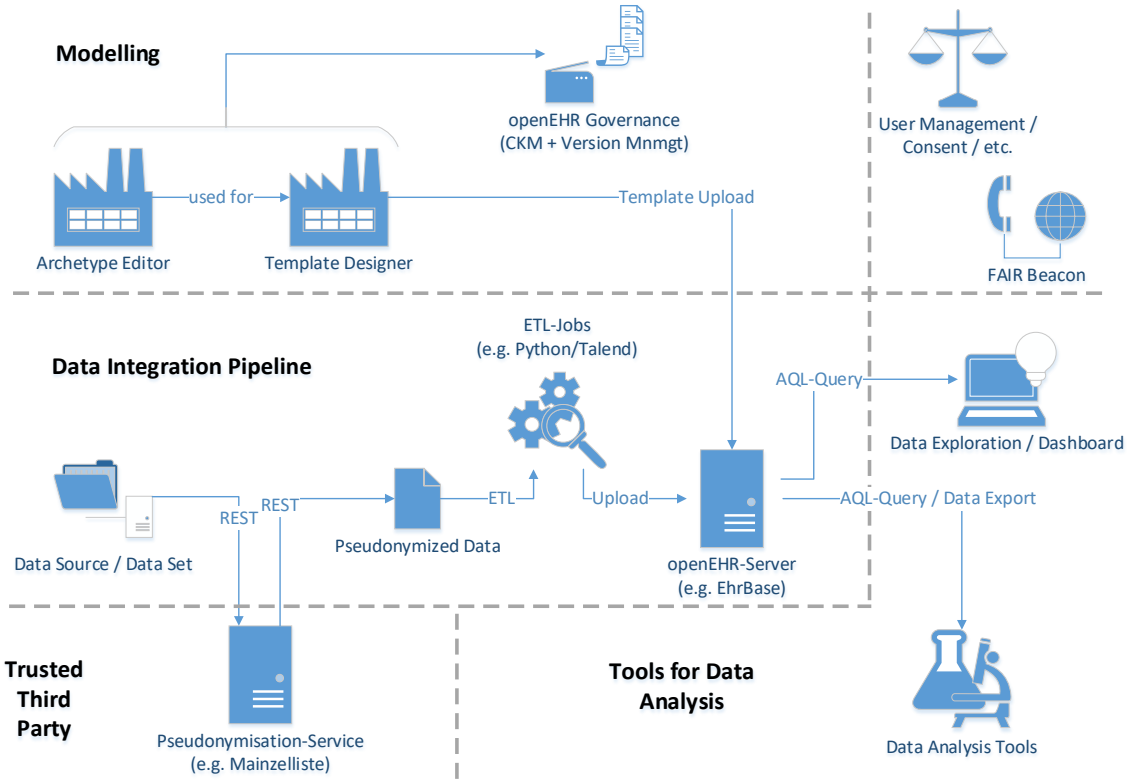

Abbildung 5: openEHR-Nutzungsprozesse mit "Data Integration Pipeline" (vereinfacht)

Umsetzungsmöglichkeit „openEHR_SDK“

Für die Umsetzung bietet sich die Verwendung einer openEHR-Library für die Verarbeitung von Templates, Archetypen, die Verwendung des Referenzinformationsmodells und letztlich die Erzeugung von Ressourcen / Compositions an. Die openEHR_SDK-Java-Library erlaubt die Erstellung von openEHR-Compositions. Mittels openEHR_SDK lassen sich zu bereitgestellten openEHR-Templates Java-Klassen generieren, die die Verwendung der Templates innerhalb der Java-Umgebung ermöglichen. Diese können mittels Java Reflection auch zur Laufzeit erzeugt und genutzt werden. Für die Umsetzung der Datenintegrations-Pipeline werden die generierten Java-Klassen verwendet.

Wenn der Nutzer sowohl das Template für einen Datensatz (OPT = Operational Template) und den Datensatz selbst (z.B. als CSV) bereitstellt, muss daraus automatisiert ein Mapping-Liste erzeugt werden, in welchem durch den Nutzer die Datenitems in der CSV den entsprechenden Datenfeldern des Template / der Java-Klassen zugeordnet werden. Mithilfe des durch den Nutzer bereitgestellten Mappings müssen für jeden Patienten / jedes Pseudonym Compositions erzeugt werden und an den openEHR-Server übermittelt werden.

HaMSTR ETL-Tool

Ein in diesem Zusammenhang interessantes Software-Tool ist HaMSTR[24] das Hannover Medical School Translational Research Framework[25]. HaMSTR ist eine prototypische Umsetzung eines ETL-Tools für die Überführung von Quelldaten in das openEHR-Format [4]. Aus einer Datenbank per SQL-Abfrage abgerufene Daten werden dabei durch eine nutzergenerierte Mappingtabelle in openEHR-Contributions zusammengefasst, welche anschließend im JSON-Format vorliegen und an einen openEHR-Server geschickt werden können.

Eine erste Inspektion des Programms ergab, dass hierbei auf das noch nicht im openEHR-Standard vorhandene FLAT-JSON-Format gesetzt wird, welches derzeit nicht durch alle openEHR-Server entgegengenommen werden kann. Die Übernahme des FLAT-Formats als SDT (Simplified Data Template) in die Spezifikation von openEHR steht derzeit noch aus[26]. Das verwendete Vorgehen deckt sich jedoch mit den Vorhaben für die Umsetzung einer

[24] https://gitlab.plri.de/tute/HAMSTRETLBuilder (abgerufen: 18.05.2021)
[25] https://plri.de/forschung/projekte/hannover-medical-school-translational-research-framework-hamstr (abgerufen: 18.05.2021)
[26] https://specifications.openehr.org/releases/ITS-REST/latest/simplified_data_template.html (abgerufen: 18.05.2021)

generischen (für beliebige Templates) ETL-Pipeline im Zukunftslabor, welches im Folgenden skizziert wird.

Mapping von Quelldaten auf openEHR-Templates

Für die direkte Erzeugung von openEHR-Compositions ist, beispielsweise mithilfe des openEHR_SDK, ein ETL-Prozess zu erproben, welcher aus beliebigen Quelldaten standardkonforme Compositions erzeugt und an den openEHR-Server schickt. Dies sollte über ein von Partner zu erzeugendes Mapping von Quelldaten auf ein zugehöriges openEHR-Template möglich sein (siehe Abbildung 6).

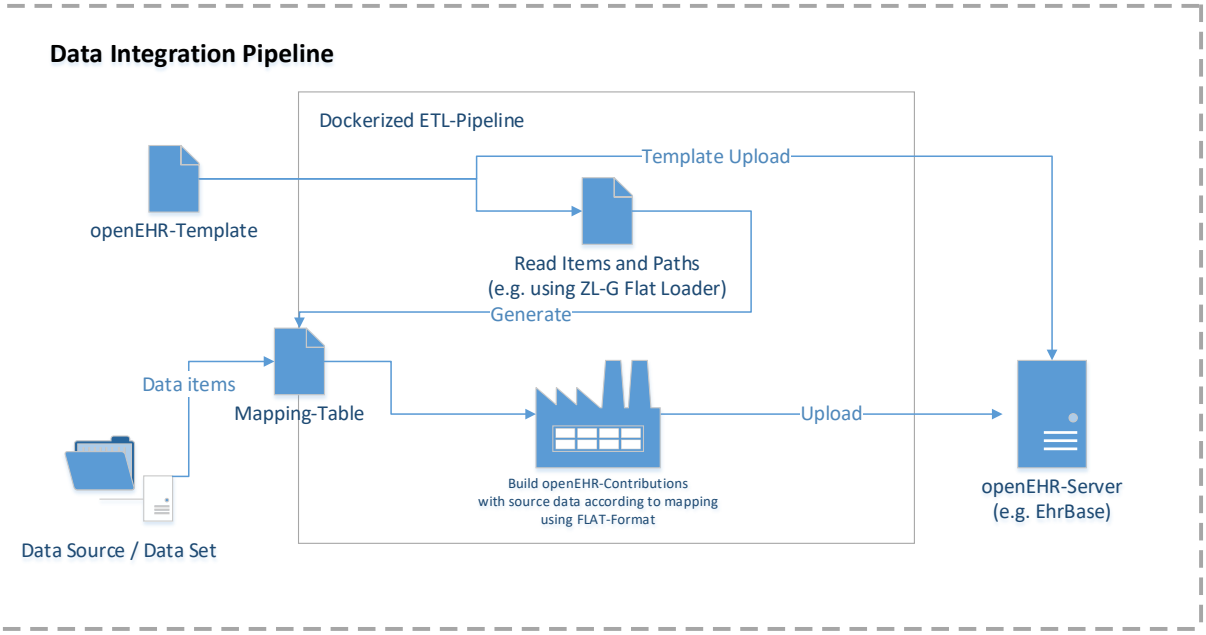

Abbildung 6: Überführung von Quelldaten in openEHR-Ressourcen via Mapping

Die Implementierung, Erprobung und Dokumentation dieser ETL-Pipeline ist die entscheidende Herausforderung für die Umsetzung der ZLG-Plattform bei den an der Infrastruktur bzw. an openEHR interessierten Partnern. Die verschiedenen Ansätze für die Umsetzung einer ETL-Strecke sind bereits in der technischen Erprobung. Einzelne verwendete Komponenten befinden sich allerdings noch in der Entwicklung und werden beständig weiterentwickelt (EHRBase, openEHR_SDK, Simplified-Data-Template). Erste Test-Implementationen ergaben, dass für die einfache Transformation von tabellarischen Input-Daten in das wenig komplexe FLAT-JSON-Format (vgl. SDT) beliebige Implementationssprachen geeignet sind und die Verwendung der openEHR_SDK-Library nicht zwingend notwendig ist.

3.2.4 Datenexploration

Für die Endnutzer (Data Scientists) ist eine gute Übersicht über verfügbare Daten sowie deren Metadaten ein wichtiges Qualitätsmerkmal einer Datenplattform. Datensätze müssen im Sinne der FAIR-Prinzipien auffindbar, zugreifbar, interoperabel und reproduzierbar sein.

Für die ZLG-Plattform ist die Erfassung von Datensatz-Metadaten im Rahmen des ETL-Prozesses bzw. Datenimports geplant. Dieser soll im Laufe der Entwicklungen in die Prozesse und Tools des ZLG integriert werden, um diese Prozesse für den Endnutzer einfach, übersichtlich und nachvollziehbar zu gestalten und dadurch die Usability zu erhöhen.

Perspektivisch ist das Vorhalten der Metadaten im openEHR-Format zu evaluieren, um den Pflege- und Wartungsaufwand der Infrastruktur zu minimieren. Bei eindeutig definierten Schnittstellen ohne viel Wartungsaufwand, kann sich eine modulare Umsetzung mit openEHR-Datenrepositorium und davon abgegrenzt dem Datensatzkatalog allerdings ebenfalls als geeignet erweisen.

3.2.5 Datenabruf/Datenexport

Der Abruf von Daten aus einem openEHR-Repository geschieht durch AQL-Queries[27]. AQL, Akronym für Archetype Query Language, ist eine SQL-ähnliche Abfragesprache, in welcher Items über AQL-Pfade für die Abfrage aus dem Repository genutzt werden. AQL-Pfade verbinden die logische Repräsentation der Datenitems in Archetypen/Templates mit dem physischen Speicherort im openEHR-Repositorium.

AQL-Abfragen können händisch auf Basis vorhandener Templates geschrieben oder bei Vorhandensein eines entsprechenden Tools in einem AQL-Query-Builder aus Templates zusammengestellt werden. Ein im Rahmen des NUM CODEX entwickeltes Web-Portal[28] könnte hier die erste generische Portal-/Dashboard-Lösung für die Anbindung und Nutzung eines EHR-Backends für die Forschung darstellen. Das Portal erlaubt den Zugriff auf openEHR-Server über vordefinierte AQL-Abfragen oder die Erzeugung von AQL-Abfragen per Drag & Drop aus ausgewählten openEHR-Templates. Weiterhin wird die Verwaltung von Nutzern und Rollen sowie openEHR-Templates und Studien ermöglicht. Die Bereitstellung als Open-Source Anwendung ist geplant.

[27] https://specifications.openehr.org/releases/QUERY/latest/AQL.html (abgerufen: 16.06.2021)
[28] https://github.com/num-codex/codex-feasibility-gui (abgerufen: 09.05.2021)

Für die Nutzung des openEHR-Backends bzw. den Zugriff auf die Plattform-Funktionalitäten durch die Endnutzer (Data Scientists) im Zukunftslabor kann perspektivisch auf dem NUM-Portal aufgebaut werden. Für die erste Iteration der Umsetzung bietet sich für die Abfrage von Daten die Verwendung vordefinierter AQL-Abfragen sowie des ETL-Tools für den Datenimport an, um in einem ersten Data-In und Data-Out die technische Umsetzung zu erproben.

3.2.6 Datenanalyse

Die Daten werden mit statistischen oder Machine Learning (ML) Modellen analysiert. Hierzu eignen sich grundsätzlich alle gängigen Anwendungen und Sprachen zur Datenanalyse, wie STATA, SAS, SPSS, R oder Python. Für Zwecke der Datenanalyse sind alle für die Analyse relevante Daten aus dem openEHR-Format in flache Formate zu bringen, d.h. strikt tabellarisch, so dass Fälle in den Zeilen und Fallmerkmale in den Spalten einer Tabelle angeordnet sind. Flache Formate werden von den meisten statistischen und ML-Prozeduren als Input erwartet. Ausnahmen können graphenbasierte Analysen sein, die aber ebenfalls eine stark vereinfachte Sicht auf die Daten erfordern, in welcher Knoten, Kanten und ggf. Kantengewichte zu spezifizieren sind. In beiden Fällen müssen aus den Daten im openEHR-Format die interessierenden Fälle, Merkmale und Beziehungen extrahiert, selektiert und vereinfacht dargestellt werden.

Die Transformation in ein solches tabellarisches oder graphenbasiertes Format muss durch einen gesonderten ETL-Job erfolgen, wenn Tools eingesetzt werden, die über entsprechende Funktionen nicht verfügen (SPSS, STATA, ARX-Tool). Entsprechende ETL-Jobs sollten im CSV-Dateien speichern.

Werden hingegen Programmiersprachen wie R oder Python eingesetzt, kann die Datenextraktion und -transformation unmittelbar durch die Auswertenden erfolgen und eine gesonderte Aufbereitung per ETL-Job zwischen Abruf und Bereitstellung ist nicht zwangsläufig erforderlich.

Die weitere Verwendung der Ergebnisse oder ein eventuelles Zurückspielen von Ergebnissen ist in der ersten Iteration der Plattformumsetzung nicht im Scope des Zukunftslabors.

Privatheitsbewahrung bei der Analyse

Für die Bewahrung der Privatheit der Patienten / Probanden zu deren Behandlung oder durch deren Mitwirkung die Daten erzeugt wurden sollen privatheitsbewahrende Mechanismen zum Einsatz kommen. Die im ZLG erprobten Mechanismen lassen sich in zwei Bereiche unterteilen

1. Privacy-preserving Mechanisms (k-Anonymität, l-Diversität, t-Nähe)
2. Differential Privacy

Die beiden Bereiche unterscheiden sich im Hinblick auf 1. privatheitsbewahrenden Mechanismen, die die Privatheit der Daten gewährleisten sollen und 2. der Anpassung der Datenanalysen selbst. In Abhängigkeiten von Definition und Grad der Privatheit, die gewährleistet werden soll, gehen die privatheitsbewahrenden Mechanismen der Datenanalyse voran oder müssen während der Datenanalyse erfolgen.

Bei der Anonymisierung von Daten sollte die Forschungsfrage bereits bekannt sein, da sonst leicht wichtige Zusammenhänge in den Daten durch die Anonymisierung verloren gehen können. Hier findet ein Trade-Off zwischen Privatheit und Potenzial der Daten für Analysezwecke statt. Aus diesem Grund ist die Untersuchung dieser Abwägung zwischen Privatheit und dem Grad der Privatheit in verschiedenen Schritten des Prozesses, auch in Kombination mit neuartigen Ansätzen und Algorithmen, ein zentraler Aspekt der Arbeiten im ZL-G.

Privatheitsbewahrung beim Datenimport

Der Datenanalyse vorangehende Mechanismen, die Privatheit garantieren, sind der einfachere Fall, da dieser Schritt in der Prozesskette einfach vorgelagert werden kann und hierfür nur eine Auswahl aus bestehenden Tools oder Implementierungen zu treffen ist. In TP1 wurde hierfür das ARX-Tool für die Ansätze k-Anonymity, l-Diversity und t-Closeness an den MIMIC-Daten erprobt. Für diesen Prozessschritt müssen die zu analysierenden Daten bereits in einem flachen Format (z. B. CSV) vorliegen.

Privatheitsbewahrung beim Datenexport

Andere Begriffe von Privatheit, wie insbesondere Differential Privacy erfordern es, dass Algorithmen zur Datenanalyse selbst modifiziert werden, da ansonsten die Genauigkeit der Analysen zu sehr durch die Privatheitsgarantien beeinträchtigt würde. Da diese Ansätze noch sehr jung sind und in gängiger Anwendersoftware keine entsprechenden Funktionen bereitstehen, müssten solche privatheitsbewahrenden Analysen für die Plattform neu

implementiert werden. Entsprechende Skripte werden in TP1 in Python implementiert und könnten in eine ETL-Pipeline eingebunden oder bereits beim Datenimport eingebunden werden.

3.3 Roadmap

Die Roadmap orientiert sich am Vorgehensplan für den Use Case „Datenanalyse“. Der Vorgehensplan ist innerhalb des TP1 abgestimmt und stellt die Ziele bezüglich der praktischen Plattformumsetzung der Partner im Teilprojekt 1 dar.

Vorgehensplan des Use Case „Datenanalyse“:

1. Auswahl, Erhalt und Explorieren von Daten / Quellformat
2. Daten modellieren bzw. bestehende Modellierung prüfen
3. Daten in openEHR-Repositorium laden (ETL-Strecke, Metadaten und Pseudonyme)
4. Daten aus openEHR-Repositorium abfragen (Export und Schnittstellen)
5. Import der Daten in Analysetools
6. Privatheitsbewahrende Datenanalyse

Die Arbeiten lassen sich in die drei Kategorien „Analyse“, „Technik“ und „Governance“ sowie „Validierung“ unterteilen. Im Projektverlauf lassen sich die Aufgaben jahresweise voneinander abgrenzen (siehe Tabelle 1).

Tabelle 1: Roadmap mit Aufgaben und Kategorisierung

Jahr 1	Vorbereitung und Planung	[Gesamt-ZL]
Jahr 2	Planung, Konzeption und Auswahl	[Analyse] [Technik] [Governance]
Jahr 3	Schnittstellen, Test und Validierung	[Analyse] [Technik] [Validierung]
Jahr 4	Fortgeschrittene Analysemethoden, Rollout und technische Weiterentwicklung	[Analyse] [Technik]
Jahr 5	Abschluss der Weiterentwicklung und Freigabe und Ergebnisaufbereitung	[Analyse] [Technik]

Für Jahr 1-3 ist die Planung und Implementierung von Basisfunktionen, ETL-Skripten und (Meta-) Datenschemata sowie jeweils zugehörigen Prozessen geplant. In Jahr 3 kommt die Kategorie „Validierung“ hinzu. Die Jahre 4-5 werden für die Erweiterung und weitere Ausarbeitung der Lösungen sowie für die Translation der Ergebnisse genutzt.

3.4 Evaluation und Weiterentwicklung

Nach der initialen Inbetriebnahme und Erprobung einer Instanz der Plattform bzw. Plattformwerkzeuge findet eine interne Evaluation der Nutzbarkeit statt. Diese wird ein einem noch auszuarbeitenden Evaluationskonzept beschrieben und darauffolgend umgesetzt. Im Anschluss an die Evaluation ist eine Weiterentwicklung der Plattformwerkzeuge und zugehörigen Dokumentation basierend auf den Evaluationsergebnissen geplant.

Die angestrebte Infrastruktur muss folgende Kern-Funktionen umsetzen:

1. **Datenhaltung**
   Forschungsrohdaten sowie Metadaten und Bilddaten müssen interoperabel, integer und sicher gespeichert werden.

2. **Auffindbarkeit und Durchsuchbarkeit**
   Gespeicherte Daten(-sätze) müssen für Nutzer auffindbar sein. Die Datensätze müssen über umfassende Metadaten verfügen, welche für den Nutzer einsehbar sein müssen.

3. **Datenmodellierung und Datenimport / Export**

   Zu speichernde Datensätze müssen standardbasiert modelliert werden. Nutzer müssen über definierte Prozesse (Meta-)Daten und ohne großen ETL-Aufwand Daten importieren und exportieren können.

4. **Datenanalyse**

   Die gespeicherten Daten sollen mithilfe von Analysetools privatheitsbewahrend analysierbar sein. Zugriffs- und Bereitstellungsprozesse müssen definiert und beschrieben sein.

5. **Dokumentation**

   Alle Funktionen müssen übersichtlich und nachvollziehbar dokumentiert sein. Auch unerfahrenen Partnern muss es möglich sein, die Plattform lokal zu nutzen und zugehörige Prozesse umzusetzen.

6. **Verwaltung- und Wartung**

   Die Verwaltung / Governance, Erstellung und Bereitstellung der Datenmodelle (openEHR-Governance / Datenmodellierung) und die technische Bereitstellung und Wartung aller Komponenten muss mit einem geringen Personalaufwand zu bewältigen sein.

7. **Reife und Open-Source**

   Die Plattform muss eine den Nutzungsszenarien angemessene Reife haben. Im Hinblick auf die zukünftige Weiterentwicklung und Open-Science-Ansätzen sollen entwickelte Komponenten auf freien Lösungen aufbauen und selbst frei verfügbar sein.

Aus Nutzersicht umfasst eine mögliche Plattform-Architektur einer Forschungsdatenplattform eine Datenhaltung mit entsprechenden Zugriffsmöglichkeiten auf die Funktionalitäten für Nutzer. Dementsprechend müssen die Hauptkomponenten der Plattform erstens die Datenhaltung und zweitens ein Web-Portal für den Zugriff sein.

4 Use Case „Datenanalyse im TP1“

Im Zukunftslabor wurde als medizinischer Kontext der Bereich kardiologischer Erkrankungen gewählt. Das Problemfeld betrifft besonders ältere Patienten über 65 Jahren. Kardiologische Erkrankungen stellen mit ca. 40% aller Todesfälle in Deutschland die häufigste Todesursache dar[29]. Besonders häufige kardiologische Erkrankungen, wie Hypertonie, Koronare Herzkrankheit und Herzinfarkt, Herzinsuffizienz und Endokarditis lassen sich hier beispielhaft nennen. Krebserkrankungen folgen mit 26% aller Tode an zweiter Stelle.

Durch Vorarbeiten im Projekt HiGHmed, in denen ein Use Case Kardiologie umgesetzt wurde, ergeben sich hier Synergien zwischen den Projekten bezüglich Modellierung und Datenverfügbarkeit.

4.1 Vorgehensplan

Die Ausarbeitung und Beschreibung des Use Case „Datenanalyse“ grenzt die zu untersuchenden Teilprozesse und Komponenten der Plattform ein. Der Use Case „Datenanalyse“ wird die beispielhafte Datenintegration von Forschungsdaten aus dem medizinischen Bereich in ein standardbasiertes Datenformat und interoperables Datenrepositorium sowie die anschließende Abfrage der Daten aus dem Repositorium und Nutzung der Daten zu Analysezwecken beinhalten. Bei der Umsetzung des Use Case erfolgt die Erforschung der dazugehörigen Technologien und Prozesse.

Vorgehensschritte im Plattform-Use Case „Datenanalyse“

1. Auswahl, Erhalt und Explorieren der Daten (Metadaten berücksichtigen)
2. Daten modellieren bzw. bestehende Modellierung prüfen
3. Daten in openEHR-Repositorium laden (Pseudonymisierung berücksichtigen)
4. Daten aus openEHR-Repositorium abfragen (Export und Schnittstellen)
5. Import der Daten in Analysetools
6. Privatheitsbewahrende Datenanalyse

[29] https://www.rki.de/DE/Content/GesundAZ/H/Herz_Kreislauf_Krankheiten/Herz_Kreislauf_Krankheiten_node.html (abgerufen 02.05.2021)

4.2 Beispieldaten

Derzeit steht die privatheitsbewahrende Analyse des MIMIC-III[30] Datensatzes im Mittelpunkt. Der Datensatz ist frei verfügbar (abgesehen von Einschränkungen, die sich bei der Verwendung der Daten für Veröffentlichungen ergeben). Parallel wurde bereits mit der Modellierung und der Umsetzung von ETL-Prozessen rund um den NATARS-Studiendatensatz in openEHR begonnen.

Die Daten mit kardiologischem Bezug sollen Versorgungsdaten aus der Universitätsmedizin Göttingen (UMG) sein. Diese wurden unter Verwendung des sogenannten Broad Consent erfasst, welcher eine Nachnutzung für Forschungszwecke erlaubt ohne den Forschungszweck bereits bei der Erteilung zu benennen. Für den Erhalt der Daten wurde ein Datennutzungsantrag an das UMG MeDIC gestellt. Für ein zum Vorhaben gehöriges Studienprotokoll mit Studienleiter Prof. Dr. Wolfgang Nejdl (Direktor des ZL-G Partners *L3S* in Hannover) wurde ein positives Ethikvotum der Ethikkommission der LUH eingeholt. Für den Datennutzungsantrag beim UMG MeDIC wurde ein positives Übernahmevotum der Ethikkommission in Göttingen eingeholt. Die beantragten Daten sollen auf einem Analyseserver des UMG MeDIC bereitgestellt werden, sodass Forscher des ZL-G per VPN die Daten explorativ analysieren können.

4.3 Akteure und Rollen

Im Rahmen des Forschungsdatenmanagements und der dazugehörigen Prozesse existieren verschiedene Rollen. Diese sind in der Literatur bereits oft beschrieben [3,5]. Im Kontext FAIR und Forschungsdatenmanagement finden diese in unterschiedlichen Ausprägungen immer wieder Verwendung. Im Kontext des beschriebenen Use Cases lassen sich folgende Rollen als relevant identifizieren. Im Rahmen der Dokumentation der Plattform sowie des Governance-Frameworks sind diese Rollen zu verwenden und im Hinblick auf die späteren Nutzungsprozesse zu beschreiben:

[30] Medical Information Mart for Intensive Care: https://mimic.physionet.org (abgerufen 03.02.2021)

- **Datennutzung (Erfassung und Analyse)**
    - **Datennutzende**

      Datenanalysten bzw. allgemein Forschende benötigen Möglichkeiten Daten zu suchen, zu finden, zu erhalten und mit geeigneten Werkzeugen zu analysieren. IT-Unterstützung muss für sie einfach, verständlich und leicht nutzbar sein. Die Unterstützung bei der Nutzung vorhandener Infrastrukturen ist für diese Gruppe von großer Bedeutung.

    - **Datenbereitstellende**

      Personen die Daten bereitstellen, fragen diese entweder als bestehende Daten aus Quellsystemen ab oder wirken bei der strukturierten Erfassung und Speicherung von Daten mit.

- **Datenmodellierung und ETL (Bearbeitung und Unterstützung)**
    - **Datenmodellierende**

      Interoperabilität und damit Austauschbarkeit von Daten wird über die Verwendung einheitlicher Standards und Terminologien ermöglicht. Bei der Modellierung von Daten werden die Datenformate in der jeweiligen Repräsentation des Modellierungsansatzes definiert. Hierbei muss umfassendes Wissen zum verwendeten Standard sowie Domänenwissen verwendet werden.

    - **Datentransformierende (ETL-Devs)**

      Die technische Überführung von bereitgestellten Quelldaten in standardkonforme Ressourcen mit einheitlicher und korrekter Verwendung der festgelegten Terminologien sowie die Erstellung und Speicherung der zugehörigen Metadaten ist Teil des ETL-Prozesses.

- **Infrastruktur und Governance (Steuerung)**
    - **Datenmanagende (Data Stewards)**

      Das Management einzelner Datensätze umfasst die ganzheitliche Begleitung der Daten von der Erfassung bis zur Nutzung. Der Weg den die Daten nehmen werden muss geplant, kommuniziert und überwacht werden.

- **Administrierende (Server and System)**

  Die Verwaltung, Wartung und der Betrieb der technologischen Komponenten liegen im Fokus dieser Rolle. Nutzer müssen betreut, Nutzeraccounts müssen eingerichtet werden, Daten gesichert und archiviert werden.

- **Datenausschuss (Use and Access-Komittee)**

  Der Zugriff auf die in der Plattform hinterlegten Daten geschieht nach festgelegten Regeln. Erfolgt die Verteilung von Zugriffsrechten nicht automatisiert, müssen Anfragen zum Export von Datensätzen hinsichtlich der rechtlichen und ethischen Zulässigkeit geprüft werden.

- **Platformbetreibende**

  Der Betrieb der Plattform umfasst die Definition von Regeln für die Governance, Finanzierung der Infrastruktur und die Steuerung und Verwaltung des meist langfristigen Betriebs.

Die Forscher des Teilprojekts 1 entwerfen, testen und optimieren in dieser Umsetzungsphase die Nutzungsprozesse rund um die entstehenden ZL-G Plattformwerkzeuge. Die begleitende und anschließende Dokumentation der Prozesse ermöglicht die Weiternutzung und Weiterentwicklung der verwendeten Komponenten.

Der beschriebene Anwendungsfall (Use Case) „Datenanalyse“ umfasst die exemplarische Umsetzung von Prozessen durch einen kleinen Beteiligtenkreis für (eine in erster Instanz experimentelle) Nutzung. Ziel im Use Case ist, dass Datennutzende aus dem Teilprojekt 1 auf Beispieldatensätzen Analysen durchführen können. Die Beispieldatensätze sollen dafür standardbasiert modelliert und abgespeichert sein. Dies schließt als Beteiligte die Rollen "Datennutzende", „Datenmanagende“, "Datentransformierende" und „Administrierende“ ein.

5 Nutzungsprozesse

Die im Folgenden beschriebenen exemplarischen Nutzungsprozesse bilden die Umsetzung der Vorgehensschritte des Use Case „Datenanalyse“ ab. Die Prozesse werden an dieser Stelle inhaltlich dargestellt und in Zukunft in der Implementierungs-Guideline TA1.4.2 dokumentiert. Die Dokumentation der Prozesse stellt hinsichtlich der beabsichtigten „Befähigung der Partner zur Nutzung von openEHR“ ein wichtiges Ergebnis dar.

Im Hinblick auf die Nutzung der am Ende stehenden Plattform sind die in Kapitel 5.3 benannten Rollen relevant. Die zu verfassenden Anleitungen zu Inbetriebnahme und Betrieb der Plattformwerkzeuge sind insbesondere als Referenz für die Rolle „Administrierende“ verfasst sowie für die Rolle “Datentransformierende”. Prozesse bezüglich Privatheit, Zugriffsrechte und allgemein die Governance der Plattform adressieren eher die Rollen „Datenmanagende“ / Data Steward und „Platformbetreibende“.

5.1 Prozess 1: Modellierung

Abbildung des Nutzungsprozesses: Die folgende Abbildung stellt die Beteiligten, Abläufe und relevante Werkzeuge dar.

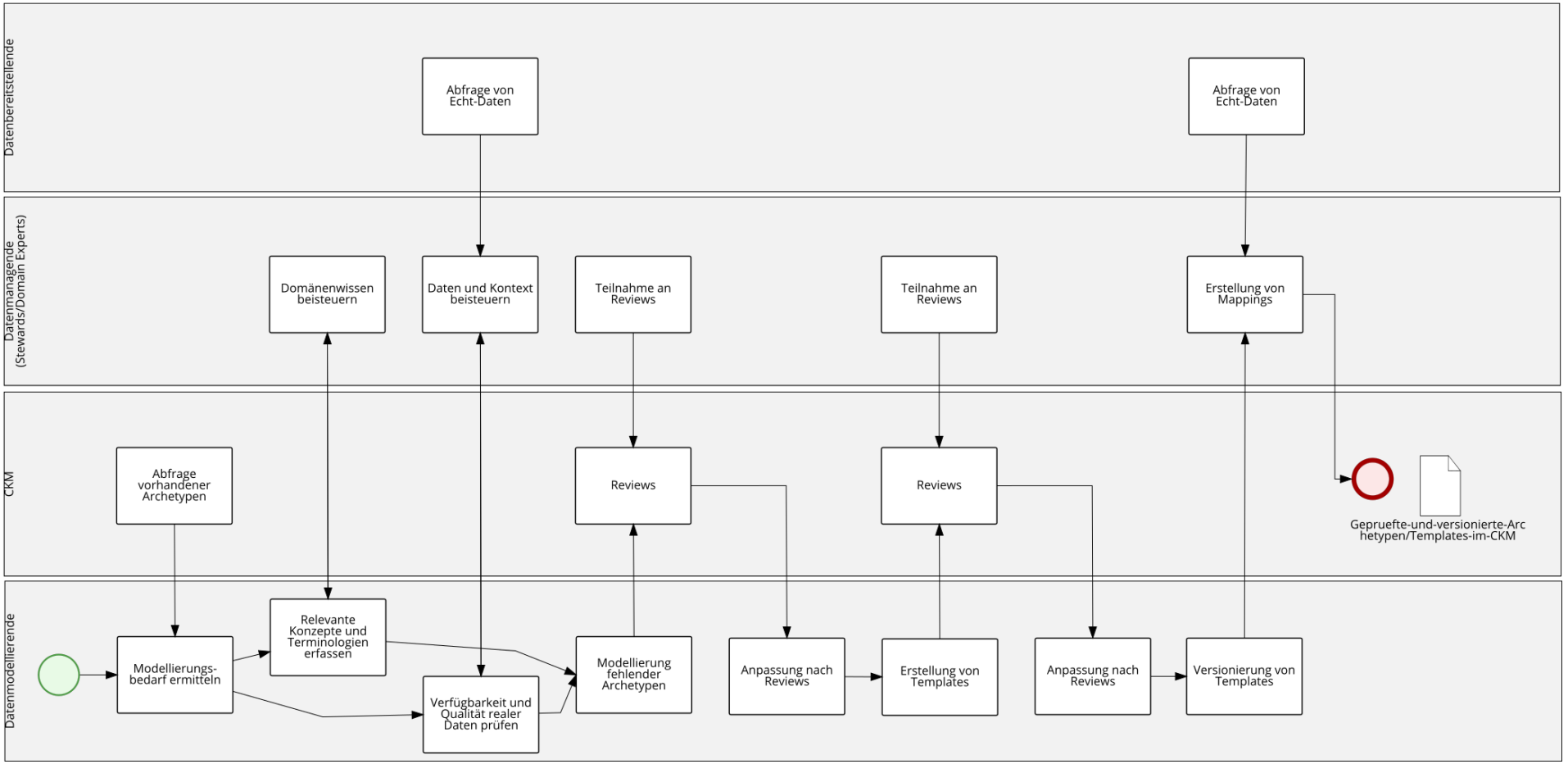

Abbildung 7: Abbildung des Nutzungsprozesses zur Modellierung

5.2 Prozess 2: Datenimport

Abbildung des Nutzungsprozesses: Die folgende Abbildung stellt die Beteiligten, Abläufe und relevante Werkzeuge dar.

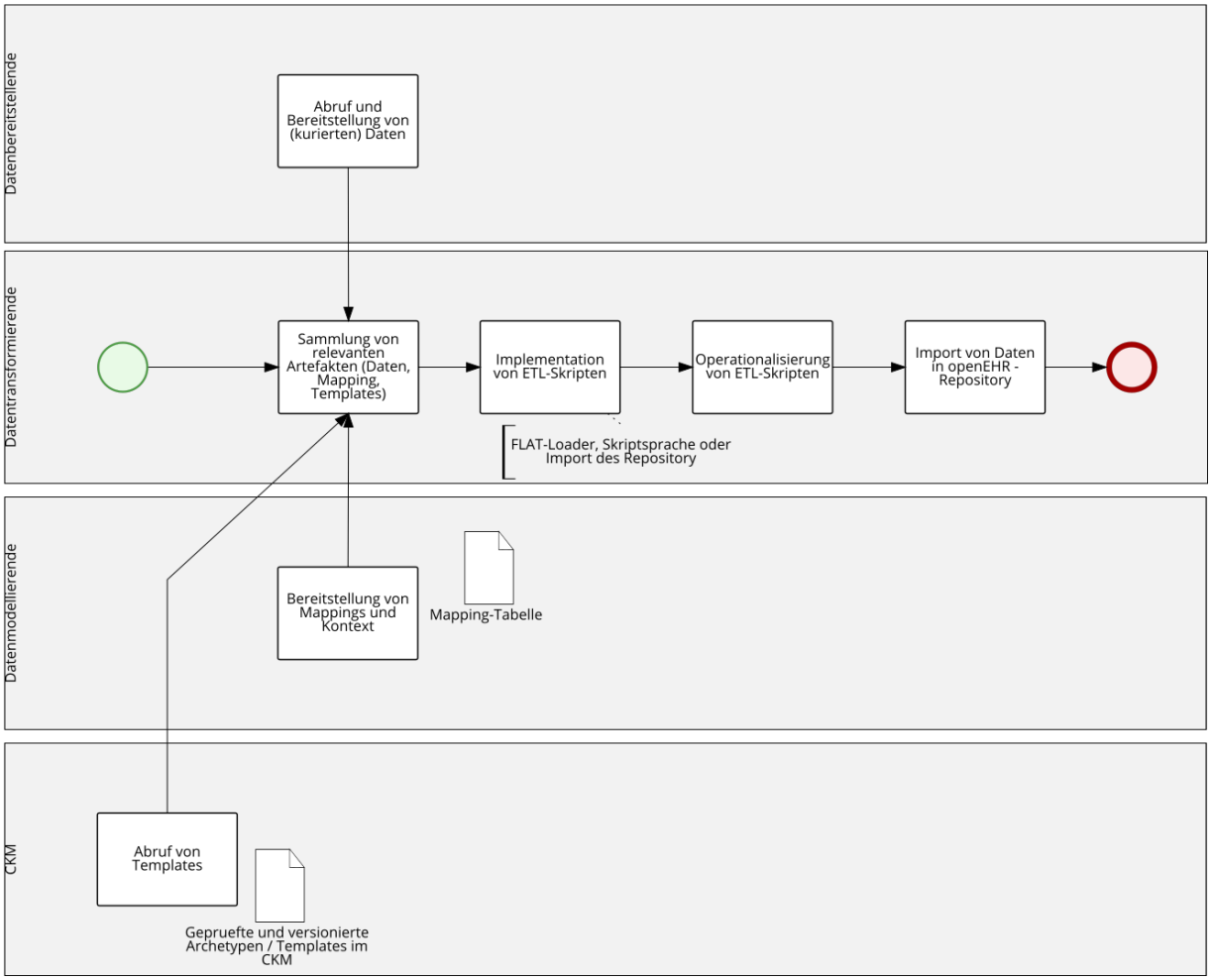

Abbildung 8: Abbildungs des Nutzungsprozesses zum Datenimport

5.3 Prozess 3: Datenexploration

Abbildung des Nutzungsprozesses: Die folgende Abbildung stellt die Beteiligten, Abläufe und relevante Werkzeuge dar.

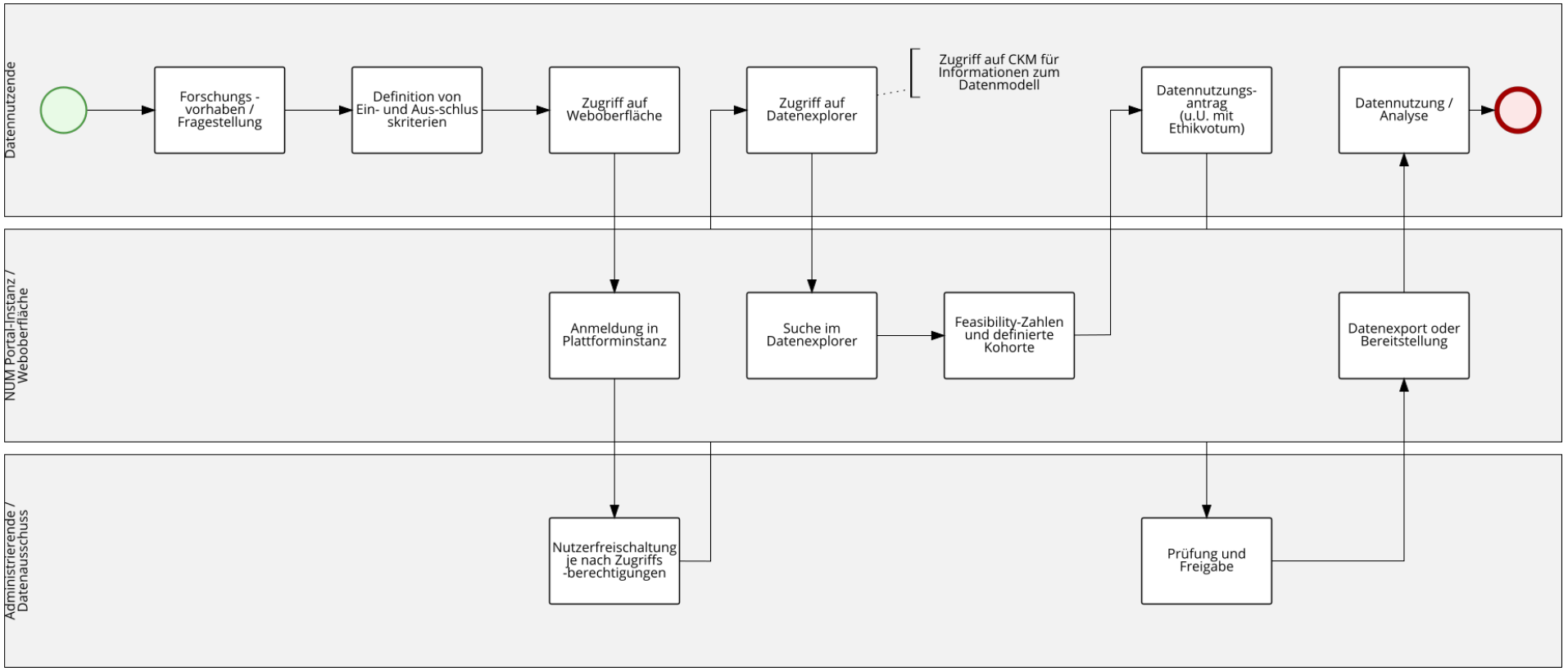

Abbildung 9: Abbildung des Nutzungsprozesses zur Datenexploration

5.4 Prozess 4: Datenabruf/-export

Abbildung des Nutzungsprozesses: Die folgende Abbildung stellt die Beteiligten, Abläufe und relevante Werkzeuge dar.

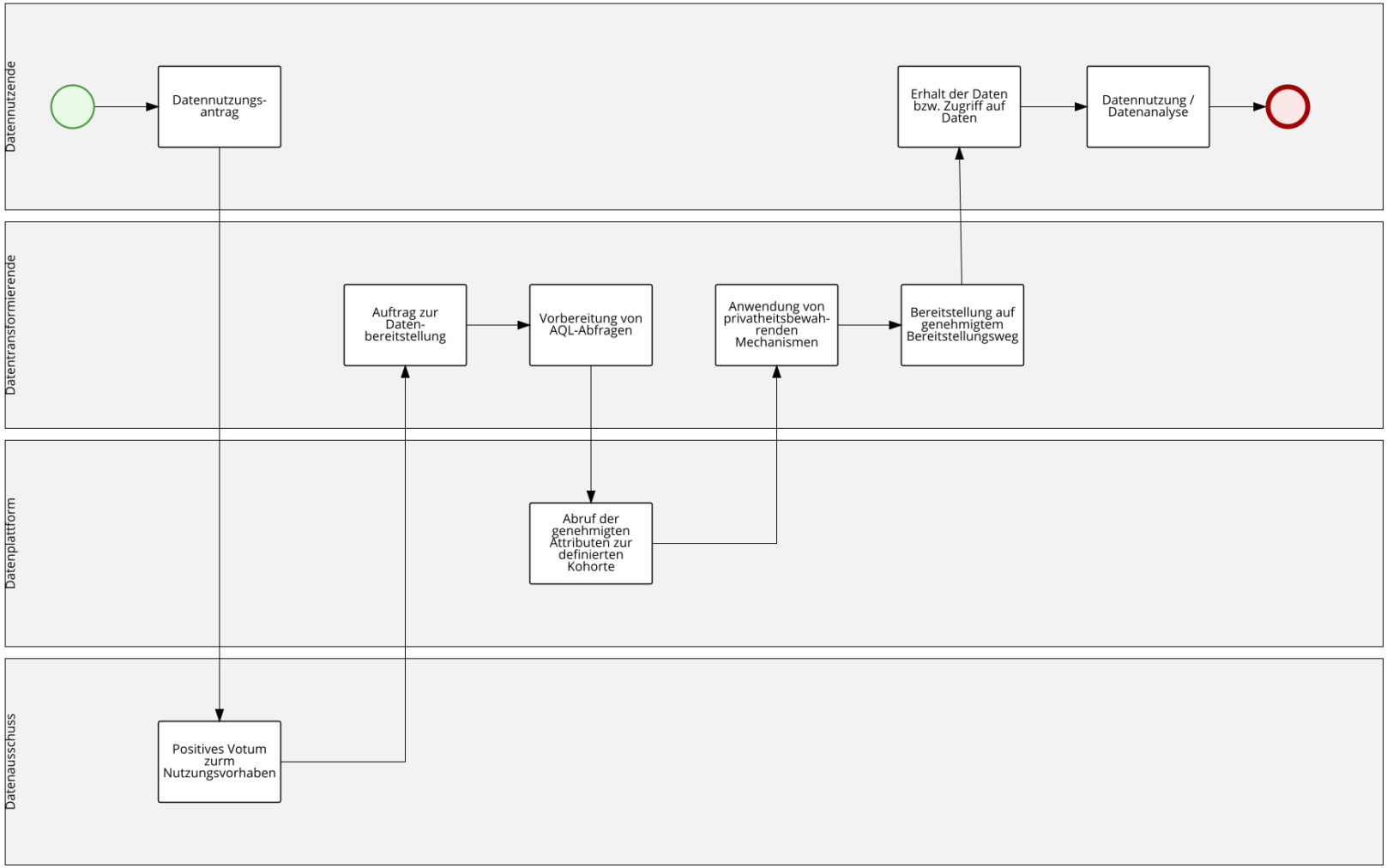

Abbildung 10: Abbildung des Nutzungsprozesses zurm Datenabruf

5.5 Prozess 5: Datenanalyse

Abbildung des Nutzungsprozesses: Die folgende Abbildung stellt die Beteiligten, Abläufe und relevante Werkzeuge dar.

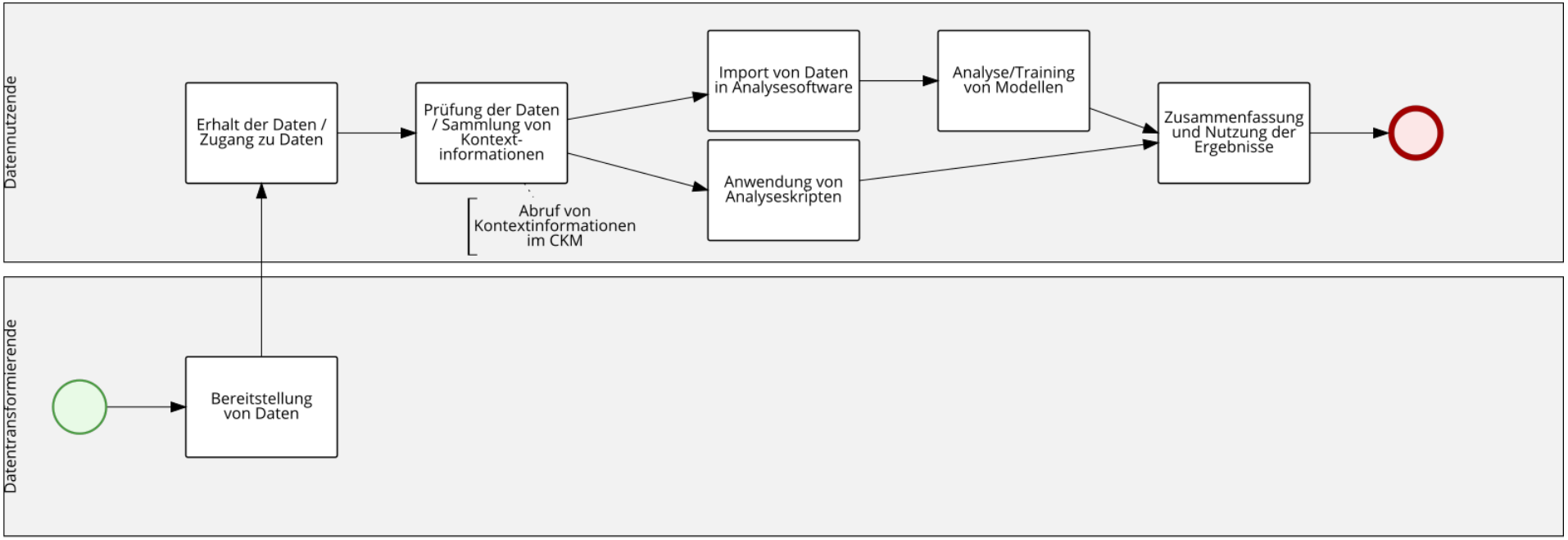

Abbildung 11: Abbildung des Nutzungsprozesses zur Datenanalyse

6 Weiterführende Arbeiten

6.1 Dokumentation zur Plattforminbetriebnahme

Bei der Inbetriebnahme einer Plattform für die Verwaltung von Forschungsdaten müssen zuerst standortspezifische Anforderungen ermittelt und analysiert werden. Auf Basis dieser Anforderungen kann eine Auswahl des zu verwendenden Modellierungsansatz und nachfolgend der zu verwendenden Werkzeuge erfolgen.

Die Auswahl von openEHR zum Aufbau einer openEHR-basierten Plattform basiert meist auf dem Vorhandensein entsprechender Modelle oder dem Einplanen und Vorhandensein von Modellierungskapazitäten in enger Zusammenarbeit mit Domänenexperten (Vergleiche HiGHmed). Im Zukunftslabor wird dazu im Projektverlauf eine Empfehlung zur Modellierung von Gesundheitsdaten erarbeitet, welche die Zusammenhänge, Erfahrungen und Besonderheiten verschiedener Modellierungsansätze darstellt.

Mit Vorhandensein bzw. Entstehen der Modelle kann die Quelldatentransformation / ETL-Prozesse entwickelt und etabliert werden. Die Verfügbarkeit (technisch sowie organisatorisch / rechtlich) und Qualität der Quelldaten ist dabei zentral für die Nutzbarmachung der Daten in der interoperablen Plattform. Parallel muss der Betrieb und die Wartung der genutzten Serverinfrastruktur und Netzwerkkomponenten, in welchen das Repository, ETL-Prozesse betrieben werden und der letztendliche Datenzugriff / Datennutzung erfolgt.

Die konkreten Schritte für Inbetriebnahme und Betrieb sind im weiteren Projektverlauf auszuarbeiten und zu dokumentieren.

6.2 Governance Framework des ZL-G

Das ZL-G hat für die Verwendung von openEHR innerhalb des ZL-G beziehungsweise in der ZL-G Plattform ein Governance Framework für die Organisation der Modellierung und die Verwaltung der openEHR-Artefakten entwickelt. Das Governance-Framework ist im Dokument „Governance Framework des Zukunftslabors Gesundheit“[31] niedergeschrieben.

Das im TP1 durch die MHH entwickelte und beschriebene Governance Framework dient einer reibungsarmen Realisierung der Prozesse, die eine gemeinsame und verteilte Nutzung

[31] Siehe Governance-Framework.pdf

medizinisch relevanter Daten ermöglichen. Es dient als eine unterstützende Struktur im Projekt und beschreibt Zuständigkeiten, Beziehungen und Hierarchien und definiert somit Verantwortlichkeiten innerhalb des Projektes. Durch diese Struktur unterstützt es beim Treffen von Entscheidungen und trägt somit maßgeblich zum Erfolg des Projektes Zukunftslabor Gesundheit bei. Als Vorbild für das Governance Framework fungierten unter anderem die im Rahmen der Medizininformatikinitiative laufenden Projekte HiGHmed, MIRACUM, DIFUTURE und SMITH. Der Fokus liegt bei der Erstellung und der Datenüberführung mithilfe der Modellierung in die Forschungsplattform, demzufolge liegt der Schwerpunkt auf dem Teilprojekt 1, ist aber über die aktuell geförderte Phase hinaus angelegt. Das Governance Framework definiert Schnittstellen zu den anderen Teilprojekten, und ermöglicht es ihnen so sich einzubringen. Durch die Schnittstellen wird der regelmäßige Austausch zwischen den Teilprojekten gewährleistet.

Die Governance-Struktur des Zukunftslabors besteht aus folgenden Komponenten:

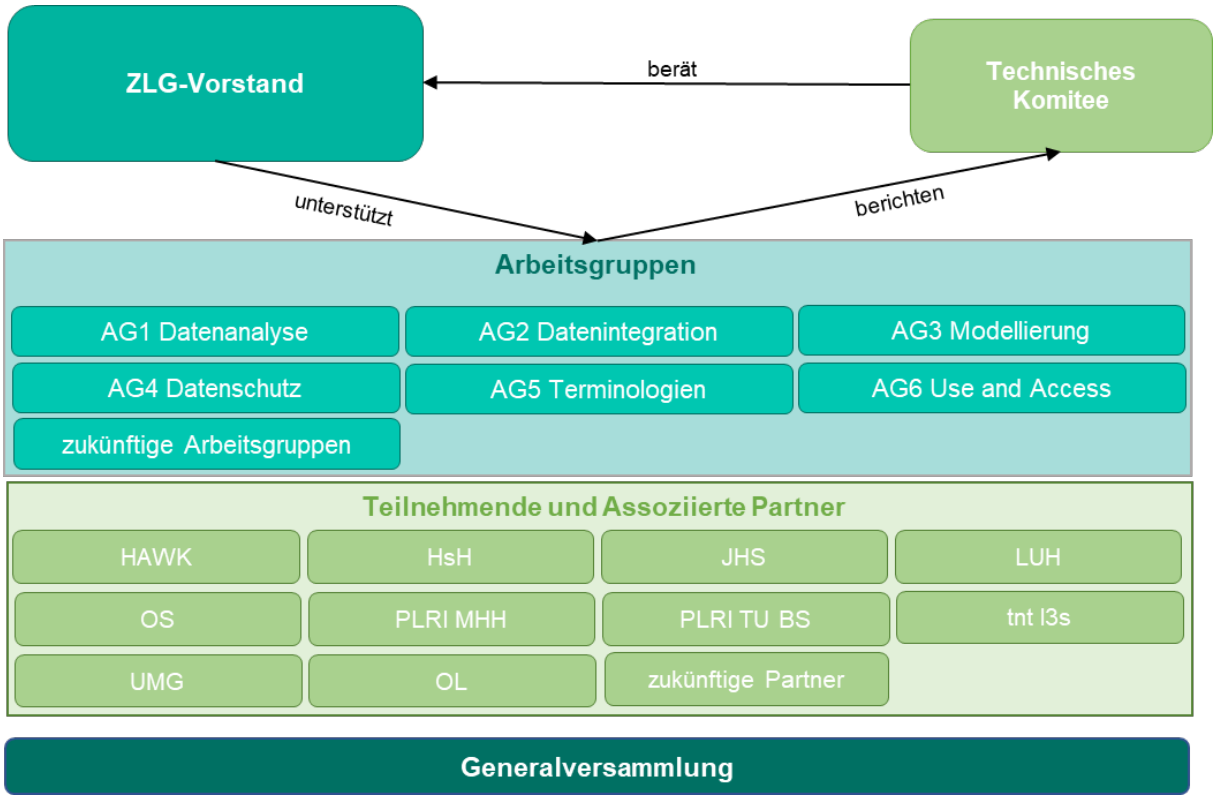

Abbildung 12: Übersicht über die Governance Struktur (aus ZL-G Governance Framework; modifiziert nach Haarbrandt et al. 2018)

Das ZL-G Governance Framework beschreibt die in Abbildung 12 dargestellten Strukturen, den ZLG-Vorstand, das technische Komitee sowie die Aufgabenbereiche der Arbeitsgruppen,

insbesondere die AG3 Modellierung. Darüber hinaus werden Rollendefinition innerhalb der Modellierung vorgenommen und diese voneinander abgegrenzt.

6.3 Pseudonymisierung und Record Linkage

Ein Pseudonymisierungsdienst stellt einen Dienst bzw. Service bereit, welcher für die Pseudonymisierung von (Patienten-) Daten genutzt werden kann. Beim Pseudonymisieren werden bestimmte, identifizierende Datenpunkte in personenbezogenen Daten, welche einen Rückschluss auf die Identität der zu den Daten gehörenden Person zulassen (wie zum Beispiel der Name eines Patienten) maskiert. Die Maskierung geschieht durch das Ersetzen des Namens und anderer Identifikationsmerkmale durch ein anderes Kennzeichen. Ziel ist es, die Identifikation der betroffenen Person unmöglich zu machen oder zumindest erheblich zu erschweren. Im Gegensatz zur Anonymisierung, bei der personenbezogene Daten derart verändert werden, dass persönliche oder sachliche Verhältnisse nur noch mit einem unverhältnismäßigen Aufwand wiederhergestellt werden könnten, ist bei der Pseudonymisierung eine Re-Identifizierung möglich, wenn eine Pseudonym-Liste vorhanden ist. Die Pseudonym-Liste (auch Patientenliste genannt) erlaubt die Zuordnung der identifizierenden Daten (IDAT) und dem dazugehörigen Pseudonym (PSN). Die relevanten Begriffe sind im §46 BDSG definiert.[32]

Ein allgemein anerkanntes Verfahren für die Pseudonymisierung wurde von der TMF in Form generischer Datenschutzkonzepte erarbeitet. Die Grundidee dieses Verfahren ist die Auslagerung von für die Re-Identifizierung notwendigen Wissens an eine vertrauenswürdige dritte Partei (Trusted Third Party, kurz TTP). Das zweistufige Verfahren wird in Abbildung 13 dargestellt. In der Abbildung ist erkennbar, dass die IDAT beim Datenerfasser verbleiben und die medizinischen Daten (MDAT) nur zusammen mit einer patientenspezifischen PID ausgeliefert werden. Die TTP erzeugt zu jeder PID eine PSN (zweite Stufe der Pseudonymisierung), mit welcher die MDAT an z.B. eine Forschungsdatenbank übermittelt werden.

[32] https://dsgvo-gesetz.de/bdsg/46-bdsg/ (abgerufen am 01.02.2021)

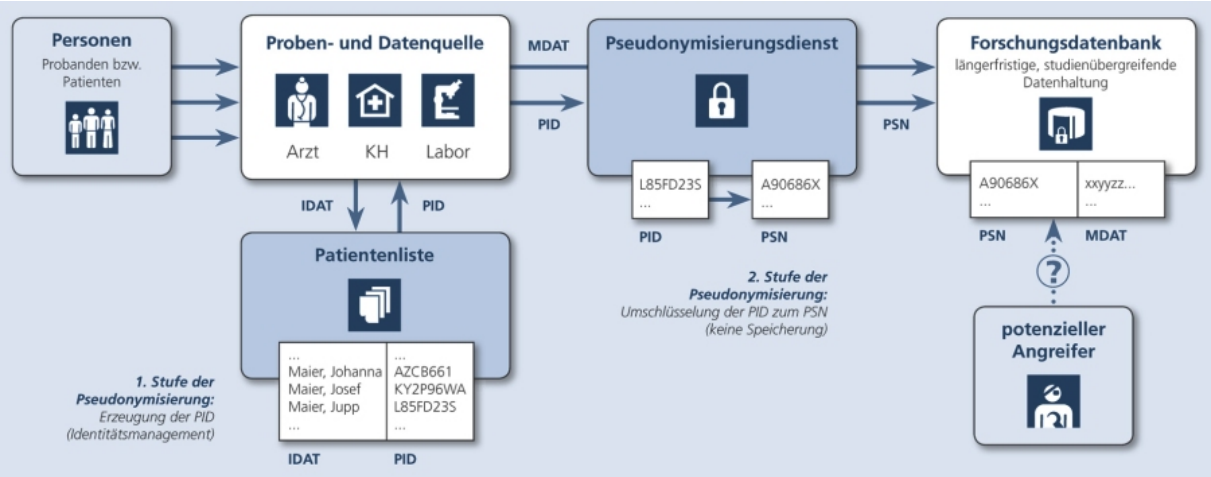

Abbildung 13: Verfahren zur zweistufige Pseudonymisierung nach TMF[33]

Über das Projekt HiGHmed bzw. die MeDICs einzelner HiGHmed-Standorte wird z.B. am Standort Göttingen ein Pseudonymisierungsdienst betrieben. Zum Einsatz kommt hier die „Mainzelliste“ eine Open-Source-Software für ID-Management, welche die Pseudonymisierung nach TMF-Vorgehen technisch umsetzen kann. Die gemeinsame Nutzung des Dienstes ist angedacht und erleichtert nicht nur das Record Linkage, sondern vermeidet zusätzlich redundante Infrastrukturen bzw. reduziert den dazugehörigen Aufwand für Wartung, Pflege und Administration.

[33] https://www.tmf-ev.de/Themen/Projekte/V000_01_PSD.aspx (abgerufen 10.05.2021)

7 Quellenverzeichnis

1. Böttinger E, Putlitz J zu. Die Zukunft der Medizin: Disruptive Innovationen revolutionieren Medizin und Gesundheit. Mit einem Geleitwort von Hasso Plattner. 1. Auflage. MWV Medizinisch Wissenschaftliche Vertragsgesellschaft; 2019. 429 S.

2. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. Npj Digit Med. 20. August 2019;2(1):1–5.

3. Praxishandbuch Forschungsdatenmanagement [Internet]. Praxishandbuch Forschungsdatenmanagement. De Gruyter Saur; 2021 [zitiert 4. Juli 2022]. Verfügbar unter: https://www.degruyter.com/document/doi/10.1515/9783110657807/html?lang=de

4. Haarbrandt B, Tute E, Marschollek M. Automated Population of an i2b2 Clinical Data Warehouse from an openEHR-based Data Repository. J Biomed Inform. 1. August 2016;63.

5. Buettner S, Hobohm HC, Mueller L. Handbuch Forschungsdatenmanagement [Internet]. Bad Honnef: Bock + Herchen Verlag; 2011 [zitiert 13. September 2022]. 224 S. Verfügbar unter: https://opus4.kobv.de/opus4-fhpotsdam/frontdoor/index/index/docId/208