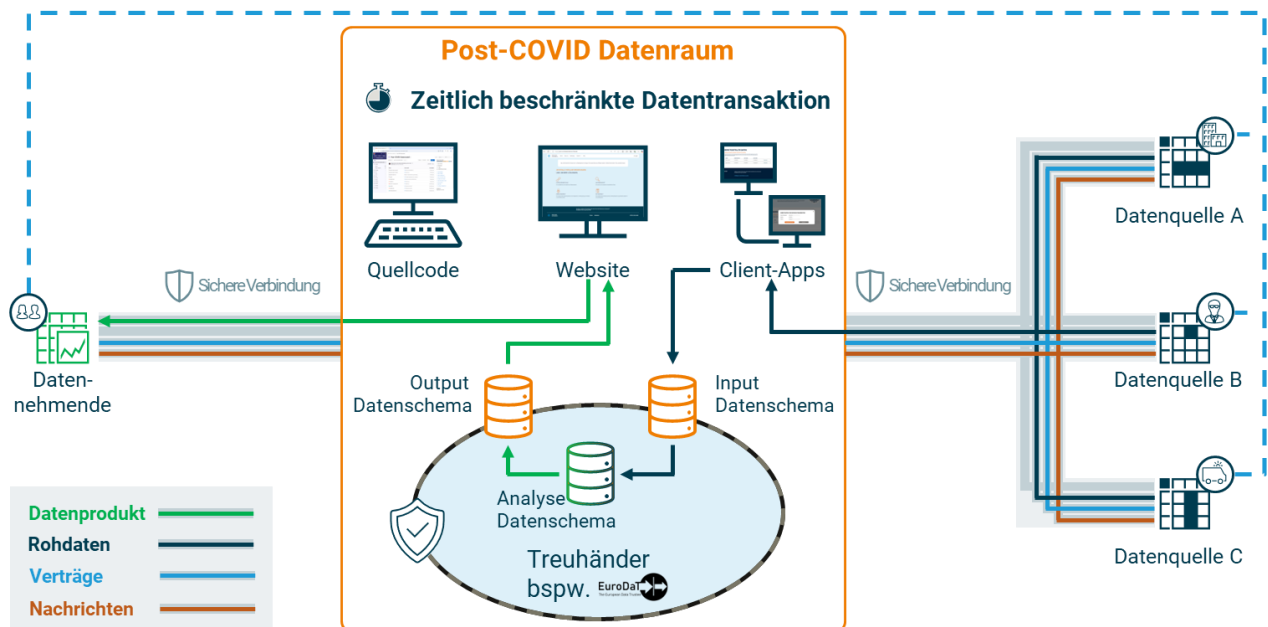


## Challenge „Post-COVID-Datenmodell“



Bericht über die abgeschlossene Stufe 3

04. April 2025

In Zusammenarbeit mit

## MANAGEMENT SUMMARY

Die COVID-19-Pandemie hat die Relevanz belastbarer, sektorenübergreifender Dateninfrastrukturen verdeutlicht. Im Rahmen der Post-COVID Challenge haben wir im Konsortium ein offenes Datenökosystem zur Unterstützung der Post-COVID-Forschung entwickelt – anschlussfähig und übertragbar auf die Entwicklung von Datenmodellen in anderen Sektoren. Dieser Bericht gibt einen Überblick über die wichtigsten Resultate aus Stufe 1 und 2 und fasst die Entwicklungen der Stufe 3 zusammen.

### HAUPTPUNKTE DES BERICHTS

Im Fokus der dritten Stufe stand die vollständige technische und prozedurale Umsetzung des MVP inklusive zentraler Komponenten: Schnittstellen zu Stakeholdern über Client-Apps und die Website, Record Linkage-Algorithmus, Backend und Datenverarbeitungslogik. Wir haben die Anbindung der Datenhaltenden und Intermediäre aktiv vorangetrieben und liefern unsere Software kontinuierlich an die datenhaltenden Stellen aus – ein Nachweis der Vertrauenswürdigkeit und Akzeptanz unserer Lösung. Ein weiterer zentraler Meilenstein unserer Arbeit war die erfolgreiche Durchführung einer Test-Datentranskation mit Record Linkage. Dies beweist, dass unser Ansatz technisch funktioniert. Dabei ist das Ökosystem modular aufgebaut: Bestehende Algorithmen wie E-PIX und Komponenten wie der NFDI4Health-Datenkatalog lassen sich integrieren, weitere Datenquellen und Standards flexibel anschließen.

### MEHRWERTE UND NÄCHSTEN SCHRITTE

Mit unserem MVP demonstrieren wir, wie vorhandene Datenquellen datensouverän und DSGVO-wahrend verknüpft (Record Linkage), besser auffindbar gemacht (Datenkatalog) und gezielt mit Forschungsbedarfen zusammengebracht werden können (Data Science Börse). Auf dieser Basis wollen wir gemeinsam mit unseren Partnern die Erprobung nach der Challenge fortsetzen. Die Bearbeitung unserer konkreten Forschungsfrage nach der Auswirkung von Post-COVID auf die Erwerbsfähigkeit bildet dabei den Ausgangspunkt. Weitere Vorhaben, z. B. mit dem ZEG Berlin, sind in Konzeption. Ziel ist die gezielte Verbreitung und aktive Nutzung des Ökosystems.

### UNSER UNIQUE SELLING POINT (USP)

Der USP unseres Ansatzes liegt in der integrierten Betrachtung technischer, prozeduraler und rechtlicher Aspekte eines dezentralen, datensouveränen Datenökosystems. Nur so lassen sich die komplexen Anforderungen der medizinischen Datenforschung ganzheitlich und praxisnah adressieren. Im Fokus stehen einfacher Zugang für Datenhaltende und -nutzende sowie die Anschlussfähigkeit an bestehende Infrastrukturen. Der erfolgreiche Testbetrieb belegt die technische Reife und hohe Nutzerakzeptanz des Ökosystems und damit die Praxistauglichkeit unseres Ansatzes.

## INHALTSVERZEICHNIS

01.	Einleitung.....	1
02.	Gesamtbeschreibung der geleisteten Arbeiten .....	2
02.01	Beschreibung der geleisteten Arbeiten .....	2
02.02	Entwickeltes geistiges Eigentum und Veröffentlichung.....	5
03.	Konzepte .....	6
03.01	Übersicht über die bisherigen Konzepte.....	6
03.02	Erweiterung des Datenmodells durch Aufnahme neuer Datensätze .....	7
04.	Forschungsobjekt .....	11
04.01	Ausrichtung des Datenmodells an den Anforderungen der Forschung .....	11
04.02	Prozesse zur Einbindung von Stakeholdern.....	14
04.03	Konzept zur Veröffentlichung des Datenmodells .....	16
04.04	Innovationsgrad des Ansatzes.....	21
05.	Datenmodell .....	26
05.01	Vorliegende Datensätze.....	26
05.02	Orientierung an Branchen-Standards .....	32
05.03	Verknüpfung verschiedener Datenquellen .....	33
05.04	Strukturierung des Datenmodells (Typ der Datenbank) .....	45
06.	Prozesse und Architektur .....	47
06.01	Aufbau der IT-Infrastruktur .....	47
06.02	Planung der IT-Infrastruktur .....	49
06.03	Prozesse zur Datenintegration.....	51
06.04	Prozesse zur Datenaktualisierung .....	57
07.	Umsetzbarkeit der Entwicklung sektorunabhängiger Datenmodelle.....	57
07.01	Das Matrixmodell .....	58
07.02	Das Vorgehensmodell.....	60
07.03	Die Qualitätssicherung.....	61
08.	Betrieb und Nachnutzung des Datenmodells .....	63

## 01. EINLEITUNG

Die COVID-19-Pandemie hat in vielerlei Hinsicht offengelegt, wo gesellschaftliche, infrastrukturelle und systemische Strukturen an ihre Grenzen stoßen. Gerade im Gesundheitswesen wurde deutlich, wie stark Entwicklung und gute Entscheidungen von der Verfügbarkeit verlässlicher Daten abhängen. Genau hier setzt unser Vorhaben im Rahmen der Post-COVID Challenge an: Ziel war es, ein **offenes, frei verfügbares und sektorübergreifendes Datenmodell** zu schaffen – zur Unterstützung der Post-COVID-Forschung und darüber hinaus anschlussfähig für weitere Anwendungsbereiche.

Unsere Arbeit ist in ein dynamisches Umfeld eingebettet: Die Pandemie hat nicht nur Schwächen offengelegt, sondern auch Entwicklungen beschleunigt, die zuvor nur zögerlich vorankamen. In Deutschland entstanden datengetriebene Infrastrukturen wie das Netzwerk Universitätsmedizin (NUM) mit dem Projekt NAPKON – heute eine zentrale Ressource für medizinische Forschung. Auch auf europäischer Ebene wächst der Handlungsdruck: Der European Health Data Space (EHDS) schafft neue Rahmenbedingungen für die Nutzung von Gesundheitsdaten. In Formaten der EU-Kommission, wie zuletzt „AI in Healthcare: EU Priorities and Ecosystem Synergies“<sup>1</sup>, zeigt sich: datengetriebene Medizin und KI stehen weit oben auf der politischen Agenda.

Vor diesem Hintergrund war es uns wichtig, nicht isoliert zu agieren, sondern **vorhandene Datenquellen und Strukturen zusammenzuführen und weiterzudenken**. Um diese übergreifend nutzbar zu machen, braucht es eine technisch, rechtlich und prozedural durchdachte Anbindung. Genau hier setzt unser Datenökosystem an.

Im Rahmen der Challenge haben wir ein offenes Datenökosystem entwickelt, dessen Kern ein transaktionsbasierter Treuhänder wie z.B. EuroDaT ist. EuroDaT setzt auf **gesicherte, souveräne Datenflüsse** statt zentrale Datenspeicherung und bildet damit das Rückgrat einer Architektur, die auf Vertrauen, Transparenz und Anschlussfähigkeit ausgelegt ist. Das Ökosystem baut auf bestehenden Vorarbeiten auf, etwa dem Datenkatalog von NFDI4Health oder Komponenten des Record Linkage Algorithmus der Treuhandstelle Greifswald, und integriert diese konsequent. Neben der technischen Anschlussfähigkeit fördern diese etablierten Komponenten auch das Vertrauen der Stakeholder – denn bereits bekannte und bewährte Lösungen stoßen auf deutlich höhere Akzeptanz als vollständig neu entwickelte Strukturen. Gerade im Umgang mit sensiblen Daten ist das ein entscheidender Faktor.

---

<sup>1</sup> European Commission, Webinar: AI in Healthcare: EU Priorities and Ecosystem Synergies. <https://digital-strategy.ec.europa.eu/en/events/ai-healthcare-eu-priorities-and-ecosystem-synergies>

In Stufe 1 haben wir das konzeptionelle Fundament des Datenraums gelegt, in Stufe 2 Stakeholderprozesse ausgestaltet und mit der technischen Umsetzung begonnen. In der sich nun abschließenden Stufe 3 liegt der Fokus auf der Weiterentwicklung zu einem funktionsfähigen MVP: Hierfür haben wir Daten von NAKO, NAPKON, FDZ Gesundheit, Rentenversicherung und Garmin eingesammelt. Bei NAPKON und NAKO konnten wir die Bereitstellung unserer Client App vorantreiben und somit das Kernstück des Datenökosystems mitsamt Backend, Verarbeitungslogik und Privacy Preserving Record Linkage (PPRL) Algorithmus weiterentwickeln. Ein zentraler Meilenstein war die **erfolgreiche Durchführung einer ersten Testtransaktion**, durch die wir die technische Machbarkeit unseres Ansatzes nachweisen konnten. EuroDaT hat sich dabei in zahlreichen Gesprächen mit Stakeholdern als vertrauensstiftende, tragfähige und anschlussfähige, zentrale Komponente des Systems bewährt. Auch mit der THS Greifswald als Intermediär haben wir unsere Zusammenarbeit fortgesetzt.

Begleitend haben wir erste Maßnahmen zur **Verstetigung des Datenökosystems** angestoßen: Durch die Veröffentlichung zentraler Komponenten wie des Codes und der Berichte, Präsentationen und Diskussionen im Rahmen des [DaTNet](#) oder mit Akteuren der [Industrielle Gesundheitswirtschaft](#) am BDI oder [sphin-x](#) haben wir die Sichtbarkeit und Nachnutzbarkeit unseres Ansatzes erhöht.

Dieses Abschlussdokument fasst die Ergebnisse der dritten Stufe und der gesamten Challenge zusammen, dokumentiert die Gesamtentwicklung des Systems und zeigt auf, wie ein übergreifendes, zukunftsfähiges Ökosystem entstehen kann – im Bereich Post-COVID aber auch für alle weiteren Sektoren.

## 02. GESAMTBESCHREIBUNG DER GELEISTETEN ARBEITEN

### 02.01 BESCHREIBUNG DER GELEISTETEN ARBEITEN

Der Hauptfokus der Stufe 3 lag in der praktischen Implementierung des von uns in Stufe 1 konzipierten und in Stufe 2 spezifizierten Datenmodells. Dafür haben wir unsere Stakeholder-Prozesse erprobt, die technische Infrastruktur entwickelt und die in den vorherigen Stufen als am relevantesten identifizierten Datensätze operativ angebunden. Dieses Kapitel beschreibt unsere geleisteten Arbeiten und fasst unsere **Erkenntnisse und Ergebnisse** zusammen, s. Tabelle 1. Für weitergehende Details werden die entsprechenden Stellen im Dokument verlinkt und auf diese verwiesen.

Da medizinisch Forschende für Datenverarbeitung ein hohes Maß an **Vertrauen** in die Sicherheitsmechanismen benötigen, ist es häufig sinnvoll, neue technische Lösungen auf bestehenden Lösungen aufzubauen und diese punktuell dort zu erweitern, wo ein spezifischer Bedarf von der wissenschaftlichen Community klar identifiziert wurde.






	<b>Stakeholder</b>	<ul style="list-style-type: none"> <li>• Festigung des Kontakts zu den Datenhaltenden <a href="#">NAPKON</a>, <a href="#">NAKO</a>, <a href="#">Rentenversicherung</a>, <a href="#">FDZ Gesundheit</a> und <a href="#">Garmin Health</a>, Datennutzenden der Universitätsmedizin Frankfurt und etablierten Intermediären in der datengetriebenen Medizinforschung wie <a href="#">NFDI4Health</a> und <a href="#">THS Greifswald</a></li> <li>• Anforderungsmanagement und Implementierungskooperation mit <a href="#">NAPKON</a>, <a href="#">NAKO</a>, <a href="#">Rentenversicherung</a>, <a href="#">FDZ Gesundheit</a> und <a href="#">Garmin Health</a></li> <li>• Vorbereitung zur Anbindung der Datenquellen an das Ökosystem</li> <li>• Kooperation mit bestehenden Initiativen der Gesundheitsdaten-Ökonomie: <ul style="list-style-type: none"> <li>• Nachnutzung des Metadatenkatalogs der <a href="#">NFDI4Health</a></li> <li>• Vorbereitung PII-Anbindung mit <a href="#">THS Greifswald</a></li> <li>• Vorbereitung eines Workshops zu Treuhänder-Lösungen in der Gesundheitsdaten-Ökonomie in Kooperation mit <a href="#">DaTNet</a></li> </ul> </li> <li>• Bestehende U&amp;A Prozesse erprobt</li> <li>• Für Dritte nachnutzbare Geschäftsmodelle formuliert</li> </ul>
	<b>Datensätze</b>	<ul style="list-style-type: none"> <li>• Datenmanagement der eingeworbenen Testdaten von <a href="#">NAPKON</a>, <a href="#">NAKO</a>, <a href="#">Rentenversicherung</a> und <a href="#">Garmin Health</a> <ul style="list-style-type: none"> <li>• Abruf der Daten von Datengebenden</li> <li>• Abgleich und Mapping der Metadaten</li> <li>• Qualitätssicherung</li> </ul> </li> <li>• Datentreuhandmodell rechtlich und prozedural ausgestaltet</li> <li>• Prozedurale Implementierung der Datenverknüpfung im Datentreuhänder <a href="#">EuroDaT</a></li> <li>• Explorative Erweiterung des Datenangebots um Open Data-Inhalte</li> </ul>
	<b>Datenökosystem</b>	<ul style="list-style-type: none"> <li>• Data Linkage-Algorithmus implementiert</li> <li>• Implementierung PPRL mit Bloomfilter-Codierung in der Client-App und PPRL-Backend in EuroDaT</li> <li>• Aufbau Data Science-Vermittlerbörse</li> <li>• Überarbeitung der <a href="#">Website</a> nach UI/UX-Gesichtspunkten</li> </ul>
	<b>Technik</b>	<ul style="list-style-type: none"> <li>• MVP der Datenraum-App fertiggestellt</li> <li>• Prototypische Transaktion mit Test-Daten durchgeführt</li> <li>• Transaktions-Lebenszyklus durch Anbindung an EuroDaT durchlaufen</li> <li>• Entwicklung einer on-premises Client-App zur Verbindung mit EuroDaT</li> <li>• Implementierung der technischen IT-Architektur inkl. <ul style="list-style-type: none"> <li>• Umsetzung und Auslieferung der Client-App für Datengebende</li> <li>• Aufbau der spezifizierten Zielarchitektur</li> </ul> </li> </ul>
	<b>Dokumentation</b>	<ul style="list-style-type: none"> <li>• Veröffentlichungskanäle etabliert für <ul style="list-style-type: none"> <li>• <a href="#">GitHub Repository</a> mit Code Basis von Backend und Client-App</li> <li>• <a href="#">Web-Frontend für integrierte Informationsbereitstellung und Zugang zum Datenökosystem</a></li> <li>• Datenmodell</li> </ul> </li> <li>• Dokumentierte Ergebnisse in dem vorliegenden Bericht</li> </ul>

Tabelle 1: Zusammenfassende Darstellung der Ergebnisse der Stufe 3.

Entsprechend haben wir während der Stufe 3 durchgängig die Erfahrung gemacht, dass die Kenntnis und **Nachnutzung bestehender Initiativen** in der Post-COVID-Forschung im Speziellen sowie in der medizinischen Datenforschung im Allgemeinen essenziell ist. Dies liegt daran, dass strikte Regulatorik wie die DSGVO allerhöchste Anforderungen an die Verarbeitung besonders schützenswerter Patientendaten stellt.

## STAKEHOLDER

Ein wichtiger Bestandteil der Stufe 3 ist die weitere Integration und Optimierung unserer Kommunikation mit den Stakeholdern, s. Kapitel 04.02. Ein zentraler Erfolg ist hierbei die Bereitstellung einer Client-App bei NAPKON und NAKO. Dies stellt einen wichtigen Schritt dar hin zu einer unmittelbaren Interaktion von Datengebenden und -nehmenden im Datenökosystem. Außerdem demonstriert die NAKO hiermit **Vertrauen in unsere Lösung und Akzeptanz in der Forschungs-Community**. Darüber hinaus verfolgen wir fortlaufend den Fortschritt unserer Daten-Nutzungsanträge für einen allgemeinen Zugang zu den Patientendaten u.a. von NAPKON und NAKO, s. Kapitel 05.01. Anhand einer realen Forschungsfrage erproben wir so den Use & Access Prozess, den Datennehmende durchlaufen, und prüfen unseren Ansatz auf Kompatibilität, s. [Abschlussbericht der Stufe 2](#). Zudem haben wir die Planungen für einen Workshop weitgehend abgeschlossen, der sich mit Treuhänder-Lösungen im Gesundheitssektor auseinandersetzt, um Synergien zu erkennen und zu fördern und den Austausch zwischen Akteuren aus Forschung und Praxis zu intensivieren, s. Kapitel 04.03.

## DATENSÄTZE

In Stufe 3 haben wir die relevantesten Datensätze an unseren Datenraum angebunden. Konkret haben wir Testdaten von einem breiten Spektrum an Stakeholdern vorliegen: von den großangelegten, stark zugangsbeschränkten Forschungsverbünden **NAPKON und NAKO**, Abrechnungsdaten der gesetzlichen Versicherten, Daten der Rentenversicherung sowie persönliche Gesundheitsdaten des Wearable-Herstellers Garmin, s. Kapitel 05.01. Von diesen haben wir prototypisch die größten Datensätzen von NAPKON und NAKO an unseren Datenkatalog angebunden. Um diesen Katalog möglichst nah an den Anforderungen und Erfahrungen der Community auszurichten, haben wir eine systematische Analyse aller existierenden Lösungen durchgeführt und den Datenkatalog des NFDI4Health als optimal für eine Nachnutzung in unserem Datenökosystem empfunden und in unsere Lösung integriert, s. Kapitel 05.01. Darüber hinaus haben wir verschiedene Open Data Datensätze mit aufgenommen und dabei prototypische Erweiterungsoptionen entwickelt, um die Flexibilität und Anpassungsfähigkeit unseres Datenmodells weiter zu erhöhen, s. Kapitel 04.01.

## DATENÖKOSYSTEM

Unsere Arbeiten am Datenökosystem umfassen alle Komponenten, die die angebundenen Datensätze managen und verknüpfen. Diese Komponenten sind essenzielle Voraussetzung, um ein **sektorübergreifendes Datenmodell** in der Post-Covid Forschung produktiv nutzen zu können. Hierfür haben wir in Stufe 3 unsere **Verknüpfungs- und Re-Identifikations-Prozesse** und Algorithmen ausgearbeitet, s. Kapitel 05.03. Wir haben einheitlich das Konzept der Privacy Preserving Record Linkage



(PPRL) mittels Bloomfiltern angewandt, eine hash-basierte Technologie, die eine **datenschutzfreundliche, sichere Verknüpfung** von Datensätzen ermöglicht, indem sie Identifikatoren anonymisiert und dabei die Vollständigkeit und Genauigkeit der Datenverknüpfung aufrechterhält, s. Kapitel 05.03 und 06.03.

## TECHNIK

Wie oben beschrieben lag der Hauptfokus unserer Arbeiten in Stufe 3 auf der technischen Implementierung des Datenökosystems. Ein zentraler Meilenstein war hierbei die Ende-zu-Ende Umsetzung der von uns konzipierten Verknüpfungs-Logik verschiedener Datensätze. Diese Umsetzung konnten wir mit der erfolgreichen Durchführung einer ersten, vollständigen EuroDaT-Transaktion nachweisen, s. Kapitel 05.03. Diese Entwicklung ermöglicht es uns, den gesamten Ablauf der Datenverknüpfung inklusive Record Linkage, s. Kapitel 04.04, zu automatisieren. Um den Nutzenden einen einfachen Zugang zum Datenökosystem anzubieten, haben wir eine on-premises Client-App entwickelt, die die Verbindung zum Datentreuhänder EuroDaT und somit dem Ökosystem herstellt. Die zugrundeliegende technische IT-Architektur haben wir hierbei entsprechend der in [Stufe 2](#) vorgestellten Spezifikationen implementiert und alle wesentlichen Komponenten entwickelt. Die so realisierte Zielarchitektur erfüllt unsere langfristigen technologischen Anforderungen und bildet die Grundlage für zukünftige Erweiterungen des Datenökosystems, s. Kapitel 06.02.

## DOKUMENTATION

Wie im [Abschlussbericht der Stufe 2](#) dargestellt, nutzen wir Veröffentlichungskanäle des Datenökosystems gleichzeitig als strukturierte Dokumentation unserer Arbeit. Die in Stufe 3 geleisteten Weiterentwicklungen unseres Webfrontends, des GitHub Repositories und unseres Workshop-Programms stellen wir in Kapitel 04.03 vor.

## 02.02 ENTWICKELTES GEISTIGES EIGENTUM UND VERÖFFENTLICHUNG

Die im bisherigen Projektverlauf und insbesondere in dieser Stufe entwickelten Ergebnisse haben wir in Kapitel 02.01 zusammengefasst. An diesen Ergebnissen haben wir **keine geistigen Eigentumsrechte angemeldet**, um die Nachnutzbarkeit des Ansatzes zu wahren und im Interesse maximaler Transparenz eine frühzeitige Veröffentlichung unserer Ergebnisse schon in der Entwicklungsphase zu ermöglichen. Ein solch transparenter Ansatz wäre nicht möglich, wenn wir am Ende der Projektarbeit ein vollständig entwickeltes Produkt als Schutzmarke anmelden würden.

Für eine detaillierte Zusammenfassung der entwickelten Veröffentlichungskanäle verweisen wir auf Kapitel 04.03.



## 03. KONZEPTE

### 03.01 ÜBERSICHT ÜBER DIE BISHERIGEN KONZEPTE

Die Erarbeitung eines neuen Datenmodells stellt vielfältige Anforderungen und benötigt die Betrachtung verschiedenster Aspekte. Diese reichen von großen, grundlegenden Aufgaben wie der Ausarbeitung eines generellen Zielbilds und eines Betriebsmodells bis zu Detailfragen, wie nötigen Personalschulungen oder IT-Infrastruktur-Architekturen. Über die drei Stufen der Post-Covid-Challenge hinweg haben wir diese und viele weitere Themen ausführlich betrachtet und die Ergebnisse in verschiedenen **Prozessen und Schaubildern** festgehalten. Die so entstandenen Konzepte machen die geleisteten Entwicklungsarbeiten nachnutzbar, dokumentieren die Umsetzbarkeit und ermöglichen eine Übertragung auf weitere Anwendungsfälle.

An dieser Stelle stellen wir eine Übersicht über die im Rahmen der Challenge erstellen Konzepte bereit. In Tabelle 2 sind alle allgemeinen Konzepte aufgelistet.

Konzept	Zu finden in
Konzeptionelle Überlegungen zur Entwicklung eines offenen und nachnutzbaren Datenmodells	<a href="#">Bericht 1</a> , Kapitel 01.01
Konzeptionelle Überlegungen zum Zielbild sowie Detaillierung des MVP	<a href="#">Bericht 1</a> , Kapitel 02.03
Identifizierte Geschäftsprozesse in einem gesamtheitlichen Zielbild	<a href="#">Bericht 1</a> , Kapitel 08.01
Data Governance Konzept	<a href="#">Bericht 2</a> , Kapitel 03.01
Konzept der Datenmodellpflege	<a href="#">Bericht 2</a> , Kapitel 03.02
Schulungskonzept	<a href="#">Bericht 2</a> , Kapitel 03.03
Grundlegende Fachprozesse	<a href="#">Bericht 2</a> , Kapitel 03.04
Erweiterung des Datenmodells / Aufnahme neuer Datensätze	Dieser Bericht, Kapitel 03.02

Tabelle 2: Übersicht der im Rahmen der Challenge erarbeiteten Konzepte.

Zusätzlich zu diesen Konzepten haben wir vier weitere Themengebiete berichtsübergreifend weiterentwickelt. In Tabelle 3 findet sich eine Übersicht zu diesen Themen und wo sie in den unterschiedlichen Berichten zu finden sind.

Konzept	Zu finden in
Prozesse zur Einbindung von Stakeholdern	<a href="#">Bericht 1</a> , Kapitel 03.02 <a href="#">Bericht 2</a> , Kapitel 04.02 Dieser Bericht, Kapitel 04.02
Konzept zur Veröffentlichung des Datenmodells	<a href="#">Bericht 1</a> , Kapitel 03.03

	<a href="#">Bericht 2</a> , Kapitel 04.03 Dieser Bericht, Kapitel 04.03
Prozesse zur Datenintegration	<a href="#">Bericht 1</a> , Kapitel 05.02 <a href="#">Bericht 2</a> , Kapitel 06.02 Dieser Bericht, Kapitel 06.03
Prozesse zur Datenaktualisierung	<a href="#">Bericht 1</a> , Kapitel 05.03 <a href="#">Bericht 2</a> , Kapitel 06.03 Dieser Bericht, Kapitel 06.04

Tabelle 3: Übersicht der im Rahmen der Challenge erarbeiteten und stetig weiterentwickelten Konzepte.

Um eine bestmögliche **Nachnutzbarkeit** der Konzepte zu ermöglichen, haben wir sie als Bestandteil der Berichte der [Stufe 1](#) und [Stufe 2](#) auf unserer [Website](#) veröffentlicht. Nach Abgabe des hier vorliegenden Abschlussberichts zur Stufe 3 werden wir auch diesen zeitnah auf der Website veröffentlichen, so dass die breite Öffentlichkeit freien Zugriff auf die erarbeiteten Informationen hat.

### 03.02 ERWEITERUNG DES DATENMODELLS DURCH AUFNAHME NEUER DATENSÄTZE

Unser Konzept zur Aktualisierung bereits integrierter Daten haben wir ausführlich im Kapitel 06.04 des [Abschlussbericht der Stufe 2](#) erläutert und verweisen daher an dieser Stelle auf den bereits vorliegenden Bericht sowie auf das Kapitel 06.04 des aktuellen Berichts für eine kompakte Zusammenfassung der aktuellen Weiterentwicklungen. Zusätzlich verweisen wir auf Kapitel 05.01 für eine Übersicht über die bereits angebundenen, bzw. sich im Anbindungsprozess befindenden Datensätze.

Neben der Aktualisierung bereits angebundener Datensätze, muss das Datenökosystem außerdem noch in der Lage sein, **nahtlos um neue Datensätzen erweitert** zu werden. Denn nur mit einer niederschweligen Möglichkeit der Erweiterung, kann das Datenmodell stets die relevantesten Datensätze bereitstellen und so **nachhaltig zukunftsfähig** bleiben. Gleichzeitig kann sich das Datenmodell nur mit einer dynamischen Erweiterbarkeit an die sich ebenso dynamisch ändernden Anforderungen der Nutzenden anpassen und sich nachhaltig in der Forschungslandschaft positionieren.

Wir haben daher ein **Konzept für die Integration neuer Datensätze** in das Datenmodell entwickelt. Dieses Konzept unterscheidet drei grundsätzliche Dimensionen, die notwendig sind, um neue Datensätze anzubinden: Die prozedurale Integration der Datengebenden, die rechtliche Prüfung der Zulässigkeit der Datenverknüpfung, und technische Integration als Umsetzung der Datenanbindung. Im Folgenden stellen wir

die **drei Integrationsdimensionen** kurz vor und gehen anschließend auf unseren konkreten Integrationsprozess für neue Datensätze ein.

## DIE DREI DIMENSIONEN DER INTEGRATION NEUER DATENSÄTZE

### Prozedurale Integration – Einbindung in die Abläufe des Datenökosystems

Unter prozeduraler Integration verstehen wir die systematische Einbindung datengebender Stellen in die **operativen Abläufe des Datenökosystems**, um eine nahtlose Verbindung zwischen den Prozessen der Datenquellen und des Datenökosystems zu gewährleisten. Nur wenn die Prozesse einfach vereinbar sind, ist sichergestellt, dass die operative Zusammenarbeit zwischen datengebenden Stellen und dem Datenökosystem reibungslos funktioniert, und keine systemischen Probleme auftreten wie z. B. uneinheitliche Datenaktualisierungen, ungenaue Datenverknüpfungen oder ein ungesicherter Datenaustausch.

Um eine prozedurale Integration zu erreichen, binden wir die datengebenden Stellen als Stakeholder in unser Datenökosystem an und **strukturieren die technische Datenanbindung einheitlich**, s. auch Kapitel 04.02 und 06.03. Dabei vergleichen wir unter anderem bestehende Prozesse der Datenaufnahme, -integration und -aktualisierung bei den Datengebenden mit den im Ökosystem etablierten Prozessen. Wir haben diesen Ansatz so gestaltet, dass sowohl datengebende als auch datennutzende Parteien aktiv eingebunden werden und alle relevanten Standards und Regularien eingehalten werden. Eine zentrale Rolle spielt dabei die prozedurale Abfolge der Datenintegration und -aktualisierung, die durch spezifische Ablaufdiagramme visualisiert wird, s. auch Kapitel 03.01. Zudem stellen wir sicher, dass die Integration neuer Datensätze sowohl konzeptionell als auch auf der Ebene der Primärdaten erfolgt, um eine umfassende und effiziente Verarbeitung und Nutzung der Daten zu gewährleisten.

### Rechtliche Prüfung – Datenökonomie von Anfang an mitgedacht

Ein besonderes Alleinstellungsmerkmal unseres Ansatzes ist es, dass wir rechtliche Überlegungen der Datenökonomie von Anfang an mitgedacht haben. Im Zuge dieser Überlegungen hat sich herausgestellt, dass bei Aufnahme neuer Datensätze auch die **rechtliche Zulässigkeit geprüft** werden muss. Prominente Beispiele sind DSGVO-Pflichten und das im Sozialgesetzbuch formulierte Verbot Krankenkassendaten mit Daten von Versorgungsträgern wie der Rentenversicherung zu verknüpfen. Ohne eine integrierte rechtliche Betrachtung neuer Datensätze kann man also leicht gegen verschiedene Regulatorik verstoßen und so das Gebot der Rechtssicherheit für die Datengebenden und -nutzenden verletzen.

Für eine solche Prüfung müssen einerseits die Daten selbst betrachtet werden, andererseits können Ausnahmetatbestände wie z. B. die Erwägungsgründe der DSGVO mit hinzugezogen werden. Ziel ist es, bei Aufnahme der Daten eine rechtliche Einschätzung für die spätere Nutzung der Daten vorliegen zu haben, sei es durch Datengebende oder den Datenraum. Dadurch wird sichergestellt, dass es nicht zu missbräuchlichen Datennutzungen kommt.

#### Datensouveränität verbleibt bei datenhaltenden Stellen

An dieser Stelle lohnt es sich noch einmal hervorzuheben, dass die **Datensouveränität zu jeder Zeit bei den datenhaltenden Stellen bleibt**. Dafür bleiben zunächst auch die Prozesse zur Datenbeantragung seitens der Datenhaltenden so bestehen wie sie aktuell existieren. Im späteren Zielbild ist hier vorgesehen, für ein vereinheitlichtes Use & Access-Verfahren entwickelte Lösungen auf unserer Website mit zu integrieren. Hierfür planen wir, die Entwicklungen von NFDI4Health und der Datenzugangs- und Koordinierungsstelle, die eine zentrale Datenbeantragung betreffen, genau zu beobachten. Je nach technischen Möglichkeiten werden die dort entstandenen Lösungen in den Datenraum integriert. Benötigte Verträge werden ebenfalls vor Aufnahme neuer Daten geschlossen. Hierzu gehören beispielsweise die zu akzeptierenden AGBs.

#### Technische Integration – Automatisierte Datentransaktionen ermöglichen

Im Datenraum sollen Datentransaktionen automatisiert durchgeführt werden können. Hierzu sind technisch viele Schritte nötig, unter anderem der automatisierte Abruf von Daten von den datenhaltenden Stellen. Diese müssen eine Client-App mit funktionaler API-Anbindung vorliegen haben. Details hierzu sind in den technischen Abschnitten 06.01 und 06.02 vorzufinden.

Die daraus folgend notwendigen Schritte für die technische Anbindung neuer Datensätze werden ausführlich in Kapitel 06.03 beschrieben. Hierzu gehören beispielsweise die technischen Details der Aufnahme von Daten in den Datenkatalog oder auch die **Installation einer Client-App** bei den datenhaltenden Stellen. Auch die Bereitstellung der PII oder die Zusammenarbeit mit einer Treuhandstelle muss entsprechend umgesetzt werden.

#### PROZESS ZUR INTEGRATION NEUER DATENSÄTZE

Nachdem die drei Dimensionen der Integration beleuchtet wurden, gilt es das Vorgehen zur Erweiterung des Datenmodells genauer zu betrachten. Wer kann das Datenmodell erweitern? Wie werden neue Datensätze aufgenommen? Wer sorgt für eine stetige Weiterentwicklung?

Grundsätzlich unterscheiden wir in unserem Konzept die Datensatzintegration nach dem Reifegrad des Ökosystems und berücksichtigen die jeweiligen Anforderungen an die drei Dimensionen. So wird die Integration neuer Datensätze in der Frühphase des Datenökosystem-Betriebs nur durch die aktive Ansprache und Überzeugung der Stakeholder vorangetrieben, wofür wir den erarbeiteten Stakeholder-Management-Prozess nutzen, s. Kapitel 04.02. Mit wachsender Bekanntheit und Verbreitung des Ökosystems werden sich **Netzwerk-Effekte** herausbilden, durch die es für Nutzende Person vorteilhaft wird, sich selbständig an das Ökosystem anzubinden.

Zu unterscheiden ist die Aufnahme neuer Datensätze von datenhaltenden Stellen, welche bereits Teil des Datenraumes sind, und von datenhaltenden Stellen, welche noch nicht angebunden sind. Bei Ersteren sind die nötigen Prozesse kürzer, da grundlegende Informationen und rechtliche Aspekte wie die Unterzeichnung der Datenraum-AGBs bereits hinterlegt sind. Bei **Aufnahme neuer Datensätze** aus solchen Quellen kann ein reduziertes Onboarding erfolgen, bei dem beispielsweise lediglich Informationen zum spezifischen Use & Access Verfahren des neuen Datensatzes abgefragt und hinterlegt werden. Bei der Aufnahme einer neuen datenhaltenden Stelle müssen hingegen alle drei Integrationsdimensionen durchlaufen werden.

DATENSATZ REGISTRIEREN

Registrieren Sie Ihre Daten unkompliziert und ermöglichen Sie deren Nutzung durch Forscher und Forscherinnen.

Name des Datensatz:

Kurzbeschreibung:

Personenbezogene Daten: ☒ Ja ☐ Nein

Use & Access Verfahren: ☒ Ja ☐ Nein

Metadaten:  Datei ins Feld ziehen und ablegen oder **DURCHSUCHEN**

DATEN ZUR ÜBERPRÜFUNG SENDEN ABBRECHEN

Abbildung 1: Beispiel Eingabemaske zur Registrierung von Datensätzen.

Im Zielbild wird es eine Eingabemaske auf der Website geben, mittels welcher sowohl neue als auch bereits angebundene Datengeber Daten zur Aufnahme bereitstellen können. Ein solches Verfahren senkt die Hürden für eine Teilnahm am Datenraum, wodurch ein rascheres Wachstum des Datenökosystems zu erwarten ist.

Abschließend möchten wir betonen, dass die Erweiterung um neue Datensätze grundlegend für den Erfolg unseres Datenmodells ist. Durch die sorgfältige Betrachtung der prozeduralen, rechtlichen und technischen Dimensionen der Integration schaffen wir eine flexible und sichere Grundlage für die kontinuierliche Aufnahme von Daten. Unser integrativer Ansatz stellt sicher, dass datengegebende Stellen stets die Datensouveränität behalten und gleichzeitig die rechtlichen Vorgaben eingehalten werden. Mit der Möglichkeit der Datenregistrierung direkt auf unserer Website senken wir zukünftig die Hürden für die Einbindung neuer Datenquellen. Dies trägt maßgeblich dazu bei, die Reichweite und den Nutzen des Datenökosystems langfristig zu steigern.

## 04. FORSCHUNGSOBJEKT

### 04.01 AUSRICHTUNG DES DATENMODELLS AN DEN ANFORDERUNGEN DER FORSCHUNG

Ein Datenmodell kann nur dann erfolgreich sein, wenn es seinen Nutzenden einen echten Mehrwert liefert. Unter dieser Prämisse haben wir in den vergangenen drei Stufen gearbeitet und unser **Datenmodell aktiv nach den Anforderungen der Forschung ausgerichtet**. Hierzu lag in Stufe 1 der Fokus zunächst auf der Identifikation vorliegender Hürden und Wünsche seitens verscheidender Stakeholder. In der darauffolgenden Stufe 2 haben wir uns vermehrt mit den daraus resultierenden Anforderungen an den Datenraum beschäftigt. **Datenschutz, Datensouveränität, Flexibilität, die Vermeidung von Doppelstrukturen** oder auch eine **einfache Handhabung** waren hierbei die Hauptpunkte. Im Zuge der letzten Stufe der Challenge haben wir uns nun einerseits um die aktive Umsetzung und Integration der zuvor identifizierten Punkte gekümmert, andererseits haben wir gezielt versucht den Mehrwert für Forschende zu erweitern. Eine Zusammenfassung des allgemeinen Innovationsgrad unserer Arbeit ist in Kapitel 04.04 zu finden.

Im Folgenden beschreiben wir die zentralen, von uns ausgearbeiteten Elemente des Datenraums zur Ausrichtung an den Anforderungen der Forschung: den Datenkatalog, die Vermittlerbörse, Transaktionen in EuroDaT sowie unsere Anbindung von Open Data.

Die zentralen Werkzeuge des Datenökosystems



Abbildung 2: Unser Mehrwertversprechen für Forschende.

## DER RESEARCH HUB – DATEN UND DATA SCIENCE VERKNÜPFEN

Ein Problem, auf das wir mehrmals gestoßen sind, ist, dass, selbst wenn Daten vorliegen, den Forschenden oftmals die Möglichkeiten für komplexe Auswertungen fehlen. Dies kann einerseits an Zeitmangel, andererseits an fehlenden Ressourcen oder Wissen liegen. Nicht immer ist die Zusammenarbeit beispielsweise mit Medizin-Informatikern möglich. Um künftig **Daten und Data Science Methoden niederschwelliger miteinander verbinden** zu können, haben wir unsere Website um einen [ResearchHub](#) erweitert. Dieser besteht aus zwei zentralen Bestandteilen, dem Data Science Marktplatz und dem Daten Marktplatz.

Der [Data Science Marktplatz](#) stellt einen Ort da, an dem beispielsweise Start-ups oder Unternehmen, welche Data Science Lösungen anbieten, ihr Angebot bereitstellen können. Somit sehen Forschende schnell und einfach ein Angebot an verschiedenen Lösungen und auch Anbietern, die ihnen bei der Beantwortung ihrer Forschungsfragen behilflich sein können. Es besteht ein Mehrwert für Forschende, da sie niederschwellig Unterstützung für ihren Auswertungen finden und ebenso für die Data Science Anbieter, da ihre Angebote mehr Nutzung finden.

Der [Daten Marktplatz](#) bietet Datenhaltenden die Möglichkeit, bekannt zu machen, dass sie Daten vorliegen haben. Hierbei geht es explizit nicht um Daten, die bereits niederschwellig in anderen Daten- und Metadatenkatalogen zu finden sind, sondern beispielsweise um neu entstehende, ggf. noch nicht ausgewertete Datenschätze. Wieso sollten Datenhaltende auf diese Daten aufmerksam machen? Oftmals haben die Datenhaltenden bereits Ideen, wie ihre Daten ausgewertet werden könnten oder was sie gerne erforschen würden, es fehlen ihnen aber die technischen Möglichkeiten. Durch das Platzieren der Daten auf dem Daten Marktplatz kann auf die Daten aufmerksam gemacht werden und **aktiv nach potenziellen Kooperationspartnern gesucht werden**. So haben einerseits Data Science Anbieter die Chance auf eine Zusammenarbeit und den Erhalt spannender Daten, andererseits bekommen Datenhalter die gewünschte Unterstützung bei der Auswertung ihrer Daten. Erneut eine klare **Win-Win Situation** für alle beteiligten Seiten. Es gilt an dieser Stelle anzumerken, dass auf dem Daten Marktplatz selbstverständlich nur über die Existenz der Daten informiert wird, keineswegs sollen die Daten selbst direkt bereitgestellt werden.

Aktuell werden auf der Website verschiedene Beispiele gezeigt, wobei an der Aufnahme beliebiger Data Science Lösungen und Datensätze sowie der Kontaktaufnahme mit möglichen ersten Lösungs- und Datenanbietern bereits gearbeitet wird.



## DER DATENKATALOG – DATEN EINFACH ZUGÄNGIG MACHEN

Ein zentraler Bestandteil unseres Datenmodells ist der Datenkatalog. Dieser wurde in den vorausgehenden Berichten ausführlich beschrieben und im Rahmen dieser Stufe in einer ersten Entwicklungsstufe umgesetzt. Der Datenkatalog verbindet dabei mehrere Prinzipien unseres Datenraums. Er ermöglicht einerseits einen **einfachen Zugriff auf und eine Sichtbarkeit von Daten**. Andererseits setzen wir mit der von uns implementierten direkten Anbindung an den **Health Study Hub des NFDI4Health** konsequent auf bereits bestehende Strukturen und Vertrauensbasen. Mehr Details hierzu sind im Kapitel 05.01 zu finden. Zusätzlich ist es insbesondere im Bereich der medizinischen Forschung von großem Vorteil auf bereits bekannte Lösungen zu setzen. Neue Lösungen, die noch kein Vertrauen in der Community haben, werden oftmals nur schlecht angenommen und benötigen dadurch besonders viel Überzeugungsarbeit, um sich am Markt platzieren zu können. Durch unsere ausführlichen Recherchen zu bereits bekannten Lösungen können wir hier **bestehendes Vertrauen nutzen** und unseren Datenraum dadurch vorantreiben.

## AUSWERTUNGEN IN EURODAT – DATENANALYSE FLEXIBEL GESTALTEN

Ein weiterer wichtiger Punkt bei der Ausrichtung des Datenmodells an den Anforderungen der Forschung ist eine **flexible Nutzbarkeit**. Hierzu zählt die Möglichkeit, verschiedenste Auswertungen in EuroDaT, beziehungsweise dem genutzten Datentreuhänder, durchzuführen. Der Wunsch nach sogenannten Trusted-Research-Environments (TRE) oder Secure-Processing-Environments (SPE) wächst sowohl in der nationalen als auch internationalen Forschungsgemeinde. Mit unserem Prinzip der sicheren Datenauswertung innerhalb eines Datentreuhänders liegen wir hier voll im Trend und können eine gute, bereits funktionsfähige Lösung bieten. Im Rahmen der Entwicklung haben wir deshalb die unterschiedlichen Konzepte so erweitert, dass perspektivisch verschiedenste Auswertungen in EuroDaT möglich sind. Kern dieses Ansatzes bleibt die Verknüpfbarkeit von Daten auf Patientenebene, das sogenannte Record Linkage, mittels PPRL-Methoden im Treuhänder, eins unserer Wertversprechen s. Kapitel 04.04.

## OPEN DATA – INNOVATIONEN FÖRDERN

Open Data (frei zugängliche Daten, welche mittels offener und diskriminierungsfreier Lizenzen frei weiterverwendet werden dürfen) wird eine allgemeine Rolle in der Forschung und Datenanalyse zugeschrieben. Durch den Austausch mit verschiedenen Stakeholdern konnten wir unseren Fokus ebenfalls auf diesen Datenbereich erweitern und Open Data in unser Datenmodell einbinden. Durch die **Bereitstellung transparenter, zugänglicher und nachnutzbarer Daten** werden Kollaboration und Innovationen einfacher möglich gemacht. Dies kann nicht nur die Effizienz der Forschung steigern,

sondern auch **neue Perspektiven** bieten, die die Entwicklungen und Implementierung von Lösungen beschleunigen. Daher trägt die Integration von Open Data dazu bei, das volle Potenzial des Datenmodells in Zukunft auszuschöpfen. Hierfür wurde von uns zunächst eine technische Machbarkeitsstudie durchgeführt. Das Zielbild der Open Data Integration wird in Kapitel 05.01 ausführlich beschrieben.

## 04.02 PROZESSE ZUR EINBINDUNG VON STAKEHOLDERN

Seit Beginn der Post-COVID Challenge wurden die Grundlagen für die Einbindung von Stakeholdern geschaffen und erprobt. Die etablierte Einteilung in Datenhaltende, Datennutzende und Intermediäre sowie Formate wie Einzelgespräche und Fokusmeetings bildeten auch in dieser Phase das Fundament unserer Arbeit.

Die **kontinuierliche Einbindung relevanter Stakeholder** war auch in Stufe 3 zentral. Ziel war es, das entstehende Datenökosystem weiterhin eng an den Anforderungen und Perspektiven der Beteiligten auszurichten und gleichzeitig Vertrauen in den Ansatz aufzubauen. Gerade in der Umsetzungsphase entstehen häufig Rückfragen oder Abstimmungsbedarfe – hier erwiesen sich Einzelgespräche und Fokusmeetings als besonders wertvoll: Wir konnten frühzeitig Impulse aufnehmen, potenzielle Stolpersteine identifizieren und zusätzliche Ansprechpersonen auf Stakeholderseite einbeziehen. So bleibt die Umsetzung praxisnah, realistisch und im Sinne aller Beteiligten.

Zentrale Fortschritte in der Stakeholder-Einbindung sind im Folgenden dargestellt.

### DATENHALTENDE – ENGE ZUSAMMENARBEIT UND ERPROBUNG DER PROZESSE

Mit **NAKO** und **NAPKON** wurden zentrale datenhaltende Partner enger angebunden. In Gesprächen wurden Prozesse zur technischen Anbindung (insb. zur Installation der Client-App, einer Schnittstelle zu den Datengebern, s. Absatz Prozesse für Datengeber in Kapitel 06.03), zur Datenfreigabe sowie zur vertraglichen Zusammenarbeit weiter konkretisiert. Die enge Abstimmung war entscheidend, um die Datensouveränität und bestehende technische als auch Governance-Strukturen zu wahren.

### INTERMEDIÄRE UND INFRASTRUKTURELLE INSTITUTIONEN – POTENZIALE NUTZEN

Die Zusammenarbeit mit der **Treuhandstelle Greifswald** und der **NFDI4Health** wurde gezielt intensiviert. Die Treuhandstelle Greifswald spielt eine zentrale Rolle in der Verwaltung personenbezogener Informationen im universitär-medizinischen Forschungsumfeld in Deutschland und bringt in diesem Bereich tiefgreifendes Know-how ein. Ihre technische und organisatorische Anbindung ist essenziell, um

datenschutzkonformes Privacy Preserving Record Linkage (PPRL) überhaupt zu ermöglichen. Darüber hinaus konnte durch die enge Zusammenarbeit ein **wichtiger Baustein des MVP realisiert** werden: Bestandteile des dort etablierten **PPRL-Algorithmus konnten übernommen**, angepasst und in unser System integriert werden (s. Kapitel 05.03).

Die **NFDI** treibt mit einer Vielzahl an Initiativen die Verbesserung der Forschungsdateninfrastruktur in Deutschland aktiv voran. Besonders hervorzuheben ist der von NFDI4Health entwickelte Metadatenkatalog, den wir im Rahmen unseres MVP erfolgreich nachnutzen und integrieren konnten (s. Absatz NFDI4Health Study Hub im Kapitel 05.01 und Kapitel 06.03).

Beide Beispiele zeigen, welches Potenzial in der bereits vorhandenen fachlichen und technischen Expertise in Deutschland steckt – und wie wichtig es ist, dieses Wissen gemeinsam weiterzudenken und durch gezielte Zusammenarbeit in die Praxis zu bringen.

#### DATENNUTZENDE – PERSPEKTIVEN EINBRINGEN

Die Anforderungen und Perspektiven der Datennutzenden wurden durch die Säule Medizin unseres Konsortiums eingebracht. In dieser Stufe haben wir auch gezielt Kontakt zu Akteuren mit sozialwissenschaftlicher Ausrichtung aufgenommen, wie z.B. dem Institut [„Sozialwissenschaftliche Perspektiven von Sport, Bewegung und Gesundheitsförderung“ der TU Chemnitz](#), das eine [einschlägige Expertise](#) in dem Bereich Post-COVID in Zusammenhang mit der Erwerbsfähigkeit aufweist. Die **sozialwissenschaftliche Perspektive** verschiedener Akteure ist besonders relevant im Hinblick auf die geplante Erprobung des Datenraums entlang unserer konkreten Forschungsfragestellung „Wie wirkt sich Post-COVID auf die Erwerbsfähigkeit aus?“. Potenziell relevant wird in diesem Zusammenhang auch die Anbindung weiterer datenhaltender Stellen aus den Forschungsnetzwerken dieser Akteure.

In Stufe 3 haben wir neben den institutionellen Stakeholdern auch die Öffentlichkeit einbezogen, um die **Ergebnisse transparent bereitzustellen** und breiter sichtbar zu machen. Details hierzu folgen im nächsten Kapitel 04.03.

Zusammenfassend zeigt dieses Kapitel, wie die in Stufe 3 intensivierte Zusammenarbeit mit datenhaltenden, -nutzenden und vermittelnden Akteuren dazu beigetragen hat, unser **Datenökosystem praxisnah und anschlussfähig weiterzuentwickeln**. Die Erfahrungen aus dieser Arbeit zeigen, dass tragfähige Lösungen nur im Dialog entstehen – technisch, prozedural und mit Blick auf Akzeptanz und Weiterverwendung.

### 04.03 KONZEPT ZUR VERÖFFENTLICHUNG DES DATENMODELLS

In diesem Kapitel fassen wir unser Veröffentlichungskonzept für das in der Challenge entwickelte Post-COVID-Datenmodell zusammen. Hierbei weisen wir darauf hin, dass wir die diesem Konzept zugrunde liegende konzeptionelle Arbeit bereits detailliert in den Abschlussberichten zu [Stufe 1](#) und [Stufe 2](#) dargelegt haben. Die aktuell abgeschlossene Stufe 3 war darauf ausgelegt, die in früheren Phasen erarbeiteten Konzepte effektiv umzusetzen, s. auch Kapitel 01 und 02.01. Daher stellen wir hier neben einer übersichtlichen Zusammenfassung der Veröffentlichungsstrategie lediglich die essenziellen Neu- und Weiterentwicklungen vor.





 <b>Leitmotive der Veröffentlichung</b>	<ul style="list-style-type: none"> <li>• Offener Zugang zu <ul style="list-style-type: none"> <li>• den Daten</li> <li>• dem Datenmodell</li> <li>• der entwickelten Software</li> </ul> </li> <li>• Community-Entwicklung bei <ul style="list-style-type: none"> <li>• Software</li> <li>• Anforderungsdefinition</li> <li>• Datenbedarfen</li> </ul> </li> </ul>
 <b>Zielgruppen</b>	<ul style="list-style-type: none"> <li>• Datengebende</li> <li>• Datennehmende</li> <li>• Datendienstleistende, z.B. Data Science-Anbietende</li> </ul>
 <b>Veröffentlichungskanäle</b>	<ul style="list-style-type: none"> <li>• Webfrontend</li> <li>• GitHub Repositories <ul style="list-style-type: none"> <li>• Post-Covid Datenmodell</li> <li>• EuroDaT</li> </ul> </li> <li>• Workshops</li> </ul>
 <b>Veröffentlichte Inhalte</b>	<ul style="list-style-type: none"> <li>• Nutzendeninformationen zur <ul style="list-style-type: none"> <li>• Anbindung an das Datenökosystem</li> <li>• Nutzung des Data Science Marktplatzes</li> <li>• Grundlage des (Privacy Preserving) Record Linkage</li> </ul> </li> <li>• Quellcode <ul style="list-style-type: none"> <li>• Datenmodell</li> <li>• EuroDaT Datentreuhänder-Plattform</li> </ul> </li> <li>• Vertrauensbildung</li> </ul>

Tabelle 4: Zusammenfassende Darstellung unseres Veröffentlichungskonzepts.

Die Zusammenfassung unseres Konzepts zur Veröffentlichung stellen wir in 04.03 vor. Hierbei schlüsseln wir das Konzept auf in die Kategorien Leitmotive, Zielgruppen, Veröffentlichungskanäle und Inhalte. Die in diesen Kategorien jeweils entwickelten Ansätze sind im Detail in unseren früheren Abschlussberichten beschrieben.

Auf diesen konzeptionellen Überlegungen aufbauend, haben wir in Stufe 3 die **Veröffentlichung unserer Arbeit vorangetrieben** und umgesetzt. Unseren Ansatz haben wir ausgebaut und in allen drei Veröffentlichungskanälen, **Webfrontend**, **GitHub**

**Repositories und Workshops**, verstetigt. Wir stellen im Folgenden unsere Weiterentwicklung an den Veröffentlichungskanälen und die jeweils neu veröffentlichten Inhalte vor.

## WEBFRONTEND – INFORMATIONEN LEICHT ZUGÄNGLICH MACHEN

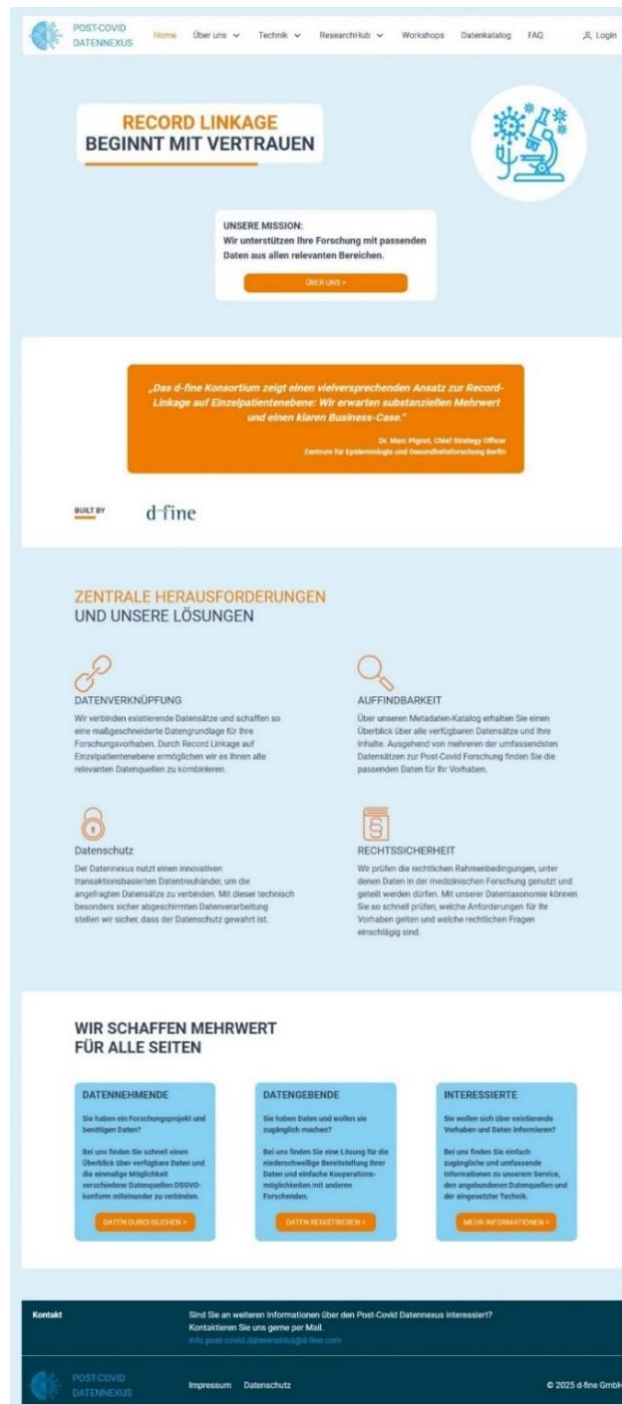


Abbildung 3: Aktualisierte Landing Page unseres Webfrontends.

Unser Webfrontend setzen wir unter anderem ein, um **einen leicht erreichbaren Zugang zum Datenökosystem** zu bieten und den Nutzenden inhaltliche Informationen bereitzustellen. Ausgehend vom im Abschlussbericht der [Stufe 2](#) vorgestellten Konzept für unser Webfrontend haben wir in Stufe 3 unseren Fokus darauf gelegt, möglichst vielen Personen umfassende Information zur Nutzung und dem Nutzen unserer Datenökosystem-Lösung zugänglich zu machen. Außerdem spielt in der praktischen Veröffentlichung von Webinhalten die empfängergerechte Präsentation eine entscheidende Rolle, weshalb wir einen Schwerpunkt auf eine gute **User Experience (UX)** gesetzt haben.

Konkret haben wir im Rahmen der Weiterentwicklung unseres Webfrontends die in Stufe 2 entwickelten Wireframes in ein funktionales und benutzerfreundlich klickbares Interface unter der bekannten Adresse <https://post-covid.dateninstitut.d-fine.dev/> umgesetzt, s. Abbildung 3. So hat sich unsere Website entscheidend weiterentwickelt, und kann den Anforderungen und Erwartungen realer Nutzender gerecht werden. Dabei haben wir eine neue UX/UI integriert, die eine intuitive und ansprechende Navigation ermöglicht.

Darüber hinaus haben wir das Webfrontend durch **neue Funktionen** erweitert, darunter Informationen zum **Research Hub**, der die zentrale Plattform für die von uns konzipierten **Data Science- und Forschungsdaten-Marktplätze** darstellt, s. Abbildung 4 und Kapitel 04.01. Außerdem haben wir im überarbeiteten Webfrontend Links zu den Code-Repositories sowohl von [EuroDaT](#) als auch unserem Datenökosystem inkl. Client-Apps und Backend (s. Abbildung 4, nächster Abschnitt und Kapitel 06.03), sowie einen Datenkatalog eingerichtet, um eine schnelle und gezielte Auffindbarkeit der im System gespeicherten Daten zu gewährleisten. Zusätzlich haben wir einen Bereich für Event-Verlinkungen eingerichtet, um **Workshops und Veranstaltungen** zu bewerben und die aktive Teilnahme der Community zu fördern, s. übernächster Abschnitt.

Diese Weiterentwicklungen stellen sicher, dass unsere Website als umfassender Zugangspunkt zu unserem Datenökosystem fungiert und den Nutzern alle relevanten Informationen und Möglichkeiten nahtlos zur Verfügung stellt.



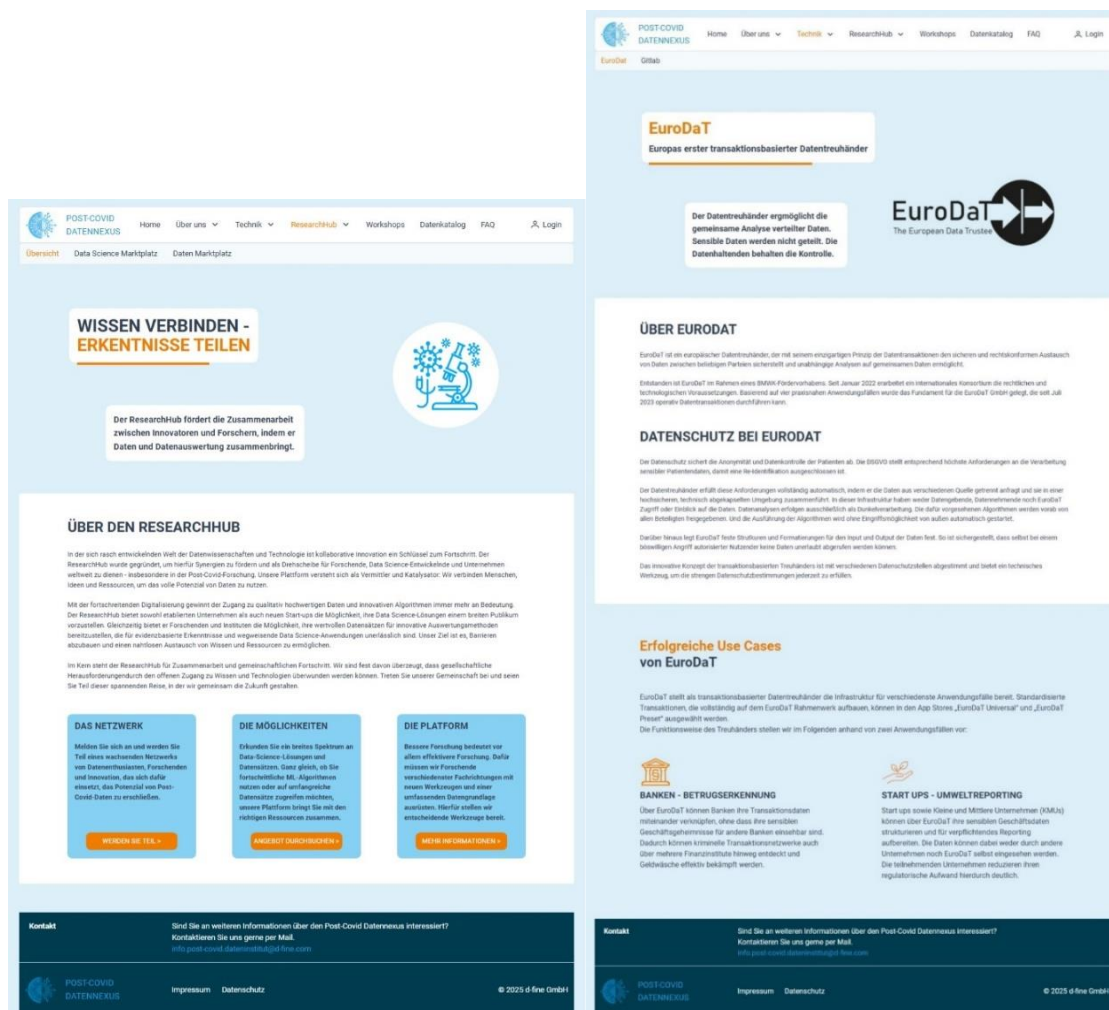


Abbildung 4: Startseite des neuen Research Hubs mit Verweisen auf die Data Science- und Daten-Marktplätze und Überarbeitete Informationsseite zur eingesetzten Technik.

## CODE REPOSITORY – INTERAKTIVE KOLLABORATION

Die Zwecke und Zielsetzungen des von uns eingesetzten Code Repositories als Steuerungs- und Versionierungstool unserer Entwicklungsarbeit, **Plattform für interaktive Kollaboration**, sowie als Werkzeug zur Erhöhung der Sichtbarkeit und Transparenz haben wir im [Abschlussbericht der Stufe 2](#) ausgeführt. Während sich an diesen Grundüberlegungen nichts geändert hat, haben wir die Stufe 3 gemäß ihrem zugrundeliegenden Prinzip genutzt, um den Betrieb des Repositories vorzubereiten. Hierfür haben wir das gesamte Repository von unserer betriebsinternen Plattform auf die **öffentliche GitHub-Plattform** umgezogen, und dort unter <https://github.com/d-fine/Post-COVID-Dateninstitut> ein neues Repository eingerichtet, s. Abbildung 5. Aktuell ist dieses noch privat, da wir abschließende Entwicklungsarbeiten und Qualitätssicherungsmaßnahmen umsetzen. Sobald diese Arbeiten aber abgeschlossen sind, lässt sich das GitHub-Repository mit einem Klick unverändert veröffentlichen und ist unter der angegebenen Adresse frei zugänglich.



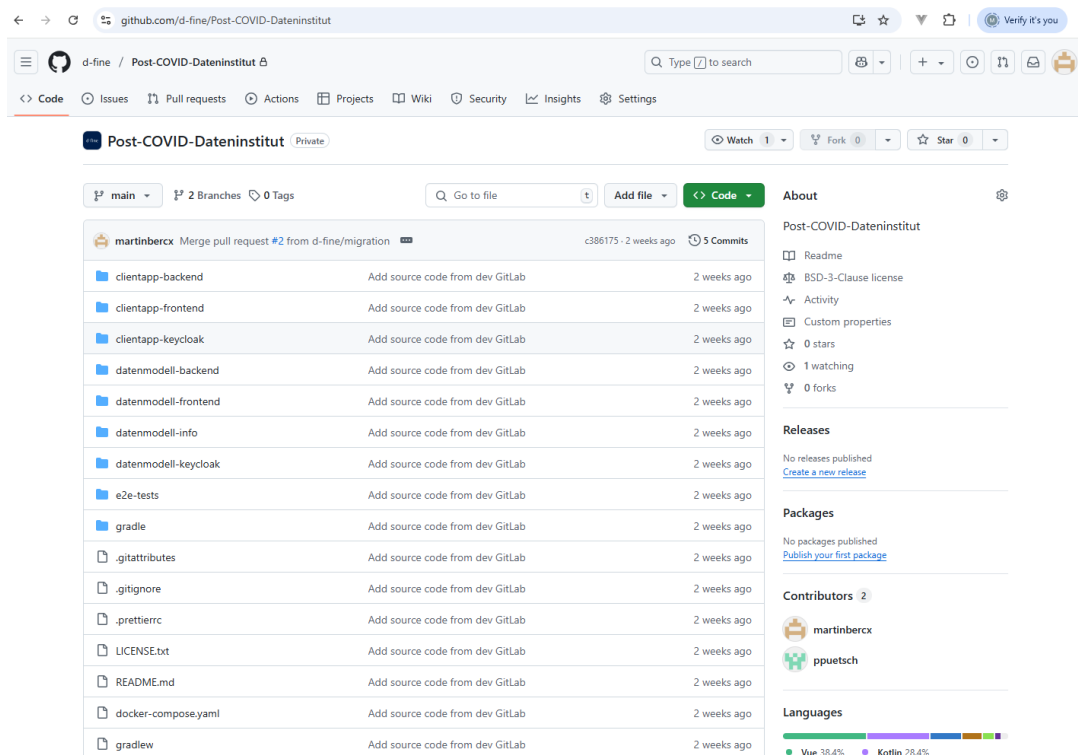


Abbildung 5: Zu veröffentlichendes GitHub Code-Repository.

## WORKSHOPS – INFORMIEREN UND DISKUTIEREN

Neben unserem eben beschriebenen Ansatz, wie wir das Datenökosystem technisch zugänglich machen, gehen wir jetzt auf Workshops als zentrales Element unseres Veröffentlichungskonzepts ein. **Persönliche Austauschformate** sind ebenso wichtig wie technische Veröffentlichungskanäle, um die Öffentlichkeit von der Existenz und dem Mehrwert unserer Lösung zu überzeugen und sie in den Diskurs über deren Nutzung mit einzubinden.

Wir haben hierzu bilaterale Workshops mit unseren assoziierten Partnern von NAKO und NAPKON sowie der Treuhandstelle Greifswald durchgeführt, welche als verwaltende Stelle der personenbezogenen Daten sowohl von NAPKON als auch NAKO fungiert. **Diese Partner haben wiederum unser Projekt auf Konferenzen und anderen Events als erfolgreiche Kollaboration vorgestellt.** Hierdurch konnten wir den Bekanntheitsgrad unseres Projektes und der entwickelten Konzepte weiter steigern. Darüber hinaus war insbesondere der Austausch mit der Treuhandstelle Greifswald von großem Wert für eine perspektivische Inbetriebnahme des Datenökosystems. Denn jegliche Datenlösung im Gesundheitssektor muss die Verwaltungsstellen von personenbezogenen Daten als zentrale Stakeholder in die Umsetzung mit einbinden. Durch unseren direkten Kontakt konnten wir die notwendigen Voraussetzungen und Anforderungen für eine **Mitarbeit einer operativen Treuhandstelle erarbeiten.**

Außerdem haben wir einen **speziellen Community-Workshop** organisiert, um **Treuhandmodelle im Gesundheitssektor allgemein und unser Post-COVID-Datenökosystem** inklusive der Client-Apps im Besonderen zu bewerben, offene Fragen einzusammeln und Anforderungen der medizinischen Forschungs-Community zu erheben. Die Organisation haben wir in Kooperation mit dem [DatenTreuhand Kompetenznetzwerk DaTNet](#) geleistet und als Termin den **13. Mai 2025** fixiert. Details hierzu werden wir in den kommenden Wochen auf unserer [Website](#) veröffentlichen. Hierdurch wird auch unser Engagement deutlich, das von uns entwickelte **Datenökosystem auch nach dem Ende der Challenge im Markt zu platzieren**.

#### 04.04 INNOVATIONSGRAD DES ANSATZES

Die zentrale Innovation unseres Ansatzes ist die **rechtssichere Zusammenführung besonders schutzbedürftiger Gesundheitsdaten**. Diese Innovation ist dabei allerdings kein Selbstzweck, sondern schafft einen klaren Mehrwert für die Nutzenden, wie mit der Jury im Rahmen des Abschlussevents der Stufe 2 diskutiert. Entsprechend ist unser Wertversprechen der Treiber unserer Innovation. So kann sich das Datenökosystem langfristig am Markt positionieren und skalieren, unabhängig vom finalen Betriebsmodell. Für eine Übersicht über die möglichen Betriebsmodelle verweisen wir auf unseren [Abschlussbericht der Stufe 1](#) und Kapitel 08 in diesem Bericht. Im Folgenden fassen wir zunächst die zentralen Mehrwerte zusammen, die Forschende aus dem von uns entwickelten Datenökosystem ableiten können:

#### UNSER WERTVERSPRECHEN ALS INNOVATIONSTREIBER

##### Record Linkage – Daten auf Patientenebene verknüpfbar machen

Als Record Linkage bezeichnet man einen Prozess, bei dem Daten aus verschiedenen Quellen auf Ebene einzelner Einträge (engl.: *records*) miteinander verknüpft werden – in der medizinischen Forschung meist Informationen zu einer natürlichen Person. Ein solcher Prozess in der medizinischen Forschung ist in Deutschland Stand heute nur mit hohem Aufwand möglich, da zum einen kein Unique Identifier existiert, der eine einfache und eindeutige Zuordnung zwischen unterschiedlichen Datensätzen ermöglicht. Zum anderen ist die Verknüpfung auf personenidentifizierenden Informationen (PII) selbst mit der Hilfe von Treuhandstellen schwierig, da die DSGVO hohe Anforderungen an die Weitergabe von Gesundheitsdaten stellt. Unsere Lösung verspricht nun als erster Datenraum eine **breit zugängliche Umsetzung von Record Linkage zwischen beliebigen Datenquellen** der medizinischen Forschung und damit einen qualitativ neuen, signifikanten Mehrwert für die Forschenden. Dies verspricht bisher unerreichbare Einsichten und enormes Marktpotenzial.

Die Nutzung eines transaktionsbasierten Datentreuhänders stellt hierbei die zentrale Innovation dar, die dieses Wertversprechen erst ermöglicht, indem sie die **DSGVO-Hürden mit einer innovativen technischen Lösung überwindet**. Hierbei ist außerdem von Vorteil, dass der von uns eingesetzte Treuhänder EuroDaT eine Open Source Software ist und daher bedeutend günstiger in der Nutzung als kommerzielle Alternativen.

Neben diesem qualitativ neuen Wertversprechen bietet die von uns entwickelte Lösung Forschenden weitere Mehrwerte, von denen wir an dieser Stelle zwei hervorheben.

#### Auffindbarkeit – Daten zugänglich und wiederverwendbar machen

Die Auffindbarkeit von Daten für Forschende stellt einen der Pfeiler der **FAIR-Prinzipien** und entsprechend auch einen zentralen Aspekt im von uns entwickelten Datenökosystem dar. Der Zugang zu vielfältigen Informationen und deren Nutzung ist nur möglich, wenn die Forschenden relevante Datensätze kennen und die jeweiligen Datenquellen auch lokalisieren können. Während die grundsätzliche Idee der **Auffindbarkeit von Daten** in der medizinischen Forschung von mehreren Initiativen bearbeitet wird, wird sie in unserem Datenökosystem durch die innovative Kombination mit der unmittelbaren technischen Nutzbarkeit zu einem operativen Mehrwert erhoben. Wir setzen hierbei auf die **Nachnutzung des etablierten NFDI4Health-Katalogs**, der im Bereich der Gesundheitsforschung bereits Vertrauen genießt, s. Kapitel 05.01. Das Einbeziehen existierender Lösungen ist dabei besonders sinnvoll, da es Ressourcen spart und **Doppelstrukturen vermeidet**, wodurch der Fokus auf die Weiterentwicklung und Integration neuer technischer Möglichkeiten gelenkt wird. Durch die Nutzung eines bewährten Datenkatalogs müssen Datengebende ihre Daten nicht doppelt registrieren, sondern können sie im Ökosystem nahtlos auffindbar machen. Gleichzeitig nutzt die Lösung damit einen Vertrauensvorschuss, indem sie auf erprobte Strukturen aufbaut. Dies trägt maßgeblich dazu bei, die Effizienz der Datenauffindbarkeit und schlussendlich -nutzung zu steigern, indem es Datennehmenden ermöglicht, Daten schnell und gezielt zu identifizieren und die zugehörigen Metadaten einzusehen.

#### Marktplatz für Data Science Lösungen – Kooperationen stärken

Wir entwickeln eine Plattform, auf der Forschende Data Science Lösungen wie Algorithmen oder Datenmodelle abrufen können und Entwickler die benötigten Datensätze finden, um ihre Lösungen voranzutreiben, vgl. Kapitel 04.01. Diese bezeichnen wir als **Marktplatz für Data Science Lösungen**. Diese Plattform bringt Parteien zusammen und ermöglicht ihnen gegenseitigen Nutzen, indem sie eine sinnvolle Aufgabenteilung unterstützt. Forschende können sich dadurch stärker auf ihre primäre Aufgabe, die Forschung, konzentrieren, während Data Scientists die

Freiheit erhalten, eigenständige Lösungen zu entwickeln. Ein solcher Marktplatz ist insbesondere für kleinere Forschungsgruppen und Unternehmen, wie beispielsweise KMUs oder einzelne Forschungsgruppen, von erheblichem Wert. Diese verfügen in der Regel nicht über die Ressourcen großer Verbünde oder Unternehmen, um dedizierte Kooperationen mit der anderen Seite aufzubauen. Hierdurch wird die **Medizindatenforschung demokratisiert**, indem ein Zugang geschaffen wird, der es auch kleineren Einheiten ermöglicht, auf spezialisiertes Know-how und umfangreiche Ressourcen zuzugreifen und konkurrenzfähige wissenschaftliche Arbeiten zu leisten, s. Kapitel 04.01.

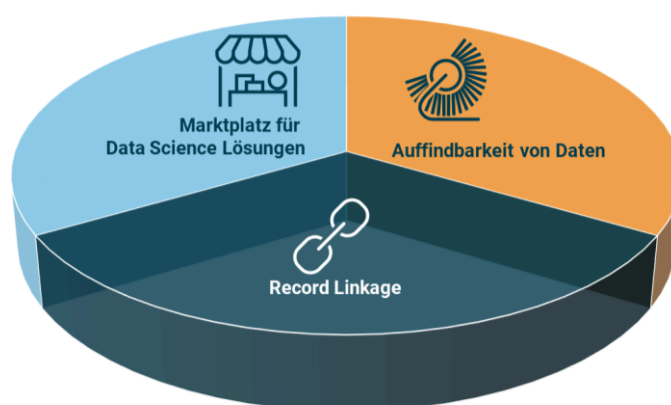


Abbildung 6: Unser dreifaches Wertversprechen.

Um dieses **dreifache Wertversprechen von Record Linkage, Auffindbarkeit der Daten und Data Science Lösungen** für die Forschenden zu realisieren, haben wir Innovationen in den zentralen Bereichen Technik, Recht und Prozesse zusammengebracht. Diese Kernbereiche des Innovationsgrads unseres Ansatzes stellen wir im Folgenden vor.

#### TECHNOLOGISCHE INNOVATION – NEUE MÖGLICHKEITEN DURCH TECHNISCHEN FORTSCHRITT

Wie oben beschrieben ist die technologische Innovation das zentrale Element unserer Arbeit, die rechtliche und prozedurale Herausforderungen in einer integrierten und innovativen technischen Lösung beantwortet. Diese technische Innovation verbindet dabei den wegweisenden Einsatz eines transaktionsbasierten Datentreuhänders, s. Kapitel 06, mit Agilität in der Anbindung neuer Datenquellen, s. Kapitel 03.02 und einer offenen Lizenzierung des Datenraums, s. Kapitel 04.03.

#### RECHTLICHE INNOVATION – SICHERHEIT DURCH DATENSCHUTZ

Neben der technischen Innovation ist die rechtliche Arbeit ein ebenso zentrales Innovationspotential unseres Post-COVID-Datenmodells. Hierbei ist die **Neubewertung der DSGVO-Vorschriften** besonders wichtig, die durch den innovativen Prozessbezug

des Re-Identifizierungsrisikos eine automatische Erfüllung der DSGVO von Datenverarbeitungsprozesse im Datentreuhänder ermöglicht. Ein weiteres innovatives Merkmal ist die **Anpassungsfähigkeit unseres Modells** an zukünftige Regulatorik, wie im [Abschlussbericht Stufe 2](#), Kapitel 04.04, dargelegt. Hierbei wurden bereits aktuelle und aufkommende regulatorische Standards wie der Europäische Gesundheitsdatenraum (EHDS), das Gesundheitsdatennutzungsgesetz (GDNG), der AI Act und der Data Act bewertet, um eine zukunftssichere Grundlage zu schaffen.

Zur Unterstützung dieser Ansätze haben wir außerdem noch konkrete rechtlich nutzbare Werkzeuge geschaffen, die die Nutzenden bei der Einschätzung und Bearbeitung rechtlicher Anforderungen unterstützen: Zunächst verweisen wir auf die von uns entwickelte **Datentaxonomie**, s. [Abschlussbericht der Stufe 2](#). Mit ihrer datenschutzrechtlichen Klassifizierung verfügbarer Datensätze bietet diese Taxonomie einen besonderen Mehrwert für die Nutzenden unseres Datenökosystems.

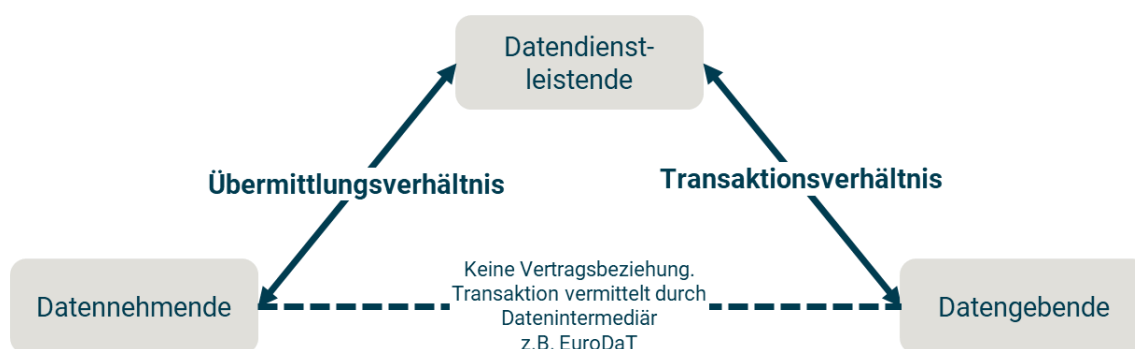


Abbildung 7: Vertragsbeziehungen im hier analysierten Szenario.

Darüber hinaus haben wir aufbauend auf Vorarbeiten aus dem EuroDaT-Konsortium und den im [Abschlussberichts der Stufe 2](#) analysierten **Vertragsbeziehungen** begonnen einen **Vertragsbaukasten** zu konzipieren. Mit diesem können Nutzende **niederschwellig und modular grundsätzliche Vertragsbeziehungen wie Datennutzungs- und -bereitstellungsverträge entwerfen** und einzelfallbezogen anpassen. Trotz der mit dem Vertragsrecht und der basalen Garantie der Vertragsfreiheit bestehenden Flexibilität und Innovationsoffenheit lassen sich hierbei im Sinne einer typisierenden und regulierenden Nutzung des Privatrechts einige Grundzuordnungen vornehmen.

Wir haben als beispielhaftes Szenario eine Datentransaktion innerhalb des Ökosystems analysiert, bei der die Betriebsstelle des Datenraums als Datendienstleisterin den Datenabruf bei den Datengebenden, die Datenanalyse und die Ausgabe der Ergebnisse an einen Datennehmenden verantwortet. Wir gehen dabei davon aus, dass die Datennehmenden alle notwendigen Use & Access Verfahren für zugangsbeschränkte Daten außerhalb des Datenraums durchlaufen haben, vgl. unser Konzept der

Datensouveränität. Innerhalb des Datenraums hingegen bestehen **vertragliche Verbindungen zwischen den Datengebenden, Datendienstleistenden (Datenraumbetreiber) und den Datennehmenden**. Typisch und für die vertragsrechtliche Einordnung elementar ist dabei die Einsicht, dass innerhalb des Datenraums vertragliche Beziehungen zwischen den Datengebenden und den -dienstleistenden sowie zwischen den Datendienstleistenden und -nehmenden bestehen, nicht hingegen zwischen den Datengebenden und den -nehmenden. Es handelt sich – jenseits des Intermediärs – somit um eine **vertragliche Dreiecksbeziehung**, s. Abbildung 7.

Das zwischen den Datengebenden und -dienstleistenden bestehende Verhältnis betrifft die Zurverfügungstellung bzw. den Erwerb von Daten und lässt sich als **Transaktionsverhältnis** umschreiben, s. Rechtsliteratur. Dass die Auswahl der Datengebenden wie gegebenenfalls auch der konkreten Daten regelhaft an den Interessen und Zielen der Datennehmenden orientiert ist, ändert insoweit nichts daran, dass ungeachtet der genauen Anbahnungssituation zu keinem Zeitpunkt eine rechtliche Sonderbeziehung zwischen den Datengebenden und Datennehmenden entsteht. Von diesem Transaktionsverhältnis ist das Verhältnis zwischen den Datendienstleistenden und -nehmenden zu unterscheiden. Es kann als **Übermittlungsverhältnis** bezeichnet werden, weil die vertragliche Verpflichtung hier darauf gerichtet ist, dass die Datendienstleistenden die Daten, die sie von den Datengebenden erhalten haben, über einen vertraglich festgelegten Algorithmus innerhalb der Infrastruktur des Intermediärs verarbeiten und das Analyseergebnis an die Datennehmenden übermitteln.

Somit lassen sich bei der transaktionsbezogenen Datentreuhand vor allem **zwei relevante Vertragsverhältnisse** ausmachen: Für das Verhältnis zwischen Datengebenden und -dienstleistenden, das **Transaktionsverhältnis**, dürften zumindest im Regelfall die mietvertraglichen Bestimmungen der §§ 548a, 535 ff. BGB passend sein. Demgegenüber fehlt eine entsprechende Passgenauigkeit für das Verhältnis zwischen Datendienstleistenden und -nehmenden, das **Übermittlungsverhältnis**. Dieses lässt sich keinem der kodifizierten Vertragstypen des BGB zuordnen. Vor dem Hintergrund der prinzipiellen Vertragsfreiheit (§§ 242, 305 BGB) bedeutet das aber keineswegs ein Handlungsverbot, sondern impliziert lediglich, dass die entsprechenden Gestaltungen als Vertrag sui generis zu qualifizieren sind.

#### PROZEDURALE INNOVATION – EFFIZIENZ STEIGERN

Sowohl die technologische als auch die rechtliche Innovation sind in unserem Projekt nicht ohne **konzeptionell neue Prozesse** zu realisieren, die wir in Kapitel 03.04 des Berichts der Stufe 2 vorgestellt haben. Hierbei haben wir insbesondere durch unser sektorübergreifende Stakeholder-Einbindung eine gemeinsame Kommunikationsbasis

aufgebaut. Nur auf deren Grundlage ist eine rechtssichere Datenintegration und -nutzung über Sektorgrenzen hinweg möglich. Darüber hinaus ist in unserm Ansatz innovativ, dass wir **Datengebende und -nutzende in unserem Datenökosystem prozedural gleichberechtigt berücksichtigen** und für beide Seiten angepasste Prozesse zur Nutzung des Datenraums entwickelt haben, s. Kapitel 03.04 des [Berichts der Stufe 2](#).

Somit bildet unser technisch, rechtlich und prozedural integrierter Ansatz die notwendige Grundlage für die in dieser Challenge gesuchte Innovation in der Post-Covid Forschung.

## 05. DATENMODELL

### 05.01 VORLIEGENDE DATENSÄTZE

In den vorhergegangenen Stufen der Challenge haben wir uns ein umfassendes Bild der Post-Covid-Datenlandschaft in Deutschland und Europa gemacht. Darauf aufbauend haben wir eine **maßgeschneiderte Auswahl der relevantesten Datensätze** getroffen und diese in unseren Fokus genommen. In der Ausgestaltung des MVP, insbesondere zur Umsetzung und Testung, haben wir diese Auswahl bewusst klein gehalten. Gleichzeitig haben wir prozedural und technisch sichergestellt, dass **beliebige Erweiterungen möglich** sind und somit die gesamte Post-Covid-Datenlandschaft perspektivisch in den Daten nexus integriert werden kann.

#### DATENSÄTZE IM FOKUS

Im Bericht zur [Stufe 2](#) der Challenge haben wir einen Überblick über die von uns fokussierten Datensätze gegeben. In Tabelle 5 ist eine Aktualisierung dieses Überblicks zu finden. Wie ersichtlich, haben wir von sämtlichen avisierten Datensätzen sowohl **Testdaten** als auch eine **Metadatenbeschreibung** vorliegen.

Datensatz	Verfügbarkeit	Dummy-Daten	Metadaten
NAPKON	Use & Access	Liegen vor	Liegen vor
NAKO	Use & Access	Liegen vor	Liegen vor
Abrechnungsdaten	Use & Access	Liegen vor	Liegen vor
Rentenversicherungsdaten	Datenabhängig	Liegen vor	Liegen vor
Wearables	Datenabhängig	Liegen vor	Liegen vor

Tabelle 5: Übersicht der für die MVP-Umsetzung priorisierten Datensätze.

Von den Datensätzen im Fokus haben wir uns in dieser Stufe besonders auf die Daten von **NAPKON und NAKO** konzentriert. Diese Entscheidung haben wir bewusst getroffen, da es sich bei diesen um die größten Studiengruppen zu Post-Covid sowie



Volkskrankheiten handelt, deren Kombination bereits einen signifikanten Mehrwert für die Forschung bringt. Im Laufe unserer Zusammenarbeit konnten wir NAPKON und **NAKO darüber hinaus dafür gewinnen unsere Client-App zur Anbindung ihrer Daten zu erproben**. Aufbauend auf diesem Erfolg konnten wir eine erste Testtransaktion mit Daten von NAPKON und NAKO durchführen, s. Kapitel 05.03.

Auf eine Einzelbetrachtung der unterschiedlichen Datensätze verzichten wir an dieser Stelle, da detaillierte Informationen hierzu bereits im Kapitel 05.01 des Berichts zur [Stufe 2](#) sowie im Kapitel 04.01 des Berichts zur [Stufe 1](#) zu finden sind.

#### Nutzungsanträge – Use & Access erproben und Hindernisse beseitigen

Damit Forschende die wertvollen zugangsbeschränkten Daten großer Forschungsverbünde nutzen können, müssen sie die **Use & Access Prozesse** der datenhaltenden Stellen durchlaufen. Um diese komplexen Prozesse frühzeitig in das Datenökosystem zu integrieren und die spätere Nutzung des Datenraums für Forschende möglichst einfach zu gestalten, erproben wir im Rahmen der Post-Covid Challenge zwei solche Use & Access Prozesse als **Proof-of-Concepts**, und zwar bei den größtangelegten Datenquellen NAKO und NAPKON.

Im Rahmen dieser und typischer anderer Use & Access Verfahren muss man eine Forschungsfrage definieren, die mit den angefragten Daten bearbeitet werden soll. In [Stufe 2](#) haben wir hierfür die Frage „**Wie wirkt sich Post-COVID auf die Erwerbsfähigkeit aus?**“ ausgearbeitet. Konkret analysieren wir die Erwerbsverläufe unterschiedlicher Bevölkerungsgruppen vor und nach einer COVID-Infektion. Hierzu nutzen wir den Datenraum inklusive Treuhänderstrukturen und PPRL, s. Kapitel 05.03. Dadurch können wir praxisnahe Erkenntnisse zur Datennutzung gewinnen und darüber hinaus die gesellschaftlichen Auswirkungen von Post-COVID genauer beleuchten.

Primärer Fokus unserer Arbeit war der **Nutzungsantrag der NAKO**. Hier konnten wir unter anderem durch intensiven persönlichen Austausch mit Mitarbeitenden der NAKO unseren technisch fokussierten Antrag weitgehend vorbereiten. Gerade planen wir einen finalen Austausch mit der Transferstelle der NAKO, um einen optimierten Antrag einzureichen. Parallel arbeiten wir auch an einem Nutzungsantrag für die Daten der **NAPKON**, welcher ebenfalls zeitnah eingereicht werden soll.

Neben medizinischen sind auch sozialwissenschaftliche Daten für die Post-Covid Forschung relevant. Um diesen Forschungsbereich mit zu berücksichtigen, stehen wir in Kontakt mit dem Arbeitsbereich **Sozialwissenschaftliche Perspektiven** von Sport, Bewegung und Gesundheitsförderung der TU Chemnitz. Prof. Dr. Torsten Schlesinger und Dr. Katrin Müller zeigten in einem ersten Austausch großes Interesse und sind

durch ihre Vorarbeiten im Bereich Post-Covid und Arbeitsfähigkeit der ideale Partner für unser Vorhaben.

## OPEN DATA – INNOVATIONEN FÖRDERN

Zusätzlich zu den aktuell im Fokus stehenden, jedoch nur mittels aufwendiger Use & Access Verfahren zugängigen Datensätzen, haben sich Open Data Datensätzen als sinnvolle Ergänzung erwiesen. Unter Open Data versteht man frei zugängliche Daten, welche mittels offener und diskriminierungsfreier Lizenzen frei weiterverwendet werden können. Beispiele hierfür sind Geodaten, Bevölkerungsdaten oder Wetterdaten. Die Bundesregierung unterstützt diesen Ansatz mit ihrer [Open Data Strategie](#). Unser Ziel ist es, **im Datenraum aktiv die Integration von Open Data zu fördern**, s. Kapitel 06.03 für eine exemplarische Diskussion. Hierzu unterstützen wir sowohl die Einbindung von Open Data Daten in den Datenkatalog als auch eine einfache Nutzung der Daten für Auswertungen. Durch die Integration von Open Data in den Datenkatalog werden Nutzende auf die Daten aufmerksam gemacht. Ziel ist es, Forschende durch diese neuen, ergänzenden Daten **zu neuen Forschungsfragen zu inspirieren**.

### Open Data Integration im MVP

Im Rahmen des MVP wurde die exemplarische Umsetzung zunächst anhand eines Open Data Datensatzes erarbeitet und durchgeführt. Hierzu fiel die Wahl auf den Datensatz zur **Abwasser-Surveillance AMELAG des RKI**. Das Abwassermonitoring wird für die epidemiologische Lagebewertung genutzt, indem es die Konzentration von Infektionserregern, insbesondere der SARS-CoV-2-Viruslast, im Abwasser wiedergibt. Diese Zahlen können beispielsweise für regionale Forschungsfragen von großem fachlichem Interesse sein. Zum Datensatz gibt es einerseits Metadaten, andererseits stehen die Daten selbst offen zu Verfügung, mit ca. wöchentlicher Aktualisierung bis Ende der Projektlaufzeit am 31.12.2025. Der Datensatz wurde von uns in den Datenkatalog mit aufgenommen, wobei hierzu im folgenden Abschnitt zum Datenkatalog weitere Informationen folgen.

Zur zukünftigen Nutzung der Daten ist folgender Prozess vorgesehen: Bei Auswahl der Daten auf der Website für eine Datenauswertung werden diese **automatisiert aus dem entsprechenden GitHub Repo des RKI geladen** und für die Verarbeitung an EuroDaT weitergeleitet. So ist es möglich, die Daten in verschiedenen Auswertungen zu integrieren, s. Abschnitt weiter [unten](#) zur Verknüpfung. Eine Aktualisierung der Daten, bzw. die Sicherstellung, dass die Daten immer den aktuellen Stand haben, erfolgt durch den automatisierten Abruf des aktuellen Stands der Daten von GitHub, s. auch Kapitel 06.03. Im Rahmen der Arbeiten am MVP wurde dieses Vorgehen erarbeitet und erprobt und kann für künftige Auswertungen niederschwellig umgesetzt werden.

### Open Data Integration im Zielbild

Im Zielbild soll das Angebot an Open Data deutlich erweitert werden. Hierzu sollen einerseits von Seiten der Betreiber des Datenraums **aktiv Open Data Datensätze in den Datenkatalog mit aufgenommen** werden. Andererseits soll es die Möglichkeit gegeben, dass Nutzende Open Data Datensätze eigenständig mit aufnehmen können.

Die Verknüpfung der Daten mit anderen Daten kann je nach Datenlage auf unterschiedliche Weise erfolgen. So haben viele Open Data Datensätze einen regionalen Bezug, wie z.B. die oben beschriebenen AMELAG-Daten, die Informationen über den Standort der Messung enthalten. Daher **lassen sie sich direkt mit anderen Datensätzen verknüpfen**, die Ortsinformationen wie z.B. Postleitzahlen enthalten. Dies erfüllen viele medizinische Studiendaten wie beispielsweise die der NAKO. Für bereits erfolgreich durchgeführte Verknüpfungen werden diese Mapping-Methoden auf der Website nachgehalten und somit für andere Datennutzende nachnutzbar gemacht. In dem Fall einer neuen Art der Datenverknüpfung muss die entsprechende Verknüpfungslogik von den Datennutzenden an den Datenraum übergeben werden, s. Kapitel 06.03 für den entsprechenden Prozess.

Für die Nutzung der Datensätze im Rahmen späterer Auswertungen sind zwei Szenarien vorgesehen.

- Daten, welche **technisch automatisiert abgerufen** werden können, wie z.B. die AMELAG-Daten, bindet der Datenraum-Betreiber technisch in den Datenkatalog ein. So kann bei späteren Datenabfragen automatisiert auf die Daten zugegriffen werden.
- Für Daten, bei welchen eine automatisierte Abfrage nicht möglich ist, ist vorgesehen auf der Website eine **Upload-Möglichkeit** zu integrieren. Hierbei kann der Datennutzer auf der Weboberfläche einen beliebigen Datensatz hochladen, welcher anschließend für die Auswertung in EuroDaT mitverwendet wird. Der Datenhalter muss hierbei die Bereitstellung der Open Data Daten übernehmen. Ein Vorteil dessen ist es, dass verschiedene Versionen der Daten verwendet werden können und es beispielsweise auch möglich ist, für Use-Cases auf historische Daten zurückzugreifen, indem diese übergeben werden.

Eine Aktualisierung der Daten, bzw. die Sicherstellung, dass die Daten immer den aktuellen Stand haben, erfolgt im ersten Szenario durch den automatisierten Abruf des aktuellen Stands der Daten. Im zweiten Szenario ist der Datennutzer für die Bereitstellung der gewünschten Version der Daten selbst verantwortlich.

## DATENKATALOG – DATEN EINFACH ZUGÄNGIG MACHEN

In den Berichten zu [Stufe 1](#) und [Stufe 2](#) haben wir bereits beschrieben, dass ein **öffentlich zugänglicher Datenkatalog** zentral für unseren Datenraum ist. Ein solcher **Katalog** ermöglicht eine **Übersicht über bereits vorhandene Forschungsdaten** und **erhöht die Sichtbarkeit** der bereitgestellten Daten. Dadurch können Forschende niederschwellig passende Daten für ihre Forschungsfragen finden. Wie im Bericht zu [Stufe 2](#) beschrieben, nutzen wir dafür den **Daten- und Metadatenkatalog von NFDI4Health** nach, den wir in Stufe 3 in unser Datenökosystem integriert haben. Somit können wir einerseits auf einen großen, bereits bestehenden Datenbestand zurückgreifen, andererseits ermöglichen wir es Datenhaltenden auch, ihre Daten gezielt in unserem Datenkatalog zu hinterlegen. Durch diese Funktion wird der Aufwand für Datenhaltende einerseits reduziert, da sie Daten, die bereits bei NFDI4Health hinterlegt sind, bei Nutzung des Datennexus nicht erneut registrieren müssen, andererseits gewähren wir ihnen eine gewisse Freiheit, indem eine Registrierung im NFDI4Health Datenkatalog nicht verpflichtend vorausgesetzt wird. Der **Datenkatalog ist für alle Interessierten auf unserer Website einsehbar**, insbesondere auch für nicht registrierte Nutzende.

### Datenkatalog im MVP

In seiner gegenwärtigen Version bietet der Datenkatalog eine detaillierte Übersicht über die im Datenraum verfügbaren Primärdaten, sowie über ausgewählte Open-Source-Datensätze. Im Datenkatalog wird hierzu jeder Datensatz durch einen Titel, eine Kurzbeschreibung, die datengegebende Stelle und einen Link auf die zum Datensatz zugehörige Website beschrieben. Die integrierten Primärdaten sind der Datensatz der **NAKO sowie die Datensätze SÜP, HAP, POP und GECCO des NAPKON**. Das für die Bereitstellung der Daten im Datenkatalog notwendige technische Onboarding durchlaufen wir aktuell mit beiden datenhaltenden Stellen. Darüber hinaus möchten wir auf die Möglichkeiten von **Open Data Daten** hinweisen und integrieren deshalb die im NFDI4Health Datenkatalog inkludierten Open Data Datensätze des Robert Koch-Instituts in unserem Datenkatalog. Diese Datensätze können für die Post-Covid Forschung von großem Interesse sein. Sie beinhalten z. B. Zahlen zu COVID-19-Impfungen in Deutschland oder die im Open Data Abschnitt besprochene Abwassersurveillance AMELAG.

Titel	Datenhalter	Website
> German National Cohort (NAKO)	NAKO	<a href="#">Mehr Informationen</a>
> Intersectoral Platform (SÜP) of the National Pandemic Cohort Network (NAPKON)	NAPKON	<a href="#">Mehr Informationen</a>
> National Pandemic Cohort Network - High-resolution Platform (HAP)	NAPKON	<a href="#">Mehr Informationen</a>
> Analysis of the Pathophysiology and Pathology of Coronavirus Disease 2019 (COVID-19), Including Chronic Morbidity	NAPKON	<a href="#">Mehr Informationen</a>
> COVIDOM: Longterm Morbidity of SARS-CoV-2 Infection and COVID-19 Disease - Consequences for Health Status and Quality of Life (NAPKON-POP)	NAPKON	<a href="#">Mehr Informationen</a>
> German Corona Consensus (GECCO)	NAPKON	<a href="#">Mehr Informationen</a>
> COVID-19 Impfungen in Deutschland	Robert Koch-Institut	<a href="#">Mehr Informationen</a>
> German Index of Socioeconomic Deprivation (GISD)	Robert Koch-Institut	<a href="#">Mehr Informationen</a>

Abbildung 8: Datenkatalog auf unserer Website.

### NFDI4Health Study Hub Integration in den Datenkatalog

Die Anbindung der Open Data Metadaten erfolgt über eine Schnittstelle des [German Central Health Study Hub](#). Der **Health Study Hub** ist eine Plattform von NFDI4Health zum Auffinden und Veröffentlichen von Metadaten, mit dem Ziel **den Zugang zu strukturierten Gesundheitsdaten aus verschiedenen Quellen zu verbessern**. Er ermöglicht Forschenden und datenhaltenden Organisationen, Metadaten, die ihre Forschungsdaten charakterisieren, sowie zugehörige Dokumente wie Fragebögen und Metadatenkataloge gemäß der FAIR-Prinzipien zu veröffentlichen. Zudem unterstützt er Forschende dabei, Informationen über frühere und aktuelle Studien zu finden, und somit bereits **bestehende Forschungsdaten für ihre eigenen Projekte nachzunutzen**.

Der Health Study Hub stellt eine API zur Verfügung, die es allen Interessierten ermöglicht, direkten Zugriff auf die über 27.000 im Hub gespeicherten Daten zu erhalten. Wir nutzen diese API, um ausgewählte Metadaten wie Titel, Beschreibung und Website der Open Source Datensätze des RKI abzurufen. **Durch die API Anbindung können wir diese Auswahl ohne Entwicklungsaufwand erweitern** und perspektivisch alle im Health Study Hub gespeicherten Daten in unserem Datenökosystem bereitstellen. Der Datenkatalog ist dadurch sehr einfach um weitere auf dem Health Study Hub veröffentlichte Daten erweiterbar. Die Prozesse zur Erweiterung des Datenkatalogs werden in Kapitel 06.03 beschrieben. Viele der Metadaten im Health Study Hub werden regelmäßig aktualisiert und angepasst. Durch diese automatisierte Anbindung wird

gewährleistet, dass die entsprechenden Metadaten nach Aktualisierung im Health Study Hub auch in unserem Datenkatalog aktuell sind.

#### Datenkatalog im Zielbild

Perspektivisch haben wir weitere Funktionalitäten geplant, um den Datenkatalog umfangreicher und aussagekräftiger zu gestalten. Zunächst planen wir eine **Suchfunktionalität** für den öffentlich einsehbaren Datenraum, durch die z.B. auch nach Keywords gesucht werden kann. Damit Nutzende schnell entscheiden können, ob ein Datensatz für sie relevant sein könnte, planen wir die Bereitstellung von **Detailseiten** für die einzelnen Datensätze. Diese beinhalten weitere Informationen, wie dem Zeitraum der Datenerhebung, detaillierte Metadatenbeschreibungen und die Mitwirkenden.

Außerdem ist geplant, dass es für registrierte Datenhaltende die Möglichkeit gibt, Informationen zu Datensätzen selbst hochzuladen oder bestehende eigene Datensätze zu aktualisieren. Diese werden dann nach manueller Überprüfung veröffentlicht, siehe Kapitel 03.02. Registrierten Nutzenden wird zusätzlich angezeigt, welche Verknüpfungen von Datensätzen bereits durchgeführt wurden, sodass diese niederschwellig nachgenutzt werden können.

## 05.02 ORIENTIERUNG AN BRANCHEN-STANDARDS

Bereits in den ersten beiden Stufen der Challenge, s. [Bericht der Stufe 2](#), wurde großer Wert daraufgelegt, bestehende **semantische und technische Standards der Datenhaltenden möglichst nahtlos in das System zu integrieren**. Ziel war es – und bleibt es – den Aufwand für die Bereitstellung so gering wie möglich zu halten, während gleichzeitig die Daten für Nutzende strukturiert, interoperabel und FAIR verfügbar gemacht werden. Der Fokus liegt dabei auf transaktionsbasierten Datenflüssen und der Vermeidung starrer Transformationsprozesse, insbesondere unter Nutzung etablierter Strukturen wie dem NFDI4Health-Metadatenkatalog, wie im vorherigen Kapitel 05.01 im Absatz NFDI4Health Study Hub beschrieben.

Darüber hinaus sind **regulatorische Entwicklungen** für unser Projekt von besonderer Relevanz, da sie maßgeblich beeinflussen, welche Standards künftig eingesetzt und gefordert werden. Mit der **in Kraft getretenen European Health Data Space (EHDS)** Verordnung nimmt ein zentraler Rahmen für die europäische Gesundheitsdateninfrastruktur zunehmend Gestalt an. Für uns und unsere Kollaborationspartner ist es von großem Interesse, welche Strukturen im Rahmen dieser Umsetzung auf nationaler und europäischer Ebene konkret etabliert werden. Gerade vor diesem Hintergrund sind anschlussfähige und zukunftsorientierte Infrastrukturen von

zentraler Bedeutung. Entsprechende Entwicklungen berücksichtigen wir daher kontinuierlich in unserer Systemkonzeption.

Ein konkretes Beispiel ist der Umgang mit dem **Metadatenstandard HealthDCAT-AP**, der im Verlauf der dritten Stufe von der **nationalen Datenzugangs- und Koordinierungsstelle (DACO) am BfArM** für den [Aufbau des eigenen Metadatenkatalogs ausgewählt](#) wurde. Unsere Anbindung an den NFDI4Health-Katalog ist bereits weit fortgeschritten; dieser wird seitens des NFDI-Teams aktiv mit Blick auf Interoperabilität und Standardkompatibilität weiterentwickelt (s. vorheriges Kapitel 05.01).

Unser Ansatz bleibt damit bewusst **adaptiv und anschlussfähig**: Ziel ist es, auf bereits vorhandenen, in der Fachcommunity breit getragenen Standards aufzubauen, Kompatibilität zu wahren und zukünftige Entwicklungen frühzeitig zu berücksichtigen.

### 05.03 VERKNÜPFUNG VERSCHIEDENER DATENQUELLEN

Grundlage der modernen Forschung sind Daten. Auf ihnen beruhen Auswertungen, Statistiken und die daraus gezogenen Schlüsse und Entscheidungen. Daten werden täglich in großen Mengen und in den verschiedensten Bereichen des Alltags erfasst. Jedoch ist es insbesondere im Bereich der medizinischen Forschung immer noch eine große Herausforderung, die gesammelten **Daten zu verknüpfen und dadurch tiefgründigere Einsichten zu erlangen**. Dies liegt an unterschiedlichen Hürden, welche in den vorangegangenen Stufen beleuchtet wurden. Mit Hilfe des von uns entwickelten Datenraums wollen wir die **Verknüpfbarkeit vereinfachen und ein Record Linkage auf Personenebene ermöglichen**. Hierbei liegt der Fokus im Rahmen des MVP auf der Verlinkung mittels Hashes (PPRL), siehe unten und Kapitel 05.03 des [Berichts zur Stufe 2](#). Perspektivisch wird auch ein Record Linkage auf personenidentifizierenden Merkmalen (PII) vorgesehen, worauf wir hier nicht weiter eingehen, sondern auf das Kapitel 05.03 im vorherigen Bericht verweisen.

#### PPRL – DATEN SICHER AUF PATIENTENEBENE VERKNÜPFEN

Wir erinnern an die wichtigsten Schritte der Verknüpfung von Daten mittels PPRL:

##### Privacy Preserving Record Linkage über Hashes (PPRL) – Eine Wiederholung

1. Basierend auf einem von uns bereitgestellten Hash-Algorithmus werden die personen-identifizierenden Daten gehasht.
2. Die Hash-Pseudonym-Tabellen sowie die Datensätze inklusive der Pseudonyme werden von den Treuhandstellen/Datenhaltenden an die [EuroDaT](#) App übermittelt.
3. Diese Hash-Pseudonym-Tabellen werden mittels eines PPRL-Algorithmus in eine Pseudonym-Matching-Tabelle überführt.
4. Die Hashes werden nach der Erstellung der Pseudonym-Matching-Tabelle gelöscht.



5. Die Datensätze werden basierend auf der Pseudonym-Matching-Tabelle miteinander verknüpft.
6. Zur Minimierung des Re-Identifikationsrisikos wird eine Re-Identifikationsprüfung (Re-ID-Check, s. [übernächstes Unterkapitel](#)) durchgeführt.
7. Basierend auf dem Ergebnis des Re-ID-Checks und den Datennutzungsbedingungen werden entweder die verknüpften Daten oder Analyseergebnisse herausgegeben.

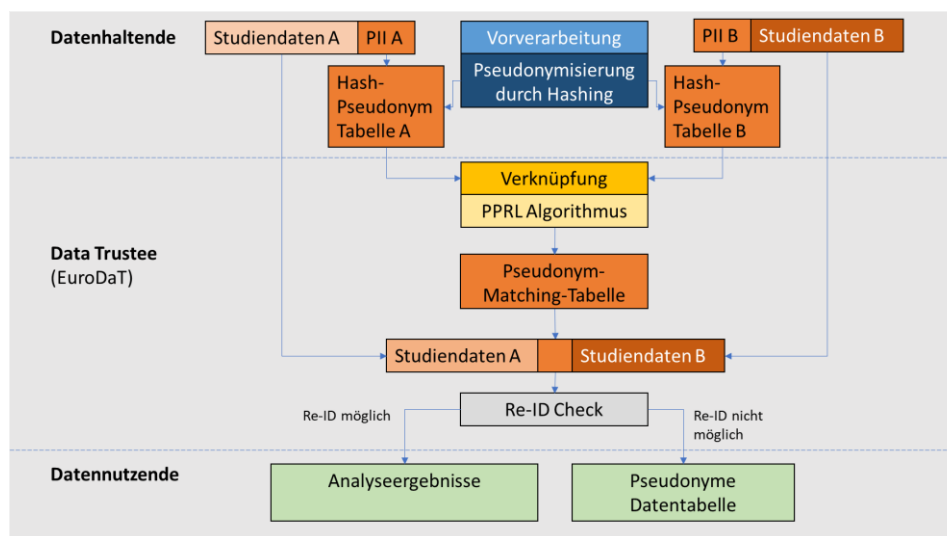


Abbildung 9: Schaubild zum Prozess der Datenverknüpfung mittels Hashes (PPRL).

### Bloomfilter – Essenziell für die sichere Datenverknüpfung

Der erste Schritt einer Datenverknüpfung ist das Hashen der PII. Hierzu können verschiedenste Hash-Algorithmen eingesetzt werden. Wir haben uns für die Erstellung der in der Medizinbranche oft verwendeten **Bloomfilter** entschieden. Entscheidende Vorteile der Bloomfilter sind ihre **Zeiteffizienz**, ihre **Skalierbarkeit** sowie der **geringe benötigte Speicherplatz**. Technisch basiert ein Bloomfilter auf einer Folge aus Nullen und Einsen, einem sogenannten Bit-Array. Die Länge wird dabei im Voraus festgelegt. Die PIIs werden in eine Bloomfilter-Datenstruktur überführt, indem eine ausgewählte Anzahl an Hashfunktionen festlegen, an welchen Positionen in der Zahlenfolge Nullen oder Einsen stehen. Diese so entstandene **Kodierung verhindert Rückschlüsse auf die ursprünglichen Identifikatoren**, ermöglicht jedoch eine fehlertolerante Verknüpfung von Datensätzen.

Bloomfilter können durch verschiedene Techniken erzeugt werden. Um den Datennehmenden möglichst viele Entscheidungsmöglichkeiten zu bieten, wollen wir ihnen auf unserer Homepage umfassend Informationen über verschiedene Einstellungen und Algorithmen zur Verfügung stellen. Anschließend **können die verschiedenen Optionen für eine Verknüpfung selbst konfiguriert werden**. Mögliche Auswahlkriterien sind hierbei die Art und Weise wie Bloomfilter generiert werden (z.B.

durch double hashing oder universal hashing), die Länge des Bloomfilters, das Setzen von Seeds (pseudozufällige Zahlen) oder Salts (zufällige Zeichenfolgen) und das Falten (halbieren der Länge) der Bitfolge. Details hierzu, inklusive einer Erläuterung der zugehörigen Vor- und Nachteile, werden im Infomaterial auf der [Website](#) zur Verfügung gestellt. Da nicht alle Datennehmenden ein Interesse daran haben, sich ausgiebig mit Bloomfiltern und Matching-Algorithmen zu beschäftigen, legen wir auch vorbereitete **Standardeinstellungen** fest.

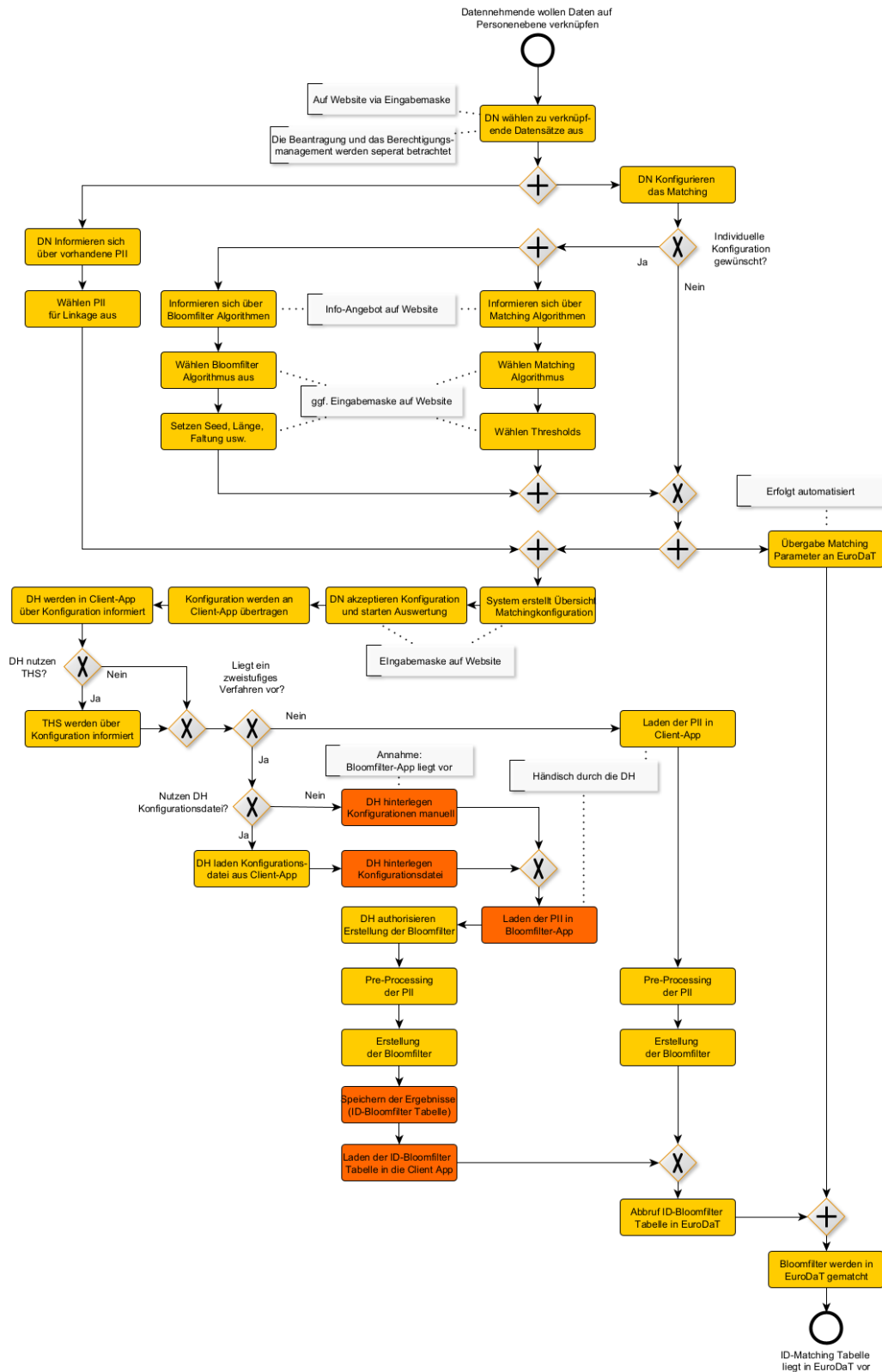
Wie erwähnt, werden Bloomfilter in der Medizinbranche vielfach verwendet. Eine etablierte Anbieterin eines entsprechenden Algorithmus ist die Treuhandstelle Greifswald, welche die Nutzung von Bloomfilter in das von ihnen entwickelte Tool [E-PIX](#) integriert hat. **E-PIX** ist ein Tool für standortübergreifendes Identitätsmanagement und wird unter anderen an vielen Deutschen Universitätskliniken bereits aktiv verwendet. Zur Wiedererkennung von Patienten und Sicherstellung, dass Patienten nicht doppelt geführt werden, nutzt das Tool Methoden des Record Linkage. Da ein wichtiger Aspekt unserer Entwicklung das **Vermeiden von Doppelstrukturen** ist, haben wir uns bei der Umsetzung stark an den Inhalten und Konfigurationen von E-PIX gehalten. Die zugrundeliegende Software wurde mit einer GPL 3-Lizenz veröffentlicht und darf **nachgenutzt** werden. Im MVP setzen wir zunächst eine einheitliche Implementierung um mit der Perspektive, beliebige Konfigurationen im Zielbild zu ermöglichen. Auf die Implementierung und technische Details gehen wir in Kapitel 06.03 näher ein.

#### Matching-Algorithmen – Fehlertolerante Verknüpfung von Datensätzen

Analog zur Bloomfilter-Konfiguration können sich Datennnehmer auf der Homepage auch über **verschiedenen Matching-Algorithmen** informieren und diese gemäß ihren Anforderungen auswählen. Hier steht zum Beispiel die Nutzung verschiedener Distanz-Metriken wie die Hamming-Distanz oder Levenshtein-Distanz zur Auswahl. Zusätzlich kann ein beliebiger Threshold für das Matching gesetzt werden. Mit Hilfe des Threshold wird festgelegt, wie hoch der Grad der Übereinstimmung zweier Bloomfilter sein muss, um sie als Match zu kategorisieren. Somit ist es möglich **variable, fehlertolerante Verknüpfung von Datensätzen durchzuführen**.

In unserem MVP setzen wir zunächst auf ein exaktes Matching, d.h. die als Bloomfilter kodierten PII werden bitweise miteinander verglichen und nur bei exakter Übereinstimmung als Match gekennzeichnet. Im Zielbild ist eine beliebige Auswahl an Matching-Algorithmen und Thresholds möglich.

## PPRL im Datenexus – Unser Zielbild als Prozess zusammengefasst



Um den **Prozess des PPRL** übersichtlich zusammenzufassen haben wir ein BPNM-Diagramm erstellt, s. Abbildung 10. Zunächst informieren sich die Datennehmer über PII, Bloomfilter und Matching-Algorithmen. Im nächsten Schritt werden die entsprechenden **Parameter festgelegt und an das System übergeben**. Daraufhin müssen die Datenhaltenden die nötigen Schritte für die Bereitstellung der PII unternehmen. In der Zusammenarbeit mit verschiedenen Datenhaltenden Stellen und Forschenden, ist dabei klar geworden, dass ein doppelter Ansatz für die Erstellung der PII benötigt wird. Deshalb bieten wir im Zielbild zwei Lösungen an: Die Erstellung der Bloomfilter in der Client-App, die auch die Anbindung an EuroDaT herstellt, sowie die Nutzung einer extra App die ausschließlich die Bloomfilter erzeugt. Im zweiten Fall muss die Anbindung an den Treuhänder getrennt erfolgen.

#### Die Bloomfilter-Erstellung – Client-App oder Bloomfilter-App

Zunächst möchten wir die Möglichkeit der Erstellung in der Client-App anbieten, so dass datenhaltende Stellen lediglich **eine App vorliegen** haben, **mittels welcher die komplette Datenbereitstellung erfolgen kann**.

In einigen Fällen kommt es jedoch vor, dass insbesondere die PII in einer geschützten Umgebung, beispielsweise einem Rechencenter liegen. Die Installation einer App mit Anbindung nach außen, wie es für die Client-App der Fall ist, ist hier oftmals nicht möglich. Um dennoch die Erstellung von Bloomfiltern zu ermöglichen, soll die **Logik in eine minimale App extrahiert** werden. Durch die Reduktion aufs Wesentliche kann diese somit auch **in besonders sicheren Umgebungen** genutzt werden. In diesem Fall sind die datenhaltenden Stellen selbstverantwortlich für die Übertragung der notwendigen Konfigurationen zur Erstellung der Bloomfilter in die Bloomfilter-App und die Überführung der erstellten Filter in die Client-App. Die Client-App führt, wie im oben beschriebenen ersten Fall, die Datenübertragung an den Datenraum aus, indem die **Bloomfilter für spätere Verknüpfungen und Auswertungen in EuroDaT genutzt** werden können. Wird hingegen die Client-App zur Erstellung der Bloomfilter genutzt, werden diese manuellen Zwischenschritte automatisch ausgeführt.

Vor der Erstellung der Bloomfilter ist im Zielbild zusätzlich ein **Pre-Processing** vorgesehen. Hiermit sollen kleine Unterschiede zwischen verschiedenen Datenquellen ausgeglichen werden. Beispielsweise können Umlaute einheitlich in ae/oe/ue geändert werden. Damit **erhöht sich die Qualität** der erstellten Filter und somit auch die Qualität des Matchings über verschiedene Datenquellen hinweg.

Im letzten Schritt werden die erstellten Bloomfilter für das Matching in EuroDaT an den Treuhänder übergeben.

## ERFOLGREICHE DATENVERKNÜPFUNG – PROOF OF CONCEPT

Im Folgenden beschreiben wir unsere erste erfolgreich durchgeführte Datenverknüpfung. Diese wurde in unserem Tech-Stack realisiert, s. Kapitel 06. Das Protokoll für die Durchführung einer Verknüpfung von zwei Studiendatensätzen, in unserem Beispiel Dummy-Daten von NAKO und NAPKON, hat folgende Schritte, mit dem Fokus auf die Datentransfers:

1. Der datennehmende User startet im Datenmodell-Frontend eine Transaktion zur Verknüpfung von **Studiendaten A** mit **Studiendaten B** und wählt dafür die existierende Analyse-Logik *PPRL-Identität-Matching* aus, s. Abbildung 11 (links). [EuroDaT](#) provisioniert daraufhin die Transaktions-DB.
2. Die Client-App bei den **Datengebenden A** übermittelt die **Studiendaten A** an den Datentreuhänder EuroDaT, s. Abbildung 11 (rechts). Inhalt sind die Spalten *StudienID* und *Wertespalten*.
3. Die Client-App bei der **Datengebenden A** übermittelt die **Bloomfilter-codierten PII-Daten A** an EuroDaT. Codiert sind die Spalten *Vorname*, *Nachname* und *Geburtsdatum*, nicht codiert ist die Spalte *StudienID*.
4. Wiederholung der Schritt 2 & 3 für **Datengebende B**, resultierend in nach EuroDaT übermittelten **Studiendaten B** und **Bloomfilter-codierten PII-Daten B**.
5. Manueller Schritt im MVP (s. Integration von Studiendaten): Nachdem die 4 Datensätze in der EuroDaT-Transaktions-DB abgelegt wurden, bestätigt der datennehmende User die Transaktion im Datenmodell-Frontend.
6. In EuroDaT wird automatisch das Python-Skript der ausgewählten Analyse-Logik *PPRL-Identität-Matching* gestartet:
  - a. Einlesen der vier Datensätze aus der Transaktions-DB in [pandas.DataFrame](#)
  - b. Identität-Matching: Full Outer Join auf den Bloomfilter-codierten Spalten für *Vorname*, *Nachname*, *Geburtsdatum*
  - c. Erstellen der verknüpften Tabelle bestehend aus: **Wertespalten-A**, *Wertespalten-B*
  - d. Schreiben des Ergebnisses in die EuroDaT-Transaktions-DB
7. Abholen der verknüpften Ergebniswerte aus EuroDaT durch das Datenmodell-Backend und Anzeige für den datennehmenden User im Datenmodell-Frontend.

Während wir uns im Rahmen des MVP auf das Identitäts-Matching eingeschränkt haben, ist die Erweiterung auf komplexere [Ähnlichkeit-Matching-Strategien](#) problemlos möglich: Alle Prozesse bleiben gleich, lediglich eine weitere Analyse-Logik wird im Datenmodell-Frontend zur Auswahl hinzugefügt (siehe Kapitel 06.03, Unterabschnitt Prozesse zur Integration der Verarbeitungslogik). Auch die Erweiterung auf

Treuhandstellen als separate Anlieferstellen für die Bloomfilter-codierten PII-Daten ist trivial, da lediglich weitere Client-Apps angebunden werden müssen.

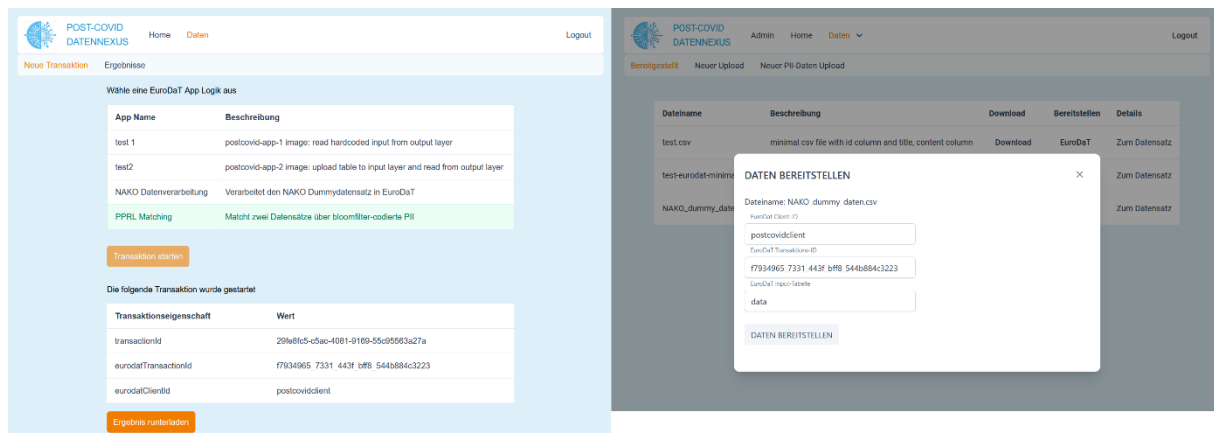


Abbildung 11: Screenshot des implementierten Webinterface für die Datennutzenden (links) und der Client-App (rechts). Der datennutzende User hat eine Analyse-Logik ausgewählt und eine Transaktion gestartet, und – als Teil eines manuellen Schritts im Rahmen des MVP (s. Integration von Studiendaten) – dem datengebenden User die Transaktionsdetails mitgeteilt. Nach Bereitstellung kann die Transaktion durchgeführt werden und der datennehmende User kann die verknüpften Ergebnisdaten runterladen.

## RE-ID-CHECK - PROZESS ZUM CHECK DER DATEN-RE-IDENTIFIZIERBARKEIT

Im Zielbild des PPRL haben wir im Rahmen des letzten Berichts ausgearbeitet, dass das Einlesen, das **Zusammenführen und die Verarbeitung von Daten im Datentreuhänder DSGVO-konform** sind, da hierbei keine personenbezogene Datenverarbeitung vorliegt, sondern lediglich eine Verarbeitung personenbezogener Daten (s. Kapitel 04.04 [Bericht Stufe 2](#)). Wobei es anzumerken gilt, dass es sich beim von uns umgesetzten PPRL-Verfahren zusätzlich um pseudonymisierte Daten in Form von Bloomfiltern handelt. Für Fragen bezüglich der Sicherheit des Treuhänders verweisen wir zusätzlich auf <https://www.eurodat.org/> und Juristen Zeitung, 79 S. 365-375 (2024).

### Re-ID-Check – Anonymität waren

Am Ende einer Datentransaktion im Treuhänder müssen die verknüpften Daten an die datennehmenden Stellen ausgeliefert werden. Sobald die Daten jedoch den Datentreuhänder verlassen und in die Hände der Datennehmer übergeben werden, handelt es sich um eine **personenbezogene Datenverarbeitung** und die DSGVO wird einschlägig. Da im Rahmen der Verknüpfung von verschiedenen Datensätzen Informationen einzelner Personen zusammengeführt werden, kann es sein, dass diese Kombination an unterschiedlichen Informationen eine natürliche Person identifizierbar machen. Man spricht in diesem Zusammenhang von Re-Identifikation. **Die Anonymität der Datensubjekte muss aber stets gewahrt bleiben**, weswegen eine Re-

Identifizierbarkeit ausgeschlossen werden muss. Deshalb stellen wir folgend unseren Re-ID-Check vor, mit dem wir prüfen, ob natürliche Personen mit den angefragten Daten re-identifiziert werden können.

Hierbei unterscheiden wir zwei Szenarien: Die **Ausgabe von verknüpften Daten** sowie die **Ausgabe aggregierter Analyseergebnisse**. Beide Szenarien unterscheiden sich in den nötigen Prozessschritten, um eine Re-Identifikation von Personen zu verhindern. Die Möglichkeiten eines Re-ID-Checks in beiden Szenarien werden folgend genauer beschrieben.

#### Ausgabe von verknüpften Daten

In diesem Szenario ist das Zielbild, dass zwei oder mehrere Datensätze miteinander verknüpft werden und anschließend der **verknüpfte Datensatz in Rohform an die Datennehmenden ausgegeben** wird. Dies setzt voraus, dass die Einverständniserklärungen der unterschiedlichen Datenquellen eine Verknüpfung erlauben. Liegt diese vor, darf theoretisch eine Verknüpfung durchgeführt werden.

Nichtsdestotrotz gilt es sicherzustellen, dass durch die Verknüpfung von Daten die **Anonymität der Datensubjekte gewahrt** bleibt. Es ist beispielsweise vorstellbar, dass durch die Auswahl an Variablen aus verschiedenen Datenquellen Informationen kombiniert werden, die zu einer einfachen Identifizierung einer natürlichen Person führt. Hierbei könnte es sich um Informationen zum Wohnort (PLZ) aus einer Tabelle und Informationen zu einer seltenen Erkrankung aus einer anderen Tabelle handeln. Je nach Einzelfall, könnte es passieren, dass nicht mehr von pseudonymen Daten gesprochen werden kann und eine Re-Identifizierung einzelner oder aller Personen im Datensatz einfach möglich ist.

- **Kritische Prüfung der ausgegebenen Variablen**

Ist das Ziel, verknüpfte Rohdaten an den Datennnehmer auszugeben, ist hierzu eine **Betrachtung der aus den einzelnen Datensätzen stammenden Variablen notwendig**. Hierbei muss eine neutrale Stelle abwägen, inwieweit die Verknüpfung der ausgewählten Informationen kritisch oder unkritisch ist. Eine Verknüpfung von Schuhgröße und Geschlecht bei großer Studiengröße (mehrere tausend Probanden) lässt kaum Rückschlüsse auf eine Einzelperson zu, während hingegen das obige Beispiel mit Wohnort und seltener Erkrankung bereits zu Problemen führen kann. Die **Einzelfallbetrachtung durch eine neutrale Stelle** kann hier Klarheit bieten. Problematisch ist jedoch, dass vor Auswertung der Daten oftmals keine Informationen über die Anzahl an Personen in verschiedenen Kategorien oder auch die Größe der Schnittmenge der zu verknüpfenden Datensätze vorhanden sind. Dadurch wird die Entscheidung über eine unkritische Zusammenführung erschwert.



- Grenzwerte für identifizierende Merkmale

Möchte man auf Nummer sicher gehen, ist es möglich **a priori Grenzwerte für die Anzahl an Patienten mit verschiedenen Merkmalen festzulegen**. Gibt es z.B. in jedem Ort min. 3 Personen mit einer seltenen Erkrankung ist die Chance der Identifizierung einer einzelnen Person gering, bei nur einer Person wäre die Zuordnung hingegen einfach möglich. Diese Grenzwerte können in Form eines Skriptes an EuroDaT übergeben werden. **Der Re-ID-Check führt dann eine Abfrage durch**, ob alle festgelegten Grenzwerte eingehalten sind und **ob die Daten ausgegeben werden dürfen**. Zeigt der Test ein Re-Identifikationsrisiko, werden den Datennehmenden keine Daten, sondern nur diese Information übermittelt. Ist der Test hingegen unauffällig und zeigt kein Re-Identifikationsrisiko, so erhält der Datennnehmer die verknüpften Daten.

- Mitbetrachtung der Bevölkerung

Zusätzlich prüfen wir auch weitergehende Re-Identifizierbarkeits-Checks, die nicht nur die Anzahl an Personen im Datensatz als Referenz betrachten, sondern auch die Anzahl der in der Bevölkerung vertretenen Häufungen. Für eine außenstehende Person ist die Identifizierbarkeit insbesondere davon abhängig, wie oft eine Kombination an Merkmalen in der Bevölkerung vorkommt und wie leicht oder schwer eine Zuordnung zu einer einzelnen Person dadurch ist. Man spricht hierbei auch von der „**k-Anonymität**“. Die grundlegende Idee hinter k-Anonymität ist, dass es schwierig sein sollte, eine einzelne Person basierend auf bestimmten Informationen (z.B. Alter, Geschlecht, Postleitzahl) zu identifizieren. Um k-Anonymität zu erreichen, muss jede Kombination von quasi-identifizierenden Merkmalen von mindestens k Individuen geteilt werden. Diese Methodik kann für den Entscheidungsprozess mit hinzugezogen werden, ist jedoch teils nicht trivial anwendbar, da ein großes Vorwissen über die Gesamtbevölkerung vorausgesetzt wird.

#### Ausgabe von Analyseergebnissen

Im Fall, dass Daten (unabhängig vom Grund) nicht verknüpft ausgegeben werden dürfen, besteht die **Möglichkeit, die Auswertung der Daten direkt in EuroDaT durchzuführen**, sodass **keine Rohdaten an die Datennehmenden überliefert** werden, sondern **lediglich Analyse-Ergebnisse**. Auch wenn diese Lösung zunächst als sichere Variante erscheint, kommt es auch hier auf die Details an. Da in diesem Fall der Sicherungsmechanismus eines fest vorgegebenen Output-Datenschemas in EuroDaT nicht unmittelbar genutzt werden kann, um die Ausgabe beliebiger Daten zu verhindern, muss der **Analyse-Algorithmus eingehender geprüft** werden. So könnten böswillige Datennehmende Auswertungsskripte so schreiben, dass sensible Informationen ausgegeben werden. Ohne jegliche Überprüfung wäre es ein leichtes, Code zu schreiben, der beispielsweise zeilenweise die Eingangsdaten zurückgibt, so dass am Ende die

Datennehmenden alle Informationen unter dem Deckmantel einer Datenanalyse erhalten. Solcherlei Missbrauch gilt es zu unterbinden.

Hierzu gibt es verschiedene Möglichkeiten und auch Entwicklungsschritte, die fortlaufend verbessert und erweitert werden können. Ein grundlegender Ansatz ist, die verwendeten Auswertungsalgorithmen vor ihrem Einsatz zu prüfen und nur freigegebene Algorithmen zuzulassen. Im Rahmen des MVP setzen wir zunächst eine solche Prüflogik und einen damit verbundenen Re-ID-Check um. Perspektivisch kann dies jedoch beliebig weiter ausgebaut werden.

- Sichere Auswertungen im MVP

Betrachten wir zunächst die Zielsetzung des Re-ID check im MVP. **Jede Art der Daten-Auswertung trägt ihre eigenen Risiken mit sich, einzelne Personen wieder identifizieren zu können.** Während es bei den Ergebnissen einer Regressionsanalyse nicht möglich ist, Rückschlüsse auf einzelne Personen zu treffen, kann die dazugehörige Punkt-Graphik bereits problematisch werden. So muss für jede Auswertung überlegt werden, ob ein Re-Identifikationsrisiko besteht und wie dieses überprüft werden kann.

Im Rahmen des MVP fokussieren wir uns auf **deskriptive Statistiken**, einen typischen Anwendungsfall in der medizinischen Forschung. Ziel ist es, Häufigkeitsverteilungen sowie Durchschnitt, Median, Minimum, Maximum und Quantile der vorliegenden Variablen zu berechnen. Zusätzlich soll es möglich sein, Kreuztabellen (Häufigkeitstabellen) für zwei beliebig ausgewählte Variablen zu erhalten. Bei allen Angaben zu Häufigkeiten wird im Vorhinein ein Grenzwert festgelegt. Alle Häufigkeiten, die unterhalb dieses Grenzwertes liegen werden vom System in der Ausgabe nicht ausgegeben. **Der Re-ID-Check im Treuhänder überprüft alle Zahlen und entfernt diese gegebenenfalls.** Im ersten Schritt der Umsetzung des Ansatzes werden alle Tabellen und Auswertungen mit Ergebnissen unterhalb des Grenzwertes nicht ausgegeben. Perspektivisch ist es möglich Zahlen zu schwärzen und um nicht mittels Kombinatorik auf die geschwärzte Anzahl an Patienten zu kommen, automatisiert noch eine weitere Ausgabekategorie zu schwärzen.

All diese Auswertungen sind über vorgefertigte Skripte, welche durch die Betreiber des Datenraums zur Verfügung gestellt oder auf Bedenkenlosigkeit geprüft werden, einfach ausführbar und zu komplexeren Auswertungen kombinierbar. **Eine Datenschutz-Überprüfung kann jeweils automatisiert durchgeführt werden.**

- Stetige Erweiterungen im Zielbild

Im finalen Zielbild kann diese **Liste erlaubter Auswertungsalgorithmen kontinuierlich erweitert** werden. Alle Auswertungen, die aus diesen zugelassenen Algorithmen

bestehen, werden automatisiert überprüft und können problemlos durchgeführt werden. Die Erweiterung des Katalogs um neue Algorithmen kann hierbei auf zwei Weisen erfolgen: Die aktive Erweiterung beispielsweise durch die Betreiber des Datenraums, oder durch das Hinzunehmen von Auswertungen, die im Rahmen einer Datennutzungsanfrage erstellt und durchgeführt wurden.

Jedoch sind Datenauswertung und Statistik ein sich stets weiterentwickelndes Gebiet, weshalb die wachsende Sammlung an Auswertungen voraussichtlich nie Vollständigkeit erreichen wird. Aus diesem Grund empfehlen wir die Möglichkeit der Überprüfung durch eine neutrale Stelle, die Auswertungen prüft, bevor diese im Treuhänder laufen dürfen. Hierdurch kann ausgeschlossen werden, dass bspw. einzelne Datenpunkte direkt ausgegeben werden. Ist eine abschließende Aussage über die Freigabe der Ergebnisdaten vor einer Auswertung nicht möglich, so ist eine zusätzliche Überprüfung durch eine neutrale Stelle vor Weitergabe der Ergebnisse an die Datennehmenden denkbar.

#### Prozess zum Re-ID Check

Im Laufe des Kapitels haben wir unterschiedliche Aspekte der Datenverknüpfung beleuchtet. Von der Erstellung von Bloomfiltern hin zur Verknüpfung der Daten und im letzten Schritt der Ausgabe der Daten unter Beachtung des Datenschutzes. Einige der hierbei beschriebenen Prozessschritte sind keineswegs trivial und erfordern zusätzlich individuelle Entscheidungen. Insbesondere im letzten Abschnitt, in dem es darum geht, die personenbezogenen Daten natürlicher Personen zu schützen, sind informierte Entscheidungen besonders wichtig. Um die nötigen Schritte übersichtlich und als Entscheidungshilfe vorliegen zu haben, haben wir den **Entscheidungsbaum** in Abbildung 12 erstellt.

Mittels mehrerer gezielter Abfragen, Überprüfungen und Entscheidungen wird **schrittweise entschieden, in welcher Form die Daten nach Verknüpfung in EuroDaT an die Datennehmer übergeben werden dürfen**. Insgesamt existieren drei Möglichkeiten, die [Ausgabe der Roh-Daten](#), die Ausgabe von gruppierten Daten oder die [Ausgabe von Analyse-Ergebnissen](#). Details zu den einzelnen Schritten und der daraus resultierenden Datenausgabe sind in Abbildung 12 zu finden.

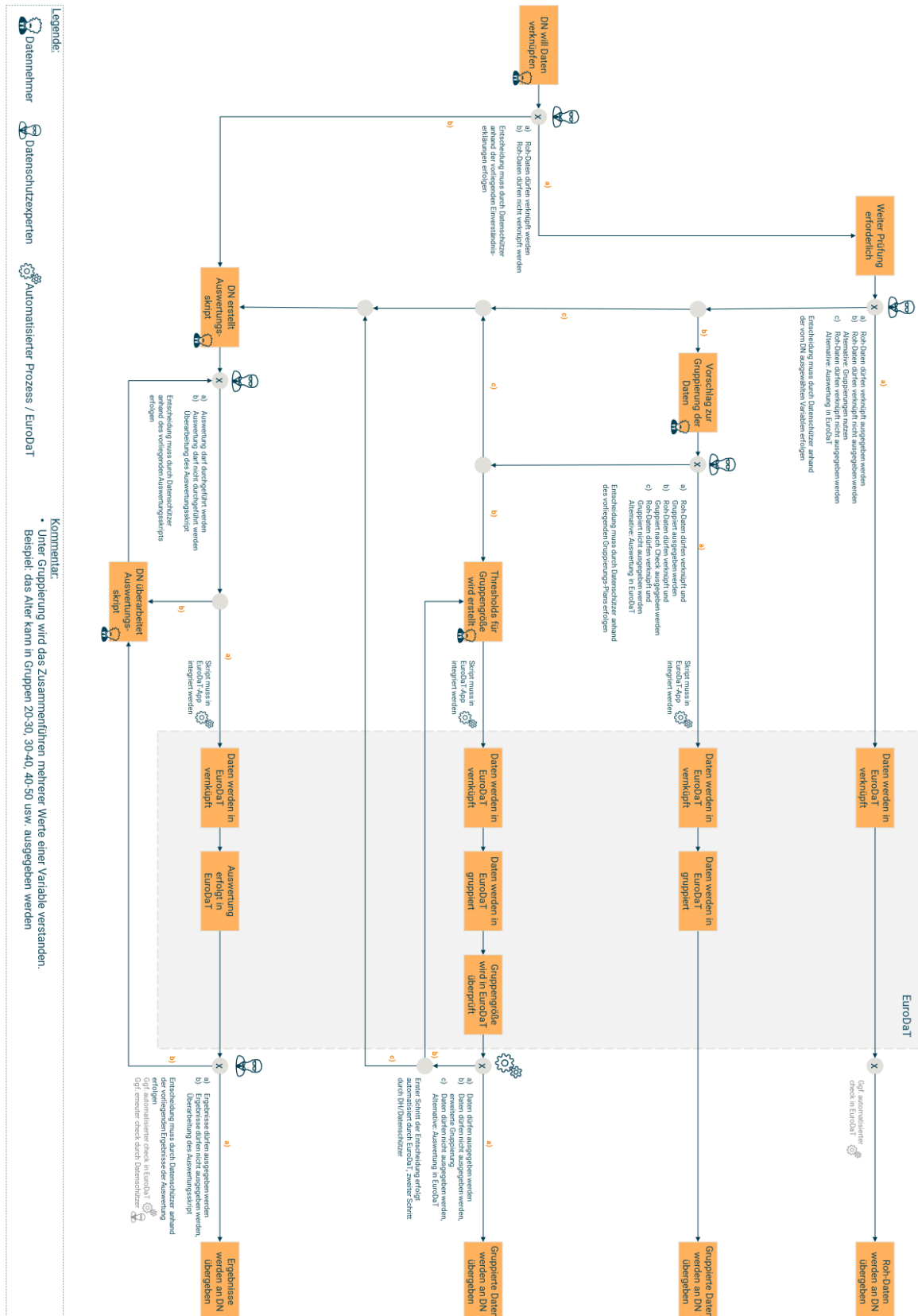


Abbildung 12: Prozessdiagramm zur Datenauswertung.

## 05.04 STRUKTURIERUNG DES DATENMODELLS (TYP DER DATENBANK)

In diesem Kapitel fassen wir das implementierte logische Datenmodell des Datenraums zusammen, aufbauend auf unserer Konzipierung, siehe [Bericht der Stufe 2, Kapitel 05.04](#). Der Fokus soll vorrangig auf den beteiligten **relationalen Datenbanken** und auf dem **Datenbank-Design** liegen, s. Tabelle 6 für einen Überblick über die implementierten Datenbanken.

DB-Name	Zugang & Berechtigungen	Beschreibung
Audit- & Management-DB des Datenraums	Nur technische Benutzer	Verwaltung der Metadaten des Datenkatalogs und der Transaktionen, sowie Monitoring
Transaktions-DB des Datentreuhänders	Nur die an der Transaktion teilnehmenden und autorisierten Datennutzenden und Datenhaltenden, nur während der Laufzeit der Transaktion, granulare Berechtigungen durch Row-Level-Security-Policies, s. <a href="#">Kapitel 06.01 im Bericht der Stufe 2</a>	Temporär existierende DB, um die Eingangsdaten aufzunehmen, zu analysieren und den Datennutzenden bereitzustellen
Lokale Client-DB der Datengebenden	Nur Datengebende	Lokale (on-premises) Ablagestelle für das Datenprodukt des Datengebenden. Das Datenprodukt kann sowohl ein pseudonymisierter Primärdatensatz als auch eine Pseudonym-Hash-Tabelle von PII-Daten sein

Tabelle 6: Überblick über die implementierten Datenbanken des Datenraums.

### AUDIT- UND MANAGEMENT-DB DES DATENRAUMS

Das **Entity-Relationship-Diagramm für die implementierten Hub-, Satelliten- und Link-Tabellen der Datenbank**, basierend auf dem [Data-Vault](#)-Ansatz, ist in Abbildung 13 zu sehen. Für die detaillierte Beschreibung der Aufgaben dieser zentralen Datenbank sei auf den [Bericht der Stufe 2, Kapitel 05.04](#) verwiesen.

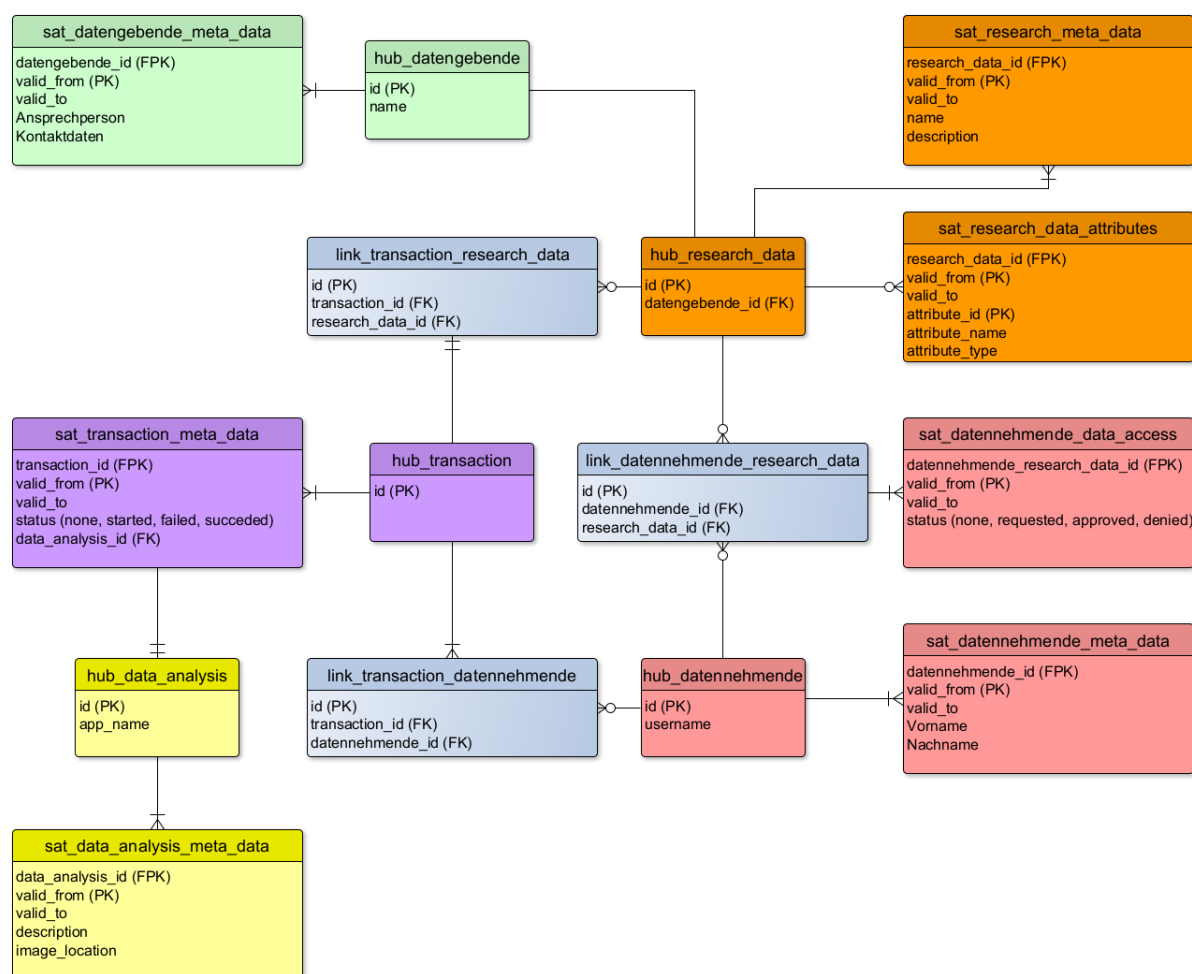


Abbildung 13: Entity-Relationship-Diagramm des implementierten Data Vaults der relationalen Datenbank des Datenraums für Audit und Metadatenmanagement. Farbliche Gruppierung in Datennehmende (rot), Datengebende (grün), Forschungsmetadaten (orange), Transaktionsmanagement (lila), Analysemanagement für Datendienstleister (gelb) und Links (grau).

## TRANSAKTIONS-DB DES DATENTREUHÄNDERS

Da der Datentreuhänder [EuroDaT](#) **transaktionsbasiert** arbeitet, wird nur für die typischerweise kurze Laufzeit einer Transaktion eine **temporäre Datenbank** provisioniert, die nach **Ende der Transaktion sofort gelöscht** wird, s. [Bericht der Stufe 2, Kapitel 05.04 und 06.01](#) für detaillierte Informationen zum logischen Datenmodell von EuroDaT. Die Struktur der Datenbank wird als Teil der EuroDaT App gespeichert.

## LOKALE CLIENT-DB DER DATENGEBENDEN

**Für die Datengebenden haben wir eine on-premises Client-App** – d.h. lokal vor Ort in der IT-Landschaft der Datengebenden integriert – mit lokaler Datenbank implementiert. Wir sind aktuell in Abstimmung mit der NAKO, NAPKON und der THS Greifswald hinsichtlich letzter Schritte zur Installation. Die lokale Datenbank dient zum einen der

**Entkopplung des Post-COVID-Datenraums von den besonders geschützten datenhaltenden Ressourcen der Datengeber, zum anderen als Einstiegspunkt zur Übermittlung der Datenprodukte in den Transaktionsprozess.** Die Datengebenden können über die Client-App die Inhalte, d.h. die Datenprodukte, der lokalen DB verwalten. Die Datenprodukte sind sowohl pseudonymisierte Primärdaten als auch codierte PII-Daten, letztere in Form von Pseudonym-Bloomfilter-Tabellen.

Da diese lokale Client-DB in die IT-Architektur der Datengebenden eingebettet ist, müssen ausreichende **Maßnahmen zur Gewährleistung der IT-Sicherheit** abgeleitet werden. Die **konzipierten technischen und organisatorische Maßnahmen (TOMs)** für die Integration der Client-App beschreiben wir in Kapitel 06.02.

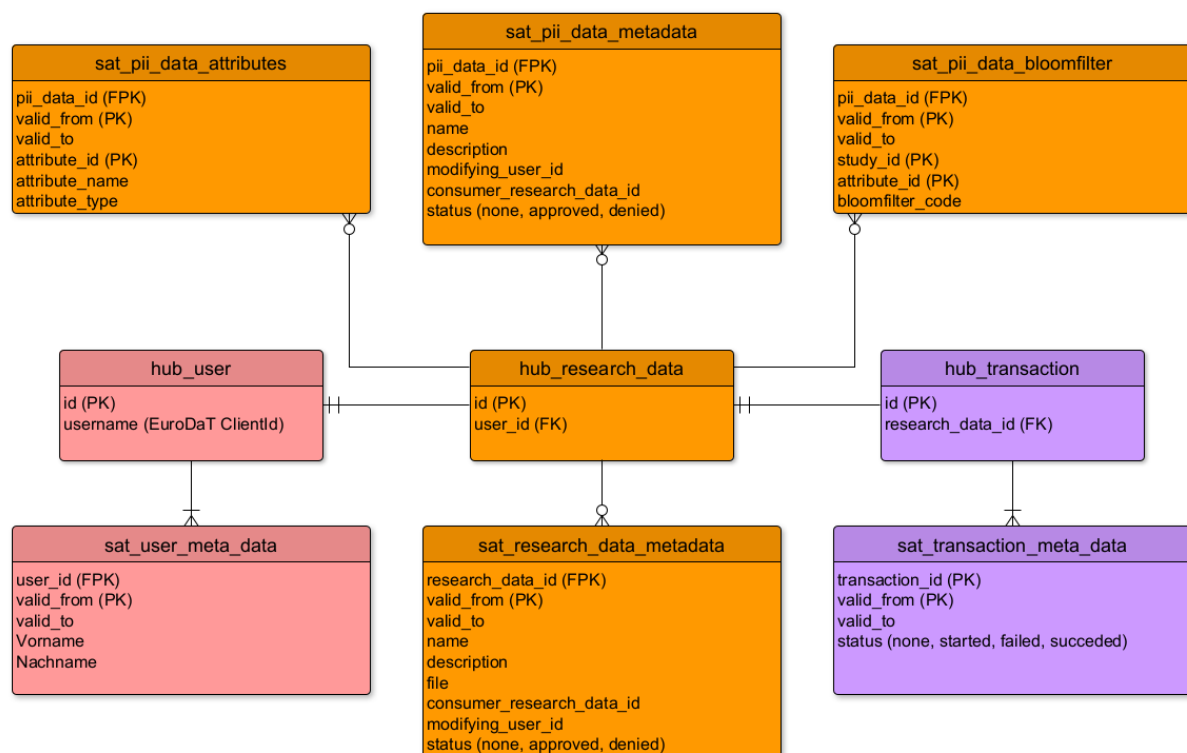


Abbildung 14: Entity-Relationship-Diagramm des implementierten Data Vaults der relationalen Datenbank der lokalen Client-App bei den Datengebenden. Farbliche Gruppierung in Client-Nutzer und Datennehmende (rot), Forschungsmetadaten und codierte PII-Daten (orange) und Transaktionsmanagement (lila).

## 06. PROZESSE UND ARCHITEKTUR

### 06.01 AUFBAU DER IT-INFRASTRUKTUR

Die IT-Infrastruktur für das Datenmodell ist die technische Realisierung der Anforderungen und Bedürfnisse von Datennehmenden und Datengebenden. Ausgehend



von unserer [IT-Architektur-Planung in Phase 2, Kapitel 06.01](#) haben wir während der Phase 3 den **Prototyp des Datenmodells ausimplementiert**. Der [Quellcode](#) für das gesamte Datenmodell wird unter der freizügigen [BSD-3-Clause Open-Source-Lizenz veröffentlicht und aktualisiert](#) werden. Abbildung 15 zeigt eine Übersicht über die Einbettung der IT-Infrastruktur in den Post-COVID-Datenraum. Die Hauptkomponenten des Datenmodells sind in Tabelle 7 aufgeführt. Für die ausführliche Darstellung des technischen und fachlichen Kontexts unserer Lösungsarchitektur sowie der Datenflüsse und Informationsobjekte verweisen wir auf den [Bericht der Phase 2, Kapitel 06.01](#).

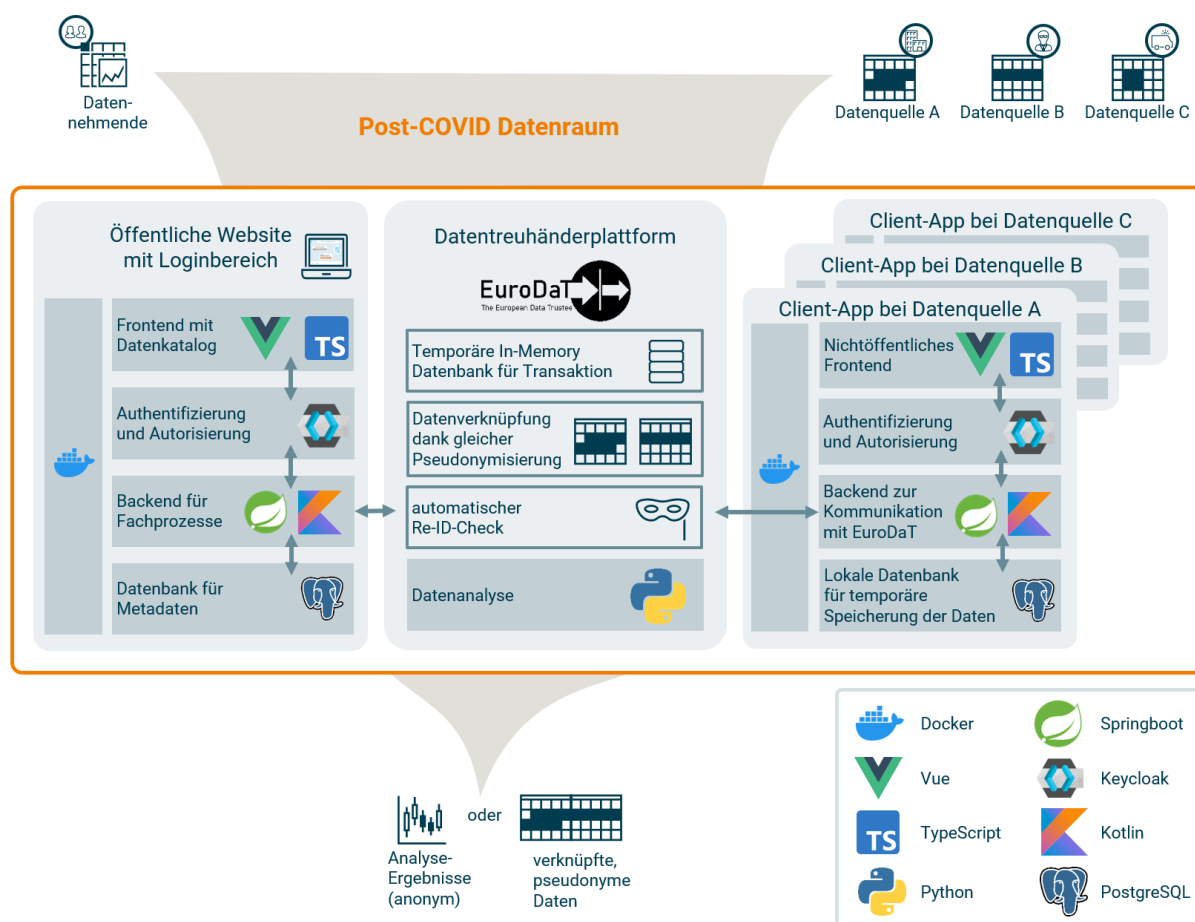


Abbildung 15: Zielbild der Einbettung der IT-Architektur in den Post-COVID-Datenraum. Im MVP ist die öffentliche Webseite noch bewusst entkoppelt von dem Login-Bereich der Datenmodell-App.

Komponente	Beschreibung	Open-Source-Technologien
Öffentliche Webseite	Öffentliche <a href="#">Website</a> des Datenraums, perspektivisch Login in die Datenmodell-App	TypeScript & Vue-Framework

Datenmodell-App	Applikation mit den Funktionalitäten Datenkatalog, Research-Hub mit Data-Science-Marktplatz für Analyse-Logiken, Datenverknüpfung, Login-Bereich für Datennutzende und Datenhaltenden, sowie perspektivisch U&A-Verfahren. Im Betrieb soll die Datenmodell-App per Login aus der öffentlichen Webseite erreichbar sein	TypeScript & Vue-Framework, Kotlin Spring Boot, PostgreSQL DB, Keycloak, Docker Container
Datentreuhänder-plattform <a href="#">EuroDaT</a>	Transaktionsbasierter Datenaustausch und Datenanalyse	Externer, komplementärer Service unter Open-Source-Lizenz, der unabhängig entwickelt und betrieben wird
Client-App	Lokale nicht-öffentliche Applikation auf Seite der Datenhaltenden, zur Aufnahme der Rohdaten (Primärdaten und Bloomfilter-codierte PII-Daten) für die Datennutzenden	TypeScript & Vue-Framework, Kotlin Spring Boot, PostgreSQL DB, Keycloak, Docker Container

Tabelle 7: Überblick über die wesentlichen Komponenten der IT-Architektur des Datenraums.

## 06.02 PLANUNG DER IT-INFRASTRUKTUR

In diesem Kapitel legen wir das Hauptaugenmerk auf die grundlegenden Konzepte zur Gewährleistung der IT-Sicherheit bei der Anbindung von datenhaltenden Stellen an das Post-COVID-Datenökosystem über die on-premises gehostete Client-Applikation. Insbesondere diskutieren wir detailliert die konzipierten technischen und organisatorischen Maßnahmen (TOMs) auf Seiten der Datengebenden bei der Integration der Client-Applikation, siehe Tabelle 8 bis Tabelle 13.

### MANAGEMENT DES PRODUKTLEBENSZYKLUS

Bereich	Technische Maßnahme	Organisatorische Maßnahme
Governance Entwicklung	Planung und Priorisierung anhand des JIRA-Boards	Geordneter Anforderungs- und Entwicklungsprozess
Vulnerability Detection	Einsatz <a href="#">OWASP-Scanner</a> und <a href="#">SonarQube</a>	Bewertung von Findings und Übergabe an Entwicklungsprozess bei Bedarf
Incident Management		Bereitstellung Kontakt seitens des Betreibers

Tabelle 8: TOMs für den Produktlebenszyklus.

## EINBETTUNG IN DIE DATENGEBER-IT-ARCHITEKTUR

Bereich	Technische Maßnahme	Organisatorische Maßnahme
Principle of Least Privilege	Einbettung der Client-Anwendung in ein VPN	
Principle of Least Privilege	Aufsetzen von geeigneten Firewall-Regeln für Inbound-/Outbound-Kommunikation, um IP- und API-Zugriffe zu beschränken	
Principle of Least Privilege	Keine Anbindung an die Forschungsdatenbank auf Client-Seite	Die Hoheit über die Datenweitergabe bleibt jederzeit bei der datenhaltenden Institution. Der bestehende Freigabeprozess der Datenhaltenden bleibt unangetastet, lediglich das resultierende Datenprodukt wird in die Client-App eingespielt

Tabelle 9: TOMs für die Einbettung in die Datengeber-IT-Architektur.

Zur Ausdetaillierung der obigen Maßnahmen befinden wir uns aktuell in enger Abstimmung mit der NAKO und dem [Rechenzentrum des DKFZ](#).

## VERTRAULICHKEIT

Bereich	Technische Maßnahme	Organisatorische Maßnahme
Zugangskontrolle	Zugang zur Client-App: Einstufige Authentisierung (Name, Passwort)	Generierung von Benutzerprofilen in Keycloak
Zugangskontrolle	Authentisierung bei EuroDaT über Client-Zertifikat: <ul style="list-style-type: none"> <li>Private-Key-JWT-Methode zur Client-Authentisierung</li> <li>TLS zur Verschlüsselung</li> </ul>	Registrierung von Client-ID und Hinterlegen des Client-Zertifikats in EuroDaT
Zugriffskontrolle	Protokollierung der Zugriffe in separater Append-Only DB	
Zugriffskontrolle	Über Row-Level-Security-Policies der lokalen DB wird die Zugriffsberechtigung auf Tabellendaten granular gesteuert	
Trennungskontrolle	Berechtigungskonzept (einfache und erweiterte Kompetenzen)	Vergabe der Rechte durch Betreiber

Tabelle 10: TOMs für die Sicherheitskomponente Vertraulichkeit.

## INTEGRITÄT

Bereich	Technische Maßnahme	Organisatorische Maßnahme
Weitergabe- kontrolle / Sicherung Datentransaktion	Verschlüsselung und elektronische Signatur während des Transports über TLS	
Eingabe/ Speicherkontrolle	Protokollierung über Historisierung der DB	Individuelle Benutzernamen
Eingabe/ Speicherkontrolle	Als Ausbaustufe sehen wir die Implementierung eines Berechtigungskonzepts vor	Vergabe der Rechte durch Betreiber
Speicherkontrolle	Als Ausbaustufe planen wir die Verschlüsselung der ruhenden Daten in der lokalen Client-DB	

Tabelle 11: TOMs für die Sicherheitskomponente Integrität.

## VERFÜGBARKEIT

Bereich	Technische Maßnahme	Organisatorische Maßnahme
Verfügbarkeits- kontrolle / Wiederherstell- barkeit	Einsatz von Docker	Neustart der Applikation bei Bedarf
Löschen von Daten	Möglich nach technisch durchlaufener Freigabe	Der Betreiber-Administrator besitzt erweiterte Rechte zum Löschen von Daten aus der lokalen Client-Datenbank

Tabelle 12: TOMs für die Sicherheitskomponente Verfügbarkeit.

## REGELMÄßIGE ÜBERPRÜFUNG UND BEWERTUNG

Bereich	Technische Maßnahme	Organisatorische Maßnahme
Datenschutz- Management	Siehe oben (v.a. Berechtigungskonzept, Zugriffskontrolle, Verschlüsselung)	Für den produktiven Betrieb soll eine Datenschutzfolgeabschätzung und ein Datenschutzkonzept erstellt werden

Tabelle 13: TOMs für die regelmäßige Überprüfung und Wartung.

## 06.03 PROZESSE ZUR DATENINTEGRATION

Die **Integration von Forschungsdaten** in den Datenraum soll **möglichst niederschwellig** funktionieren, von der initialen Informationssuche bis hin zur Freigabe durch die

Datenraum-Betreibenden, s. [fachlicher Gesamtprozess im Bericht der Stufe 2, Kapitel 06.03](#). In diesem Kapitel soll der Fokus auf der technischen Umsetzung der Prozesse zur Datenintegration liegen.

Die Integration von Datensätzen lässt sich in zwei Stränge unterteilen. Zum einen ermöglicht der Datenraum den Datensuchenden das **Auffinden von relevanten Datensätzen und Kollaborationspartnern**. Dieser Aspekt führen wir im Abschnitt Prozesse zur Erweiterung des Datenkatalogs aus und legen den Schwerpunkt dabei auf öffentliche Metadaten zu vertraulichen Datensätzen und Open Data.

Zum anderen benötigt der Datenraum eine **technische Anbindung der datengebenden Stellen**. Für diesen Zweck haben wir eine on-premises Client-App implementiert. Die implementierte Integration von sowohl Primärdaten als auch PII-Daten über die Client-App beschreiben wir im nachfolgenden Abschnitt Prozesse für Datengebende.

## PROZESSE ZUR ERWEITERUNG DES DATENKATALOGS

Integration von öffentlichen Metadaten, am Beispiel des Health Study Hub der NFDI4Health

Zusätzlich zur Aufnahme von Datensätzen einzelner Datengeber ist unser Datenraum auch darauf ausgelegt, **auf schon existierende externe öffentliche Metadatenquellen zuzugreifen und auf diese Weise die Datenauffindbarkeit und Vernetzung zu fördern**. Exemplarisch für öffentlich verfügbare Daten haben wir – in enger Abstimmung mit den Leitungs- und Entwicklungsteams des NFDI4Health – den [NFDI4Health Health Study Hub](#) als **etablierte Plattform für Studienmetadaten im deutschen Gesundheitssektor angebunden**. Die technische Integration beruht auf der Implementierung der [öffentlichen API](#) des Health Study Hub Resource-Servers in unserem Datenmodell-Backend.

Technische Integration von Open Data

Der **Datenraum-Tech-Stack ist versatil** genug, um **Metadaten von Open-Data im Datenkatalog zu integrieren** und die Datensätze an EuroDaT anzubinden. Im Rahmen einer Transaktion wird der aktuelle Stand aus dem jeweiligen externen Open-Data-Repository abgerufen. Technisch können wir diese Anbindung realisieren, indem die Datenmodell-Betreibenden selbst eine Client-App für Open-Data betreiben, welche als Proxy für nicht direkt angeschlossene Datensätze fungiert. Das Datenmodell-Backend befüllt mittels der Schnittstellenimplementierung des Open-Data-Repository diese Client-App mit dem aktuellen Datensatz. Daraufhin kann sie dann als datengebende Client-App am Transaktionsprozess teilnehmen und den Datensatz in den Datentreuhänder laden.

## PROZESSE FÜR DATENGEBENDE

### Installation der Client-App

Die Client-App ist Teil des Datenraum-Tech-Stacks und ihr Quellcode ist ebenfalls im [GitHub-Repository](#) unter der [BSD-3-Clause](#) Open-Source-Lizenz veröffentlicht. Aufgrund der Containerisierung des Client-Stacks **erfordert die Installation der Client-App eine minimale Standardkonfiguration** ([Docker Engine](#), [npm](#), sowie Freischaltung des Netzwerkzugriffs auf die EuroDaT-API) der IT-Umgebung der Datengebenden. Teil des Onboardings für die Inbetriebnahme sind die einmalige Registrierung bei dem Datentreuhänder [EuroDaT](#) und bei den Betreibenden des Datenraums.

Als erste datengegebende Stelle hat die NAKO zugesagt, unsere Client-App on-premises in der eigenen IT-Infrastruktur zu installieren und zunächst mit Dummy-Daten zu verproben. Für die Installation wird eine Virtuelle Maschine provisioniert, die vom [Rechenzentrum des DKFZ](#) betrieben wird. Zusätzlich befinden wir uns in Abstimmung mit NAPKON und der THS Greifswald.

Um den Einstieg in die Welt des Datenraums zu erleichtern und das Kennenlernen der **Prozesse der Client-App zu befördern, bieten wir allen interessierten datengebenden Stellen eine Cloud-Version der Client-App an**. Für den Betrieb empfehlen wir jedoch die on-premises Lösung der Client-App, damit die vollständige Souveränität der Datenhaltenden über ihre eigenen Daten gewahrt bleibt.

### Integration von Studiendaten

Zur Anbindung von Studiendaten besitzt die Client-App ein Frontend, mittels dem Datensätze in eine lokale Datenbank gespeichert werden können. **Berechtigungsmanagement, Versionierung und Verwalten von Metainformationen ist ebenfalls Teil der Lösung**. Diese lokale Datenbank dient zum einen der Entkopplung von der geschützten Forschungsdateninfrastruktur der Datengebenden, und zum anderen als Einstiegspunkt zur Datentransaktion mit dem Datentreuhänder [EuroDaT](#).

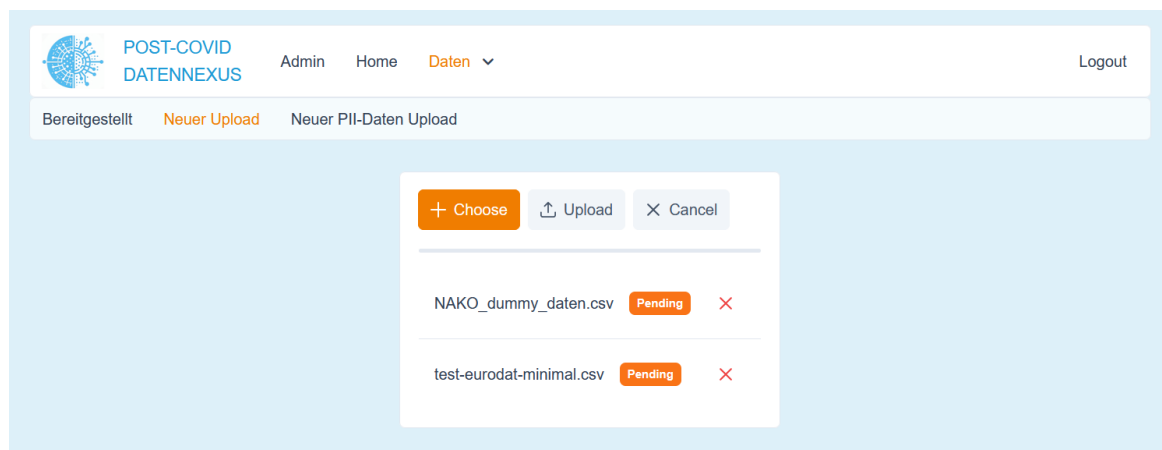


Abbildung 16: Eine Upload-Maske ermöglicht den Upload von Dateien in die Datenbank der Client-App. Ausgehend von der Übersicht aller bereitgestellter Dateien können diese in den Datentreuhänder [EuroDaT](#) geladen werden.

Um nun Daten für eine Transaktion zur Verarbeitung zur Verfügung zu stellen, kann aus der Client-App ein **Datensatz direkt in den Datentreuhänder [EuroDaT](#) hochgeladen** werden. Im Rahmen unserer MVP-Umsetzung erhalten die Datennutzenden beim Start einer Transaktion technische Informationen wie bspw. Transaktions-Id der Transaktion im Webinterface. Diese müssen dann den datenhaltenden Stellen mitgeteilt werden. Mit diesen Informationen kann die Client-App die aktuelle Version des Datensatzes in die temporäre Transaktionsdatenbank von [EuroDaT](#) laden. Dies kann man in Abbildung 11 sehen. Als Ausbaustufe planen wir, diesen momentan noch manuellen Schritt durch einen von [EuroDaT](#) vermittelten Kommunikationskanal zwischen Datenraum-Betreibenden und Client-App abzulösen: Dafür sehen wir die Implementierung der [EuroDaT-Messaging-Resource-API](#) vor, die den **Nachrichtenaustausch während der Transaktionslaufzeit automatisiert**. Das Berechtigungsmanagements des Datentreuhänders [EuroDaT](#) stellt dabei sicher, dass nur autorisierte Client-Apps in der Lage sind, der Transaktion Daten anzuliefern.

Nach erfolgreichem Upload können die Daten aus der Client-App gelöscht werden. Falls die Daten jedoch erneut angefragt werden, müssen sie wieder aus der ursprünglichen Quelle in die Client-App gebracht werden. Aufgrund der umfangreichen Sicherungsmaßnahmen **kann die Client-App aber auch zur Verwaltung der eigenen Datensätze verwendet werden**. Per Design löscht der Datentreuhänder [EuroDaT](#) nach erfolgreicher Transaktion immer die gesamte Verarbeitungsumgebung inklusive aller hochgeladenen Daten.

#### Integration von PII-Daten

Im Zentrum unseres Datenökosystems steht die **datenschutzkonforme, sichere Verknüpfung von Datensätzen unterschiedlicher Herkunft**. Für die Schlüsselfelder der



Verknüpfung, wie etwa sensitive personenbezogene PII-Daten, sehen wir eine Pseudonymisierungslogik mittels Bloomfiltern vor, s. Kapitel 05.03 zur ausführlichen Beschreibung der Verknüpfung von Datensätzen und [Bloomfiltern](#).

**Query 1: SELECT \* FROM sat\_pii\_data\_bloomfilter**

pii_data_id	valid_from	valid_to	study_id	attribute_id	bloomfilter_code
uid	timestamp with time zone	timestamp with time zone	text	uid	text
09cef46b-d56e-4d1a-b81c-5e97505dd046	2025-04-03 09:25:35.317017+00	null	1	e4b95eb1-303a-418f-9e2c-155739163eb3	000001000100000000010000010000010000100000000001100010
09cef46b-d56e-4d1a-b81c-5e97505dd046	2025-04-03 09:25:35.369383+00	null	1	b3fe2b04-03c7-4000-b525-0568eb3a038c	000000000100000000010100000000000001010010000000000000010
09cef46b-d56e-4d1a-b81c-5e97505dd046	2025-04-03 09:25:35.410594+00	null	1	c5525f5f-a88e-4b62-8104-4eef39a2d5b7	000000001000000001010000000100010100110110000000000000000000
09cef46b-d56e-4d1a-b81c-5e97505dd046	2025-04-03 09:25:35.449448+00	null	1	e53f1997-9ec1-46f4-b7a7-5d3cc1530679	0100100001001000

**Query 2: SELECT \* FROM "sat\_pii\_data\_attributes"**

pii_data_id	valid_from	valid_to	attribute_id	attribute_name	attribute_type
uid	timestamp with time zone	timestamp with time zone	uid	text	text
09cef46b-d56e-4d1a-b81c-5e97505dd046	2025-04-03 09:25:35.300819+00	null	e4b95eb1-303a-418f-9e2c-155739163eb3	Vorname	String
09cef46b-d56e-4d1a-b81c-5e97505dd046	2025-04-03 09:25:35.357118+00	null	b3fe2b04-03c7-4000-b525-0568eb3a038c	Nachname	String
09cef46b-d56e-4d1a-b81c-5e97505dd046	2025-04-03 09:25:35.398037+00	null	c5525f5f-a88e-4b62-8104-4eef39a2d5b7	Geburtsdatum	String
09cef46b-d56e-4d1a-b81c-5e97505dd046	2025-04-03 09:25:35.438912+00	null	e53f1997-9ec1-46f4-b7a7-5d3cc1530679	Geburtsort	String
09cef46b-d56e-4d1a-b81c-5e97505dd046	2025-04-03 09:25:35.469211+00	null	27f68098-138d-447a-a798-f64652373af3	Geschlecht	String
09cef46b-d56e-4d1a-b81c-5e97505dd046	2025-04-03 09:25:35.494676+00	null	926e0571-45d9-4a53-b8a6-c2d345cad78c	Wohnort	String

Abbildung 17: Ansicht der lokalen DB der Client-App: Ein PII-Datensatz wurde in die Client-App geladen und automatisch in einen Bloomfilter vor dem Speichern in der Datenbank codiert.

Die Client-App liest die Klartexte der **PII-Daten** ein, **transformiert sie mit einem Bloomfilter-Algorithmus und speichert die resultierenden Bloomfilter-Codierungen** in der lokalen Datenbank der Client-App. Die **Klartexte der PII-Daten werden hingegen nicht gespeichert und können daher auch nicht die IT-Infrastruktur der Datengebenden verlassen**. Als mögliche Ausbaustufe über den MVP hinaus kann die Bloomfilter-Codierungslogik in eine separate Anwendung ausgelagert werden bzw. in bestehende Codierungstools der Datengebenden integriert werden, um so bestehende Sicherheitsanforderungen zu erfüllen.

Im Rahmen der Datentransaktion schreibt die Client-App die erzeugten Bloomfilter-Codierungen in die temporäre Transaktionsdatenbank von [EuroDaT](#). Basierend auf Metriken für ein Ähnlichkeits-Matching vergleicht unser PPRL-Algorithmus in [EuroDaT](#) die erzeugten Bloomfilter-Codierungen der verschiedenen beteiligten Client-Apps und erzeugt eine **Zuordnung von gleichen bzw. ähnlichen Codierungen**. Die beteiligten **Apps können dabei sowohl von den Treuhandstellen als auch den Datenhaltenden betrieben werden**. Da die Pseudonym-Bloomfilter-Tabellen vorliegen (s. Kapitel 05.03) lässt sich die Verknüpfung verschiedener Primärdatensätzen über die PII-

Daten bewerkstelligen, ohne dass PII-Daten jemals außerhalb der geschützten Umgebung der Datengebenden gespeichert werden. Für den MVP der Client-App wurden für die ausgewählte Variante des Bloomfilter-Algorithmus die Konfiguration fest eingestellt, s. Tabelle 14.

Parameter	Ausprägung
q-gram Wahl	2-gram
Länge des Bit-Arrays	100
Bits per q-gram	2
Hash-Funktionen	<a href="#">SHA1</a> , <a href="#">MD5</a> , Anwendung der Double-Hashing-Strategie

Tabelle 14: Feste Konfiguration des Bloomfilter-Algorithmus für den MVP.

## PROZESSE ZUR INTEGRATION DER VERARBEITUNGSLOGIK

Um die Daten zu verarbeiten, stellt man EuroDaT die Analyselogik in Form eines Docker-Images zur Verfügung. Dazu haben wir ein Basis-Image entwickelt, das die Anbindung an die Transaktionsinfrastruktur übernimmt. Die eigentliche **Datenanalyse wird dabei auf ein einfaches Python-Skript reduziert**, da die hochgeladenen Daten als [Pandas](#)-DataFrame von dem Image bereitgestellt werden. Da Pandas als Standardbibliothek für die Analyse von Datensätzen in Python weit verbreitet ist, **können Forschende ihre Datenauswertung mit vertrauten Werkzeugen durchführen**.

Um Verarbeitungslogik und die eingespeisten Daten aufeinander abzustimmen, sehen wir Konfigurationsmechaniken vor, damit Datennutzende die Auswertung und Anpassung bei Änderungen im Datensatz vornehmen können.

Eine **Historisierung der erfolgten Transaktionen** ermöglicht die Wiederverwendung der angewandten Logik zur Untersuchung ähnlicher Fragestellungen. Im Zielbild entsteht ein **Katalog von Verarbeitungslogiken**, aus dem Nutzende passende Komponenten auswählen können. Beispiele hierfür sind Implementierungen von Re-ID-Checks (siehe 05.03) oder Bloomfilter-Verknüpfungen. Zusammen mit Metriken zu bereits erfolgten Transaktionen, die etwa die Güte eines Bloomfilter-Matchings beschreiben, können Datennutzende **von den Erfahrungen anderer profitieren**.

Ein solcher Katalog eröffnet zudem die Möglichkeit, am Datenraum als Anbieter von Verarbeitungslogiken teilzunehmen. Dies umfasst sowohl das Bereitstellen von Docker-Images für [EuroDaT](#) als auch die Vernetzung von Datennutzenden mit Experten für Datenanalysen, die komplexere Auswertungen durchführen können.

## 06.04 PROZESSE ZUR DATENAKTUALISIERUNG

Ein wichtiger Aspekt für die dauerhafte Nutzung des Datenraums ist die **Sicherstellung der Aktualität der enthaltenen Daten**. Nur so können kann das Vertrauen der Nutzenden in die Daten und den Datenraum langfristig gesichert werden. In den vergangenen Stufen haben wir deshalb ein ausführliches Konzept zur Aktualisierung bereits angebundener Datensätze erstellt. Dies ist in Kapitel 06.04 des Berichts der [Stufe 2](#) sowie im Kapitel 05.03 des Berichts der [Stufe 1](#) zu finden.

Unser Fokus lag hierbei auf der **Aktualisierung der Primärdaten, der Transaktionsdaten, der Metadaten, sowie der Datenmodellpflege** selbst. Der zentrale Aspekt unseres technischen Ansatzes ist dabei der Einsatz eines transaktionsbasierten Datentreuhänders. Im Rahmen einer Datentransaktion werden immer die jeweils aktuellen Datenbestände der Datengebenden abgerufen, wodurch wir die **Pflege der Primärdaten automatisieren und garantieren**, dass die Datennutzenden jederzeit Zugriff auf die aktuellste verfügbare Datengrundlage haben. Insbesondere die in unserem Datenkatalog eingepflegten Datensätze von NAPKON und NAKO werden so aktuell gehalten. Die Mechanismen zur Aktualisierung der von uns angebotenen Open Data Datensätze beschreiben wir in Kapitel 05.01.

## 07. UMSETZBARKEIT DER ENTWICKLUNG SEKTORUNABHÄNGIGER DATENMODELLE

Im Folgenden stellen wir unseren Ansatz zur Entwicklung eines Datenmodells vor. Dabei stellen wir sowohl unsere Entwicklungsarbeit am Post-Covid Datenmodell vor als auch eine Generalisierung zu einer **Entwicklungsblaupause für die Entwicklung von Datenmodellen in beliebigen Sektoren**. Die Übergänge sind dabei aufgrund unseres dynamischen Entwicklungsansatzes teilweise fließend.

Wir stellen hierbei drei Aspekte vor: Zunächst erläutern wir **unseren Strukturierungsansatz**, den wir als Matrixmodell in Scope und Zeit entwickelt und mit dem wir unsere Arbeit übersichtlich geordnet haben. Als nächstes stellen wir unser konkretes **Vorgehensmodell** dar. Final stellen wir unser **Qualitätskonzept** vor, mit dem wir sichergestellt haben, dass das Datenmodell den Ansprüchen aller Stakeholder genügt.

## 07.01 DAS MATRIXMODELL

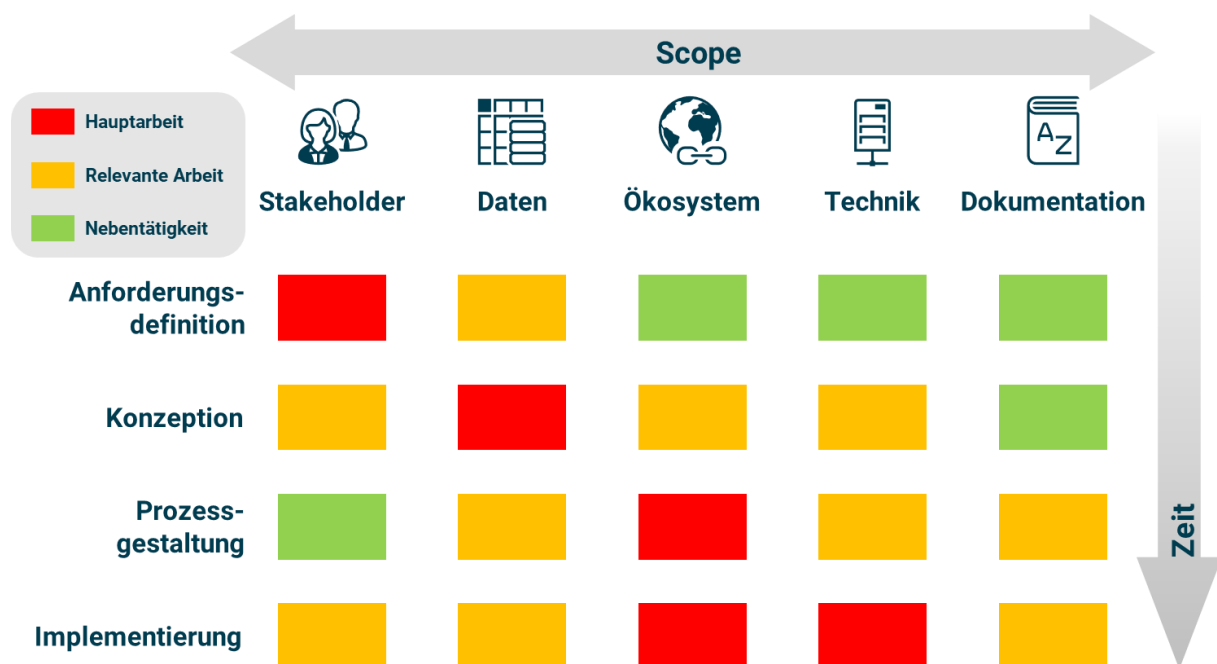


Abbildung 18: Visualisierung des Matrixmodells zur Strukturierung unserer Entwicklungsarbeit in Zeit und Scope.

Grundsätzlich hat die Entwicklung eines Datenmodells ein klar definiertes Ziel sowie eine begrenzte Laufzeit und kann somit als ein **Entwicklungsprojekt** verstanden werden, auf das sich die umfangreiche Methodik des Projektmanagements anwenden lässt. Projekte werden typischerweise in den **drei Dimensionen Kosten, Zeit und Scope** gesteuert. Da die Kosten der Entwicklungsarbeit extern vorgegeben sind, haben wir unser Vorgehen in den verbleibenden Dimensionen Zeit und Scope strukturiert und hierfür ein **übersichtliches Matrixmodell entwickelt**, s. Abbildung 18.

### Scope – Ziele und Anforderungen des Projekts

Die erste Strukturdimension **Scope** umfasst hierbei **die wesentlichen Ziele und Anforderungen des Projekts**, und ist entsprechend der Klassifizierung unserer Gesamtarbeiten, s. Kapitel 02.01, unterteilt in Stakeholder, Datensätze, Datenökosystem, Technik und Dokumentation. Diese Struktur hilft, die **Komplexität zu reduzieren und die Übersichtlichkeit zu wahren**. Stakeholder beziehen sich auf die Einbindung relevanter Akteure und die Berücksichtigung ihrer Perspektiven und Anforderungen. Datenquellen betreffen die Auswahl und Nutzung von Primärdatensätzen. Das Datenökosystem beschreibt die Vereinheitlichung und Verknüpfung verschiedener Datenquellen. Technik beschreibt die IT-Infrastruktur und Sicherheitsstandards und die

Dokumentation beschreibt, wie die Projektergebnisse hochqualitativ und nachvollziehbar festgehalten werden.

### Zeit – Die Ablaufphasen des Projekts

In der zweiten Strukturdimension **Zeit** legen wir **verschiedene Ablaufphasen des Projekts** fest. Wir haben dabei die Stufen der Post-Covid Challenge, Konzeption, Vorbereitung der Infrastruktur und Aufnahme von Daten, kritisch gewürdigt und als Grundlage unserer Strukturierung berücksichtigt. Gleichzeitig sehen wir einen Mehrwert darin, vor der Konzeption noch eine separate Phase der Anforderungsdefinition vorzusehen, da dieser Schritt wesentlich für die Ausrichtung der Entwicklungsarbeit, aber nicht zwingend in der Konzeption enthalten ist.

- **Anforderungsdefinition**

Gemäß unseres Modells konzentriert sich diese Phase darauf, die Marktbedürfnisse zu bestimmen, rechtliche und fachliche Expertise einzubinden, Use Cases zu definieren, festzulegen, welche Datenprodukte entwickelt werden sollen und allgemein ein konkretes Zielbild des Mehrwerts des Datenmodells zu entwickeln.

- **Konzeptions-Phase**

In der Konzeptions-Phase werden detaillierte Konzepte ausgearbeitet, die beschreiben, wie die Datenmodellierung z.B. in Form relevanter Datendimensionen oder Ontologien die vorab formulierten Forschungsfragen adäquat unterstützen kann und wie die identifizierten Datensätze harmonisiert werden können.

- **Prozessgestaltung**

In der dritten Phase Prozessgestaltung werden operative Abläufe optimiert, Prozesse einheitlich modelliert und visualisiert (z. B. als BPMN-Diagramme), um eine gemeinsame Verständigungsgrundlage zu schaffen. Regulatorische Vorgaben werden integriert, und die Datenmodelle verschiedener Quellen verzahnt, um eine konsistente und rechtssichere Datenstruktur zu gewährleisten.

- **Implementierung**

Die finale Implementierungsphase umfasst schließlich die praktische Umsetzung der entwickelten Konzepte in Software, Hardware sowie die Integration dieser Elemente. Da die Implementierung stark von den Anforderungen des konkreten Datenmodells abhängt, verweisen auf Kapitel 06 für einen Überblick unserer Arbeiten.

### Matrixmodell – Korrelation von Scope und Zeit

Diese Unterkategorien sind jeweils übersichtlich und trotzdem umfassend genug, um die gesamte **Komplexität der Entwicklung abzubilden**. Das Matrixmodell erlaubt nun, die **Projektdimensionen miteinander zu korrelieren**, indem die Arbeit der Projektphasen in der Zeit schwerpunktmäßig einzelnen Handlungsfeldern des Scopes zugeordnet

werden. Diese Zuordnung ist an einem klassischen Projektplan als orientiert, lässt aber gleichzeitig mehr **Freiräume bei der Schwerpunktsetzung und Ausgestaltung der Entwicklungsarbeiten** für Datenmodelle in beliebigen Sektoren. Die in Abbildung 18 konkret dargestellte Zuordnung hat sich für die Post-Covid Challenge als sinnvoll erwiesen. Für die Entwicklung anderer Datenmodelle kann es sinnvoll sein, das Mapping anzupassen, wofür das Matrixmodell eine einfache Strukturierungshilfe bietet.

## 07.02 DAS VORGEHENSMODELL

Im diesem Kapitel erläutern wir, wie die durch das Matrixmodell vorgegebene Projektstruktur effektiv in Arbeitsprozesse übersetzt werden kann. Ausgangspunkt ist dabei wieder unsere Arbeit an der Post-Covid Challenge.

Grundsätzlich hat es sich als vorteilhaft erwiesen, ein **iteratives Vorgehensmodell** anzuwenden, da **das Datenmodell so anhand von Feedback und neuen Erkenntnissen kontinuierlich an sich ändernde Anforderungen und gesetzliche sowie technologische Veränderungen angepasst** werden kann. Das so entwickelte offene Datenmodell ist **flexibel, nachnutzbar** sowie **anpassungsfähig** für verschiedenste Anwendungsfälle. Ein iteratives Vorgehen ermöglicht es darüber hinaus, über die schrittweise Entwicklung das Datenmodell dynamisch an neue Rahmenbedingungen und Erkenntnisse aus der Projektarbeit anzupassen, was im sich aktuell weiterentwickelnden Umfeld der Datenökonomie von Vorteil ist. In einer späteren, reiferen Phase der deutschen Datenökonomie lassen sich andere Vorgehensmodelle denken, auf die wir hier aufgrund der aktuellen Marktlage aber nicht eingehen.

Unabhängig vom genauen Vorgehensmodell sollten bei der Entwicklung eines neuen Datenmodells, unserer Erfahrung nach, folgende Aspekte berücksichtigt werden:

### 1. Konzeptionelle Modellierung

Zu Beginn sollte ein konzeptionelles Datenmodell erstellt werden, das unabhängig von spezifischen Technologien ist und einen Überblick über Entitäten und deren Beziehungen bietet.

### 2. Inkrementeller Aufbau

Das Datenmodell sollte schrittweise erweitert werden, beginnend mit einem Kernsystem und sukzessiver Ergänzung weiterer Funktionen.

### 3. Dokumentation

Eine gründliche Dokumentation jeder Iteration ist wichtig, um die Nachnutzbarkeit zu gewährleisten.

### 4. Standardisierung

Es sollten standardisierten Methoden wie Entity-Relationship-Modellen (ERM) oder

unified modeling language (UML) verwendet werden, um die Interoperabilität zu fördern.

## 5. Entwurfsmuster

Nutzung von bewährten Entwurfsmustern für häufig auftretende Strukturen wie Historisierung oder Mehrsprachigkeit.

Konkret haben wir für die Umsetzung dieser Prinzipien in unserer Projektarbeit ein iteratives Vorgehensmodell genutzt und kontinuierlich verfeinert, welches wir entsprechend zur Nachnutzung empfehlen.

Das gesamte Vorgehen war durch die Implementierung von **zweiwöchigen Sprints** gekennzeichnet. Diese kurzen Entwicklungszyklen ermöglichen eine schnelle Reaktion auf neue Erkenntnisse und Herausforderungen sowie eine **kontinuierliche Anpassung und Verbesserung der Projektarbeit**. Zur agilen Planung der Projektarbeit haben wir ein **Kanban Board** und das Tool **Jira** genutzt. Diese visuelle Planungshilfen ermöglichten es, Aufgaben zu organisieren und den **Fortschritt transparent zu gestalten**, wodurch alle Teammitglieder den aktuellen Stand des Projekts jederzeit einsehen können.

Zur persönlichen Abstimmung haben wir **regelmäßige Jour Fixes** mit den Fachgruppen des Konsortiums eingerichtet, in unserem Fall mit den Säulen Recht und Medizin. Diese Gremien dienen dazu, spezifische Fachfragen zu diskutieren und sicherzustellen, dass das Projekt in diesen zentralen Bereichen fundiert bleibt. Außerdem sorgten wöchentliche **Austauschformate zwischen den technischen und fachlichen Entwicklern** für eine enge **Verzahnung der Arbeiten**. Während sich die fachliche Entwicklung auf inhaltliche Themen und die Umsetzung von Anforderungen fokussierte, umfasste die technische Entwicklung die Implementierung und Optimierung der technischen Infrastruktur und Systeme. Eine enge Abstimmung zwischen beiden Arbeitssträngen hat sich als essenziell erwiesen, um effektiv auf die Projektziele hinzuarbeiten.

In all diesen Aspekten hatten die Entwicklerinnen und Entwicklern weitgehende **Autonomie** bei der Auswahl ihrer Arbeitspakete und Bearbeitungsmethoden. Diese Freiheit hat durchgängig die **Kreativität, Eigenverantwortung und Motivation gefördert**, was zu einer flexiblen und effektiven Problemlösung und Verantwortungsbewusstsein für die entwickelte Lösung geführt hat.

## 07.03 DIE QUALITÄTSSICHERUNG

Um eine **einheitliche hohe Qualität und Nutzbarkeit** der Arbeitsergebnisse zu garantieren, haben wir eine **Qualitätskontrolle etabliert**, die die Stakeholder-Perspektive als zentralen Bestandteil des Wertversprechens unserer Arbeit ins Zentrum stellt. Im



Folgenden beschreiben wir, unsere Arbeiten, s. Kapitel 02.01, und deren Übertragbarkeit in die fünf oben spezifizierten Handlungsfelder.

#### Stakeholder – Absicherung des Wertversprechens

Um unsere Arbeiten mit den Stakeholdern kontinuierlich auf ihre Qualität zu prüfen, haben wir verschiedene qualitätsgesicherte Maßnahmen entwickelt. Grundsätzlich hat sich als zentrale Strategie unserer Arbeit die **Nachnutzung bestehender Strukturen** bewährt, die durch ihre Akzeptanz im Markt bereits als qualitätsgesichert wahrgenommen werden. Dies ist insbesondere in stark regulierten Sektoren wie der Medizin von Vorteil. Bereits im Markt etablierte Ansätze bieten hier nämlich einen Vertrauensvorschuss, den neu eingeführte Lösungen nur schwer aufholen können, vgl. auch Kapitel 04.04 und 05.01. Des Weiteren ist ein **kontinuierlicher Informationsaustausch** wichtig, um alle Stakeholder regelmäßig über Fortschritte und Entwicklungen zu informieren und ihr **Feedback und Anforderungen in den Entwicklungsprozess zu integrieren**. Wir haben hierfür Review-Meetings eingerichtet, z.B. bei der sektorübergreifenden Entwicklung der Datentaxonomie, s. Kapitel 04.04.

#### Datensätze – Hochqualitative Daten für hochqualitative Forschung

Die Qualitätssicherung angebundener Datensätze und ihrer Datenmodelle ist in stark regulierten Sektoren eine Herausforderung, da man weitgehend keinen Einblick in die Daten nehmen darf. Als Lösung empfehlen wir zunächst nur Datenquellen von Einrichtungen anzubinden, die intern eine hohe Datenqualität sicherstellen. Weitere Maßnahmen, die Datenquellen sektorunabhängig qualitätssichern können sind z.B. **semantische und inhaltliche Checks**, bei denen geprüft wird, ob bereitgestellte Daten und Metadaten den vorgegebenen Standards entsprechen und innerhalb fachlich definierter, sinnvoll erwartbarer Grenzen liegen. Diese Maßnahmen können als Bestandteil eines **allgemeinen Datenmonitorings** insbesondere in weniger regulierten Sektoren die Datenqualität sichern.

#### Technische Lösung – Sicherheit und Zuverlässigkeit garantieren

Zur Qualitätssicherung der entwickelten technischen Lösung kann man zahllose etablierte Methoden der Softwareentwicklung einsetzen, von denen wir im Folgenden die von uns genutzten vorstellen. Da diese keinen fachspezifischen Bezug zum Gesundheitssektor haben, lassen sie sich **unmittelbar als Blaupause auf die sektorunabhängige Entwicklung beliebiger Datenmodelle übertragen**.

In unserer Arbeit hat sich ein **strukturierter Anforderungs- und Entwicklungsprozess** bewährt, den wir zur Nachnutzung empfehlen. Konkret haben wir die Planung und Priorisierung der Entwicklungsarbeit mithilfe eines **JIRA-Boards** koordiniert, was Aufgaben und Zuständigkeiten transparent ordnet. Außerdem steigern regelmäßige

**Code-Reviews** die Qualität und Zuverlässigkeit des entwickelten Codes. Mit der Implementierung von **Integrationstests** und **Unit-Tests** ist darüber hinaus sichergestellt, dass alle Softwarekomponenten wie erwartet funktionieren und miteinander harmonieren. Zur Koordinierung dieser Maßnahmen haben wir **eine CI-Pipeline** eingerichtet, die unter anderem die **Code-Qualität automatisiert überprüft**. Insbesondere für Datenmodellierungen, die größere Codeentwicklungen erfordert, empfehlen wir eine solche Pipeline. Für die notwendige Identifikation und Schließung von Sicherheitslücken eignen sich ebenso **automatisierte Code-Analysen**, für die z.B. die Tools SonarQube und OWASP Dependency Check empfehlen können.

In Summe haben wir in diesem Kapitel ausgeführt, wie sich die Entwicklung eines Datenmodells sektorunabhängig umsetzen lässt. Wir haben dabei auf den Erfahrungen aus unserer Arbeit an der Post-Covid Challenge aufgebaut und unser Matrixmodell als Strukturierungsansatz, unser iteratives Vorgehensmodell und die von uns etablierten Qualitätssicherungsmechanismen beschrieben.

## 08. BETRIEB UND NACHNUTZUNG DES DATENMODELLS

Der vorliegende Bericht führt zu einem Abschluss unserer Arbeiten an der Post-Covid Challenge hin. Ein zentrales Ziel der Challenge war es hierbei, **als Pilot Use Case Verbesserungspotentiale für eine sektorübergreifende und gemeinwohlorientierte Datennutzung in Deutschland zu identifizieren**. Dieses Ziel kann mit Abschluss der Challenge nicht final erreicht werden, da die deutsche Datenökonomie zum einen erst im Aufbau ist, und sich zum anderen auch in Zukunft dynamisch weiterentwickeln muss, um jederzeit auf die sich ändernden Anforderungen des Marktes und der Nutzenden eingehen und so mehrwertstiftende Angebote entwickeln zu können.

Aufgrund dieser Notwendigkeit haben wir einen wesentlichen Teil unserer Überlegungen auf die **Fortführung und Nutzbarkeit unserer Arbeit nach dem Ende der Challenge** ausgerichtet. Hierbei unterscheiden wir zwischen dem Betrieb des von uns entwickelten Datenökosystems, worunter wir operative Aufgaben wie Finanzierung laufender Kosten und Bereitstellung notwendiger Ressourcen verstehen, und der Nachnutzung durch Datengebende und -nehmende, worunter wir sowohl die Nutzung des Datenökosystems als Forschungsdatenplattform als auch die Verwendung der von uns entwickelten Prozesse und technischen Komponenten in der Entwicklung neuer Angebote verstehen. Wir gehen im Folgenden auf beide Aspekte getrennt ein.

### DAS BETRIEBSMODELL

Im Betrieb eines Datenökosystems müssen die **Finanzierung laufender Ausgaben und personelle sowie technischen Ressourcen** bereitgestellt werden. Die sichert z.B. die

Infrastruktur für das Hosting, Sicherheitsupdates und Weiterentwicklungen des Datenmodells, Anpassung der Datenmodellierung an neue Use Cases und das Onboarding neuer Datengebender und Datensätze.

Die Finanzierung ist dabei die zentrale Herausforderung, da die Bereitstellung technischer Werkzeuge im Gegensatz zur allgemeinen Forschungsförderung keine hoheitliche Aufgabe darstellt und somit Betriebsressourcen auch extern beschafft werden können. Für die Finanzierung haben wir im [Abschlussbericht der Stufe 1](#) verschiedene Betriebsmodelle skizziert, die wir im Verlauf der Challenge weiter ausgearbeitet haben und hier kurz zusammenfassen. Aufgrund der Begebenheiten sind **für die Zeit nach Abschluss der Challenge noch viele Fragen ungeklärt**, weswegen wir an dieser Stelle keine finale Empfehlung aussprechen, sondern lediglich folgende drei, **grundsätzlich alternative Betriebsansätze** beleuchten: Eine öffentliche, eine kommerzielle private und eine gemeinnützig private Trägerschaft.

#### Öffentlicher Träger mit Staatsauftrag

Unter dem Betrieb durch einen öffentlichen Träger verstehen wir die **Verwaltung und Durchführung von Aufgaben im Zusammenhang mit dem Datenökosystem durch eine staatliche Einrichtung oder Behörde**, welche eine vertrauenswürdige und neutrale Handhabung sicherstellt. Ein Staatsauftrag bezeichnet hierbei eine förmliche Verpflichtung der Regierung, die entweder aus öffentlichen Mitteln oder aus einem Gebührenmodell ohne Gewinnerzielungsabsicht finanziert wird, um spezifische Projekte oder Ziele zu unterstützen, wie z.B. eine solche datenbezogene Initiative von öffentlicher Bedeutung. Eine solche Struktur könnte einerseits eine direkte Förderung und Unterstützung durch die Regierung ermöglichen und andererseits die Integration digitaler und datenschutztechnischer Expertise innerhalb der Regierung stärken.

Der Betrieb des Datenökosystems könnte auf verschiedene Weisen realisiert werden, wobei deren Priorisierung unter anderem auch von den Ergebnissen der aktuell laufenden Regierungsbildung abhängt. Eine Möglichkeit wäre die Einrichtung des **Dateninstituts** als eigene Behörde oder als Abteilung einer bestehenden Bundeseinrichtung. Alternativ könnte das Datenökosystem einer neuen Regierungseinheit als Bestandteil zugeordnet werden, wie z.B. dem angedachten Digitalministerium.

Die **Vorteile** einer staatlichen Lösung liegen in der erhöhten Transparenz und dem **öffentlichen Vertrauen**, das solche Einrichtungen genießen. Sie könnte die **Gewährleistung des Datenschutzes** und die verantwortungsvolle Behandlung personenbezogener Daten unterstützen, was für Datengebende und -nutzende von großer Bedeutung ist. Ein möglicher **Nachteil** könnte jedoch in den administrativen

Aufwänden liegen, die manchmal mit staatlichen Projekten verbunden sind. Die Flexibilität und schnelle Anpassungsfähigkeit könnten, im Vergleich zu privatwirtschaftlichen Ansätzen, eingeschränkt sein. Zudem können politische Einflüsse und Regierungswechsel die langfristige Datenökosystemstrategie beeinflussen, wie im Gründungsprozess des Dateninstituts aktuell zu beobachten.

#### Kommerzieller privater Träger

Unter einem kommerziellen privaten Träger für den Betrieb verstehen wir eine **privatwirtschaftliche Organisation, die das Datenökosystem betreibt, um mit den angebotenen Services Einnahmen zu erzielen**, die sowohl die laufenden Kosten decken als auch darüber hinaus einen Gewinn versprechen. Möglichkeiten zur Realisierung einer solchen Trägerschaft wären die Übergabe des Datenökosystems an ein bestehendes Wirtschaftsunternehmen oder die Gründung einer eigenständigen Betriebsgesellschaft mit eigener Rechtsform, wodurch die betriebliche Verantwortung klar abgegrenzt und die wirtschaftlichen Interessen gewahrt werden können.

Bei diesem Betriebsmodell sind vor allem die **Monetarisierungsoptionen des Datenökosystems** von zentraler Bedeutung, um den kommerziellen Erfolg und somit einen nachhaltigen Betrieb sicherzustellen. Wir haben folgende, in der Literatur diskutierten Alternativen exemplarisch betrachtet, vgl. den [Business Model Navigator](#) der Hochschule St. Gallen:

- Eine **kommerzielle Kooperation** mit Unternehmen, beispielsweise aus der Pharma-Branche, könnte dabei helfen, Erlöse zu generieren, da diese für die Nutzung des Datenraums zahlen können, während akademische Forschende kostenfrei Zugang erhalten.
- Ein **Freemium-Modell** bietet einen Basis-Service kostenlos an, wobei Zusatzleistungen gegen Bezahlung angeboten werden.
- Ein **two-sided market** könnte etabliert werden, bei dem ein kostenloses Angebot bereitgestellt wird, aus dem Daten gewonnen werden, die dann kostenpflichtig an Interessenten verkauft werden.
- Weitere Geschäftsmodell-Typen, die in Betracht gezogen werden könnten, sind Abonnement-Modelle, Pay-per-use, und die Lizenzierung der Daten für Drittanbieter.

Der wesentliche **Vorteil** des Betriebs durch einen kommerziellen privaten Träger liegt in der **möglichen finanziellen Unabhängigkeit und Profitabilität**, die erreicht werden kann. **Nachteile** könnten sich durch mögliche Interessenkonflikte ergeben, da kommerzielle Interessen möglicherweise die Neutralität und Gemeinwohlorientierung beeinträchtigen, verbunden mit Bedenken hinsichtlich des Datenschutzes und ethischem Umgang mit personenbezogenen Daten.

### Gemeinnütziger privater Träger

Unter Gemeinnützigkeit verstehen wir in diesem Kontext eine **privatwirtschaftliche Organisation als Betreiber des Datenökosystems, die durch ihren Betrieb lediglich die laufenden Kosten decken möchte** und dabei keine Gewinnerzielungsabsicht verfolgt. Dieses Modell kann als eine Mischform betrachtet werden, bei einer Grundfinanzierung durch die öffentliche Hand erfolgt und Weiterentwicklungen des Datenökosystems durch die projektbezogene Einwerbung von Drittmitteln abgedeckt werden müssen.

Bei der administrativen Betriebsstelle könnten verschiedene Varianten in Betracht gezogen werden. Eine Option ist die öffentliche Hand, die das Dateninstitut beispielsweise als Anstalt des öffentlichen Rechts (AöR) einrichten kann. Dies ist etwa sinnvoll, wenn das Dateninstitut zwar in öffentlicher Trägerschaft verbleiben, aber einen Teil seiner Kosten selbst decken soll. Alternativ könnte eine privatwirtschaftliche Betriebsgesellschaft eingerichtet werden, die als Vertrauensanker für die fehlende Gewinnerzielungsabsicht auch in öffentlicher Trägerschaft bleiben kann. Ein vorbildhaftes Beispiel hierfür ist die EuroDaT GmbH, die dem Privatrecht unterliegt, aber vollständig im Besitz des Landes Hessen ist.

Die **Vorteile** dieser Lösung liegen in der Möglichkeit, durch die Kombination von öffentlicher Grundfinanzierung und projektbezogenen Mitteln flexibel auf finanzielle Anforderungen zu reagieren, ohne Gewinninteressen unterliegen zu müssen. Zusätzlich kann die Einbindung öffentlicher Akteure die Vertrauenswürdigkeit und Risikoabsicherung erhöhen. **Nachteile** könnten sich im administrativen Aufwand und der Komplexität der Finanzierung über verschiedene Quellen ergeben, die zu einer gesteigerten Bürokratie führen könnten. Auch die begrenzte Gewinnorientierung könnte in manchen Fällen die Innovationsbereitschaft und die Geschwindigkeit der Weiterentwicklung einschränken.

### NACHNUTZUNG

Wie oben beschrieben beziehen wir die Nachnutzbarkeit unseres Datenökosystems auf die Möglichkeit, die von uns entwickelten Prozesse und **Konzepte sowohl für die wissenschaftliche Forschung als auch für die Entwicklung vergleichbarer Datenmodelle und -ökosysteme in verschiedenen Sektoren nutzen zu können**. Unser hierauf ausgerichtetes Nachnutzungsmodell haben wir im Abschlussbericht der [Stufe 1](#) umfassend dargelegt und im Abschlussbericht der [Stufe 2](#) weiter ausgeführt. Dabei sind wir insbesondere auf zwei Fragen eingegangen:

1. **Wie** können Nachnutzende unsere Entwicklungsergebnisse einsehen und weiterverwenden?
2. **Warum** ist eine Nachnutzung für diverse Anwendungen vorteilhaft?

Die resultierenden Konzepte, die als Antwort auf die erste Frage darlegen, wie die von uns entwickelten Komponenten für eine Nachnutzung technisch zugänglich sind, haben wir in Kapitel 04.03 dieses Berichts übersichtlich beschrieben.

Als Antwort auf die zweite Frage haben wir den Mehrwert unserer Arbeit für Datennehmende wie z.B. Forschende bereits im Abschlussbericht der [Stufe 1](#) quantitativ und qualitativ beschrieben. Wir greifen diese Mehrwert-Diskussion für Nutzende des Datenökosystems in Kapitel 04.04 dieses Berichts noch einmal auf. Darüber hinaus kann unsere Arbeit durch das Dateninstitut oder vergleichbare Enabler der Datenökonomie nachgenutzt werden. **Insbesondere können die von uns ausgearbeiteten strukturierten Ansätze für die Entwicklung neuer Datenmodelle und Datenökosysteme direkt für die Erstellung weiterer Angebote durch Dritte genutzt werden.** Die entsprechenden Überlegungen fassen wir Kapitel 07 dieses Berichts zusammen.

# Ihr Kontakt

**Dr. Robert Görke**  
Partner  
+49 162 263 1426  
[Robert.Goerke@d-fine.com](mailto:Robert.Goerke@d-fine.com)



d-fine GmbH  
An der Hauptwache 7  
60313 Frankfurt  
Deutschland

d-fine

analytical. quantitative. tech.